When Do Protests Work? *

Bocar A. Ba † Tony Cheng ‡ Iris Chang § Maxime Menuet ¶ Alexander P. Whitefield ||

March 26, 2026

Abstract

Why do protest movements fail to shift entrenched preferences yet trigger durable institutional change on adjacent margins? We develop a dynamic Bayesian persuasion model in which salience shocks interact with institutional thresholds to produce two regimes: transitory attitude shifts that fade with attention and persistent institutional adoption sustained by feedback. Under polarization, persuasion is asymmetric and concentrated among sympathetic partisans. We test the model using the 2020 Black Lives Matter protests. Protest exposure produced both a partisan shift in attitudes toward policing that dissipated within two years and simultaneous adoption of crisis response programs that persisted as attitudes reverted.

*Thanks to Naomi Murakawa, - Ba gratefully acknowledges financial support from the National Science Foundation under Grant No. 2314013 and the Duke Economics Department. Any errors are our own

†Duke and NBER. Email: bocar.ba@duke.edu

‡Duke. Email: iris.chang@duke.edu

§Duke. Email: tony.cheng@duke.edu

¶University Côte d'Azur, CNRS. Email: maxime.menuet@univ-cotedazur.fr

||Wharton. Email: whalex@wharton.upenn.edu

1 Introduction

Mass protest is one of the most visible forms of political action, but its consequences are strikingly uneven. Some protest movements appear to transform politics durably. The Civil Rights Movement helped trigger landmark federal legislation that reshaped American institutions for decades. The Tea Party mobilization left a lasting imprint on Republican electoral outcomes and congressional behavior. By contrast, other massive protest waves generate intense public attention yet leave only weak or temporary traces on public attitudes and policy. The anti-war protests of the Vietnam era altered the political climate only gradually and unevenly. Occupy Wall Street made inequality newly salient without producing commensurate policy reform. The protests following the murder of George Floyd in 2020 shifted public discourse on policing with extraordinary speed, yet much of that attitudinal change later receded. Across these cases, the central puzzle is not whether protest matters, but why some protest shocks fade while others become self-sustaining.

This paper studies protest-based persuasion both theoretically and empirically. The core theoretical question is why the same protest movement can fail to shift entrenched policy preferences while simultaneously triggering durable institutional change on a related margin. We argue that the answer lies in the interaction between salience dynamics and institutional thresholds.

We develop a dynamic model with three actors: a protest movement that shapes public information, a continuum of citizens who update beliefs and decide whether to support reform, and an institutional environment that translates sufficiently strong public support into policy responses. The central mechanism is a feedback loop between persuasion, mobilization, and institutional response. A salience shock—such as a highly publicized act of police violence—increases the prominence of a policy issue on the public agenda. Higher salience expands the set of belief distributions the movement can induce, in the spirit of Bayesian persuasion (Kamenica and Gentzkow, 2011). As more citizens update toward supporting reform, mobilization grows, which raises the probability of an institutional response. Institutional responses, in turn, reinforce salience, potentially sustaining a self-reinforcing cycle.

Whether this cycle is sustained or collapses depends on the strength of the feedback relative to the natural decay of public attention. The model generates two qualitatively distinct regimes. In the *fading* regime, salience declines faster than institutional responses can reinforce it: Persuasion attenuates, mobilization shrinks, and attitudes revert toward their preshock levels. In the *persistent* regime, the early mobilization is large enough to trig-

ger institutional responses that sustain salience above the threshold needed for continued persuasion. Small differences in the initial shock can therefore produce large differences in long-run political outcomes—a feature that helps explain why protest movements with similar levels of public support can yield very different policy legacies.

The model also incorporates polarization. Citizens differ in their prior beliefs about whether reform is warranted, and we allow these priors to vary systematically across partisan groups. Under polarization, persuasion is asymmetric: Citizens whose priors are already closer to the reform margin are more responsive to the movement's signals, while those with strong priors in favor of the status quo are largely unaffected. This generates a prediction that protest-induced attitudinal change is concentrated among sympathetic partisans, with little cross-partisan persuasion—even when the salience shock is large and the information environment is common to both groups.

A key implication concerns the distinction between attitudes and institutions. Attitudinal change requires sustained salience: Once attention fades, beliefs drift back toward prior levels. Institutional adoption, by contrast, is absorbing: Once a government creates a program, it does not automatically dismantle it when public attention moves on. This asymmetry means that even transitory persuasion can leave a permanent institutional footprint—but only on policy margins where the threshold for adoption is sufficiently low. On margins where adoption requires sustained majority support or large fiscal commitments, the same transitory persuasion produces no lasting effect. The model therefore predicts that protest movements are likeliest to generate durable change not on their primary demand but on adjacent margins where the political cost of action is lower.

We take these predictions to data on the protests that followed the murder of George Floyd in May 2020. The killing of Floyd by a Minneapolis police officer catalyzed the largest protest movement in US history, with an estimated 15 to 26 million participants (Buchanan et al., 2020). A central demand of activists at the movement's core was to "defund the police," that is, to reduce police budgets and reallocate resources toward community-based alternatives. The geographic unevenness of the protests creates cross-county variation in exposure that we exploit in a difference-in-differences (DID) design. We study two outcomes that map directly onto the model's distinction between attitudinal change and institutional adoption: public support for police funding and the adoption of police alternative programs.

For the attitudinal analysis, we use the Cooperative Election Study (CES), a nationally representative survey conducted biennially from 2014 to 2024, linked to county-level

protest data from the Global Database of Events, Language, and Tone (GDELT) database. We find that protest exposure reduced support for maintaining or increasing police funding by 3–4 pp in 2020, relative to a preperiod mean of 0.93. The effect is entirely concentrated among Democratic voters; Republican respondents show no detectable response. The estimates for 2022 and 2024 attenuate substantially, indicating that much of the initial persuasion dissipated within one to two survey waves.

The attitudinal results are not driven by a general shift in preferences on public spending. We implement a triple-difference design using attitudes toward education and infrastructure spending as placebos. Neither outcome exhibits a comparable shift, confirming that the effect operates specifically through attitudes about policing. This rules out the concern that protest exposure simply made respondents more liberal or more willing to cut government budgets across the board.

For the institutional analysis, we construct a novel county-level panel tracking the adoption of crisis response programs from 2016 to 2023. These programs represent a concrete institutional response: a new organizational unit, typically housed outside the police department, that handles a defined set of emergency calls. We find that protest exposure increased the probability of adoption of such a unit by 5.9 pp on average. In contrast to the attitudinal effects, this response has compounded over time, rising from roughly 1 pp in 2020 to 11 pp by 2023—consistent with the absorbing nature of institutional adoption in the model.

The heterogeneity in program types sharpens the interpretation. Adoption is concentrated in civilian-only, community-based programs—the organizational form most consistent with the demand for alternatives to armed police response. We observe no differential adoption of programs housed within police departments. The partisan pattern mirrors the attitudinal results: Adoption overwhelmingly occurred in Democratic-leaning counties, consistent with the model's prediction that institutional change requires the crossing of a local political threshold that is lower where prior support for reform is higher.

The contrast between the two outcomes is the paper's central empirical result. On the margin where the political and institutional thresholds for change were high, that is, police budgets, reductions in which would have required sustained public support, city council majorities, and large fiscal reallocation, the protests produced transitory persuasion that faded with salience. On the margin where those thresholds for change were low, adoption of alternative programs, which required only a discrete administrative act and modest funding typically from outside existing police budgets, the same transitory persuasion was

sufficient to trigger institutional change that persisted and compounded.

The framework's logic extends well beyond policing. Protest movements routinely produce divergent outcomes across policy margins, consistent with the distinction between transitory persuasion and durable institutional adoption. The #MeToo movement increased reporting of sex crimes by roughly 10%, an effect that persisted for at least 15 months and preceded any legislative change, suggesting that the movement shifted social norms before triggering institutional reform (Levy and Mattsson, 2025). The Civil Rights Movement followed a pattern closer to our persistent regime: Early mobilization triggered federal legislation—the Civil Rights Act of 1964 and the Voting Rights Act of 1965—that reinforced salience and sustained further institutional change, including minimum wage extensions that closed over 20% of the racial earnings gap (Derenoncourt and Montialoux, 2021), even as white racial attitudes evolved slowly and partisan realignment proceeded over decades (Kuziemko and Washington, 2018). Wasow (2020) documents this asymmetry directly: The nonviolent protests of the 1960s shifted votes toward Democrats, while violence shifted attention toward social control. Madestam et al. (2013) show that Tea Party rallies had lasting effects on Republican vote shares and congressional voting, demonstrating that even a single protest event can trigger durable political change when it activates organizing along margins where the institutionalthresholds for change are low. In Hong Kong, Cantoni et al. (2019) show that protest participation is strategic and belief dependent, and Bursztyn et al. (2021) demonstrate that one-time mobilization shocks generate persistent political engagement—but only when a sufficient share of an individual's social network is also mobilized, a micro-level analog of our salience-feedback mechanism. These cases suggest that the tension between fading persuasion and absorbing institutional adoption is a general feature of protest-driven reform.

Related Literature. This paper contributes to several strands of literature at the intersection of political economy, persuasion, and collective action.

First, it relates to a large literature on persuasion and media effects, surveyed in DellaVigna and Gentzkow (2010). A broad set of empirical studies shows that changes in the information environment can influence political behavior, including voting and policy preferences (e.g., DellaVigna and Kaplan, 2007; Enikolopov et al., 2011; Adena et al., 2015. More recent work emphasizes the limits of persuasion and the importance of prior beliefs and polarization in shaping responsiveness to information (e.g., Allcott and Gentzkow, 2017; Gentzkow and Shapiro, 2011). A key feature of most existing designs, however,

is that the informational treatment is persistent: exposure to media or communication is either continuous or repeated over time. As a result, it is difficult to distinguish between durable belief change and the mechanical effect of an ongoing information flow. In contrast, we study a setting in which persuasion arises from a transient shock—a protest event—and show that persistence is endogenous and depends on whether the initial shock generates reinforcing feedback through institutions and media.

Second, the paper is closely related to the theory of Bayesian persuasion (Kamenica and Gentzkow, 2011) and its subsequent developments (e.g., Gentzkow and Kamenica, 2016; Bergemann and Morris, 2019; Lipnowski and Ravid, 2020). We build on this framework by embedding persuasion in a dynamic political environment with endogenous salience and institutional thresholds. In our model, the key departure from the canonical persuasion framework is that the sender's payoff depends on an endogenous coordination outcome—the size of the protest—which itself depends on beliefs. This introduces a novel interaction between persuasion and collective action. Moreover, because institutional responses are threshold-based and discontinuous, the sender's payoff is generally non-concave in posterior beliefs, generating a role for concavification in a setting with political thresholds. While recent work has applied persuasion to political environments, it typically abstracts from collective action or institutional feedback. We integrate these elements in a unified framework.

Third, the paper contributes to the literature on protest, collective action, and political mobilization. A number of studies examine how communication technologies and information affect protest participation. For instance, Enikolopov et al. (2020) show that social media increases protest activity, while Acemoglu et al. (2018) study how information diffusion shapes protest dynamics during the Arab Spring. Theoretical work emphasizes how information can help resolve coordination failures (e.g., Shadmehr and Bernhardt, 2019; Edmond, 2013. Other contributions study the political consequences of protest, including effects on voting, policy, and political engagement (e.g., Madestam et al., 2013; Wasow, 2020; Cantoni et al., 2019; Bursztyn et al., 2021). This literature typically focuses either on the determinants of mobilization or on its average effects.

We complement this work by focusing on the dynamics that follow a protest shock. Even when a protest successfully coordinates large-scale participation, its long-run consequences are not guaranteed. We show that the persistence of protest effects depends on the interaction between salience dynamics and institutional thresholds. In particular, we introduce a distinction between two margins of political change: attitudes, which revert when salience

fades, and institutions, which are absorbing once adopted. This distinction implies that the same protest shock can generate transitory changes in public opinion while producing durable institutional change, depending on the cost and threshold of policy adoption.

More broadly, the paper contributes to a growing literature on belief dynamics and persistence. Existing work shows that beliefs can be shaped by long-term exposure to information or formative experiences (e.g., Alesina and Fuchs-Schündeln, 2007; Malmendier and Nagel, 2011. Our framework highlights a complementary force: belief changes induced by highly salient but transient shocks may decay unless reinforced by endogenous feedback. By explicitly modeling the interaction between persuasion, coordination, and institutional response, we provide a unified explanation for why protest movements can generate large short-run effects yet limited long-run attitudinal change, while still leaving a durable institutional footprint.

Plan The rest of this paper is organized as follows. Section 2 sets up the model with three actors—citizens, a protest movement, and an institutional authority—and defines the equilibrium. Section 3 derives the optimal persuasion strategy. Section 4 analyzes salience dynamics and persistence, introduces polarization and asymmetric protest effects, and states the testable implications. Section 5 presents the empirical application: background on the "defund the police" movement and police alternatives, the data and empirical strategy, and results on both attitudinal shifts and policy adoption. Section 6 concludes with a discussion of the implications of our findings.

2 Model

We consider a discrete-time dynamic framework in which protest operates as a mechanism of political persuasion in a polarized democracy. Time is indexed by $t \in \mathbb{N}$. The economy is populated by three types of actors. The first actor is a continuum of *citizens*. The second is a *protest organizer*, broadly defined as a collective actor seeking to promote or resist policy reform. Examples include the civil rights movement, Black Lives Matter, climate activism, gun rights organizations such as the NRA, and anti-immigration mobilizations. These actors do not directly set policy but influence political outcomes by shaping the public information environment through demonstrations, media engagement, and the dissemination of evidence. The third actor is an *institution* that aggregates citizen support and may translate sufficiently strong public pressure into observable political responses. We

interpret institutions broadly as the set of public authorities—such as governments, legislatures, courts, or administrative bodies—that can react to social mobilization through institutional reforms or adjustments to public programs.

We now describe the behavior of the three actors in the model. We then define the equilibrium concept and characterize the different dynamic outcomes generated by the political feedback described above.

2.1 Citizens

The political conflict studied in this paper concerns whether a reform should be implemented relative to the status quo. Let the latent state of the world be $\theta \in \{0,1\}$, where $\theta = 1$ indicates that reform is socially warranted and $\theta = 0$ that the status quo should be maintained. The state captures the underlying social desirability of reform, which may depend on factors such as the effectiveness of existing institutions, the severity of social problems, or the expected benefits of policy change. Reforms in this framework may concern a broad range of political issues, including policing and criminal justice, redistribution, climate regulation, gun control, immigration, or social policy. In each case, citizens face uncertainty about whether reform is truly justified relative to the status quo. All agents share a common prior belief about the state $\mu_0 = \Pr(\theta = 1)$, which reflects the prevailing public perception before the protest movement releases new information.

There is a unit-mass continuum of citizens indexed by $i \in [0,1]$. Each citizen is characterized by a type $\tau_i \in (0,1)$, which we interpret as an individual *policy threshold*. The threshold τ_i represents the minimum posterior belief that reform is socially warranted required for citizen i to support reform. It threshold summarizes a variety of determinants of political behavior, including ideological preferences, perceived costs of reform, trust in institutions, or risk aversion toward policy change. Citizens with lower thresholds are more easily persuaded, whereas citizens with higher thresholds require stronger evidence before endorsing reform.

Individual Decision Given a posterior belief μ about the state $\theta = 1$, citizen i supports reform if and only if $\mu \geq \tau_i$. Formally, the action of citizen i is

$$a_i(\mu) = \mathbf{1}\{\mu \geq \tau_i\}, \tag{1}$$

where $a_i = 1$ denotes support for reform and $a_i = 0$ denotes opposition.

Policy thresholds are distributed in the population according to a continuously differentiable cumulative distribution function G on $[0,1]$ with strictly positive density g.[1]

Aggregate Support From (1), the share of citizens who support reform conditional on belief μ is

$$G(\mu) = \int_0^{\mu} g(\tau)\, d\tau.$$

Citizens may hold heterogeneous posterior beliefs. Let $\nu \in \Delta([0,1])$ denote the distribution of beliefs in the population. Aggregate support for reform is therefore

$$P(\nu) \equiv \int_0^1 G(\mu)\, d\nu(\mu). \tag{2}$$

The function $P(\nu)$ maps the distribution of beliefs into aggregate political support. Persuasion increases support by shifting beliefs upward and moving more citizens above their evidentiary threshold.

2.2 Protest Movement

We model the protest movement as a strategic sender of public information, in the spirit of the Bayesian persuasion framework of Kamenica and Gentzkow (2011). The movement seeks to increase public support for reform by shaping the information environment in which citizens form beliefs about whether reform is socially warranted.[2]

Information Design Formally, in each period t, the protest movement chooses an information structure that induces a distribution of posterior beliefs $\nu_t \in \Delta([0,1])$. This reduced-form representation is standard in the Bayesian persuasion literature. Rather than modeling messages or signal realizations explicitly, we characterize the movement's action by the distribution of posterior beliefs it induces. Because signals are public, all citizens share the same posterior within a realization, while different realizations generate different posteriors ex ante.

[1] These regularity conditions ensure that aggregate support varies smoothly with posterior beliefs and rule out mass points in policy thresholds.

[2] Protest activity includes demonstrations, marches, symbolic actions, media engagement, the production or dissemination of evidence, and sustained attention to salient events. These actions do not directly determine policy outcomes; rather, they affect the content, salience, and credibility of public signals observed by citizens.

Feasibility Constraints Not all distributions over posterior beliefs are feasible. First, information structures must satisfy Bayes plausibility, that is, consistency with Bayesian updating from the common prior μ_0:

$$\int_0^1 \mu \, d\nu_t(\mu) = \mu_0. \tag{3}$$

Second, producing informative persuasion is costly. These costs reflect the organizational, informational, and reputational resources required to attract and sustain public attention in a congested information environment. Increasing the intensity or precision of persuasion—such as maintaining large-scale mobilization or inducing beliefs closer to certainty—requires disproportionately greater resources. We capture these costs by imposing an information budget measured by the Kullback–Leibler (KL) divergence between posterior and prior beliefs.[3] For any posterior $\mu \in (0,1)$, let

$$\mathrm{KL}(\mu \| \mu_0) = \mu \log \frac{\mu}{\mu_0} + (1-\mu) \log \frac{1-\mu}{1-\mu_0}. \tag{4}$$

The KL divergence implies that marginal deviations from the prior are relatively inexpensive, whereas inducing posterior beliefs close to certainty is increasingly costly. This convexity property yields a tractable characterization of feasible information structures while abstracting from the details of signal generation.

An information structure ν_t is feasible in period t if

$$\int_0^1 \mathrm{KL}(\mu \| \mu_0) \, d\nu_t(\mu) \leq s_t, \tag{5}$$

where $s_t \geq 0$ denotes the level of *informational salience*. This variable s_t summarizes the protest movement's effective attention capacity in period t: the amount of public attention, media coverage, and cognitive bandwidth available to process information about the reform. High-salience moments—such as widely publicized incidents or viral footage—relax the information constraint (5) and allow sharper belief updating. In our setup, salience is modeled as a stock variable, reflecting accumulated public attention and political relevance inherited from past periods. In the absence of reinforcement, salience depreciates

[3]The use of the KL divergence to model information costs is standard in the Bayesian persuasion and rational inattention literatures (e.g., Sims, 2003; Kamenica and Gentzkow, 2011). The KL divergence measures the statistical distance between posterior and prior beliefs, independently of the specific realization of signals, and depends only on the induced distribution of beliefs.

over time as public attention shifts to other issues.

Payoff Given an information structure ν_t, aggregate support for reform is $P(\nu_t)$ as defined in (2). The protest movement values aggregate support because it generates political pressure on institutions. Its per-period payoff is

$$u(s_t, \nu_t) = R(s_t)\, P(\nu_t) - C\left(\int_0^1 \mathrm{KL}(\mu \| \mu_0)\, d\nu_t(\mu)\right), \tag{6}$$

where $R(s) > 0$ scales the contemporaneous value of political support and $C : \mathbb{R}_+ \to \mathbb{R}_+$ is a cost function. Suppose that $R(\cdot)$ is bounded and continuous, $C(\cdot)$ is increasing and strictly convex with $C(0) = 0$.

Conditional on salience s_t, the protest movement solves the constrained information design problem

$$\max_{\nu_t \in \mathcal{N}(s_t)} u(s_t, \nu_t), \tag{7}$$

where the feasible set is

$$\mathcal{N}(s_t) = \left\{ \nu \in \Delta([0,1]) : \int_0^1 \mu\, d\nu(\mu) = \mu_0,\ \int_0^1 \mathrm{KL}(\mu \| \mu_0)\, d\nu(\mu) \le s_t \right\}. \tag{8}$$

Lemma 1. *Fix $\mu_0 \in (0,1)$. For every $s \ge 0$, the set $\mathcal{N}(s)$ is nonempty, convex, and weakly compact in $\Delta([0,1])$. Moreover, the correspondence $s \mapsto \mathcal{N}(s)$ is increasing and has a closed graph.*

Lemma 1 describes how the set of feasible posterior beliefs expands with the salience parameter s. Bayes plausibility fixes the average posterior at the prior μ_0, while the KL constraint bounds how far beliefs can deviate from it. A higher s relaxes this constraint, allowing more informative and more dispersed posterior distributions. Formally, this is captured by the monotonicity of $s \mapsto \mathcal{N}(s)$: Greater salience enlarges the feasible set without altering its convex and compact structure, with $\mathcal{N}(0) = \{\delta_{\mu_0}\}$ as the no-salience benchmark.

2.3 Institutional Response

Institutions translate public support into observable political responses. We interpret "institutions" broadly as the decision-making layer capable of reacting to protest pressure, including elected officials, legislatures, courts, and administrative authorities. Institutional

responses may take the form of policy reforms, legislative changes, or adjustments in public programs. Examples include the introduction of alternative crisis response programs following Black Lives Matter mobilizations, the expansion of civil rights legislation during the Civil Rights Movement, tax reductions announced in response to the Yellow Vest protests in France, or policy decisions related to the end of the Vietnam War following sustained antiwar protests. Such responses increase the visibility and perceived relevance of the issue in the public debate and can sustain political attention over time.

General Response Function Institutional reinforcement is given by

$$F(v) = f\big(P(v)\big), \tag{9}$$

where $f : [0,1] \to [0,1]$ is continuous and weakly increasing. We assume that there exists $\bar{P} \in (0,1)$ such that $f(p) = 0$ for $p \leq \bar{P} - \varepsilon$, and $f(p) = 1$ for $p \geq \bar{P} + \varepsilon$, for some $\varepsilon > 0$.

This formulation generalizes the deterministic threshold rule $\mathbf{1}\{P(v) \geq \bar{P}\}$ by allowing for implementation frictions and uncertainty around the threshold. When aggregate support is well below $\bar{P}$, institutional action does not occur; when it is well above $\bar{P}$, reform is implemented with certainty. For intermediate levels of support, responses are gradual. The threshold $\bar{P}$ summarizes institutional inertia and political constraints and can be interpreted as the level of public support required for reform to become politically viable given electoral incentives, veto points, and opposition from organized interests.[4]

Salience Dynamics Salience evolves according to

$$s_{t+1} = \delta s_t + \eta F(v_t) = \delta s_t + \eta f\big(P(v_t)\big). \tag{10}$$

This law of motion formalizes how attention to the issue evolves over time. The parameter $\delta \in (0,1)$ captures the natural depreciation of attention: In the absence of institutional reinforcement ($F = 0$), salience decays geometrically and converges to zero, reflecting the transitory nature of attention. Higher values of δ imply greater persistence of political relevance. On the other hand, $\eta \geq 0$ captures institutional reinforcement. When institutions respond to public support through policy announcements, investigations, or legislative ac-

[4] In a fully specified voting or bargaining model (which we do not develop here for tractability), $\bar{P}$ would be endogenously determined by electoral competitiveness and the distribution of political power. In the present framework, it serves as a parsimonious reduced-form measure of the political difficulty of reform.

tion, they sustain attention by generating new information and legitimacy. A larger η therefore implies a stronger feedback from institutional action to future salience.

Eq. (10) highlights a key mechanism of the model: Protest-induced persuasion can temporarily raise support and salience, but persistence requires institutional reinforcement. Absent such reinforcement, salience decays and the protest movement's capacity to persuade diminishes over time.

Lemma 2. *For any two information structures ν and ν' such that $P(\nu) \leq P(\nu')$, we have $s_{t+1}(\nu) \leq s_{t+1}(\nu')$.*

Lemma 2 establishes a monotone feedback between persuasion and future attention. Information structures that generate higher aggregate support lead to greater institutional reinforcement and, in turn, to higher future salience. As a result, successful persuasion today expands the movement's future informational capacity. This complementarity underlies the dynamic incentives faced by the protest movement.

2.4 Timing

In each period t, events unfold as follows:

1. The current level of salience s_t is realized and observed by all agents.
2. The protest movement chooses an information structure $\nu_t \in \mathcal{N}(s_t)$.
3. Citizens update beliefs according to Bayes's rule and choose whether to support reform.
4. Aggregate support $P(\nu_t)$ is realized.
5. Institutions respond according to $F(\nu_t) = f(P(\nu_t))$.
6. Salience updates to s_{t+1} according to $s_{t+1} = \delta s_t + \eta F(\nu_t)$.

2.5 Equilibrium

Given the timing described above, the protest movement faces a dynamic information design problem with salience as the state variable. We focus on stationary Markov equilibria.

Definition 1 (Markov-Perfect Bayesian Persuasion Equilibrium)**.** *A Markov-perfect Bayesian persuasion (MPBP) equilibrium consists of*

1. *a stationary persuasion policy* $\nu^* : \mathbb{R}_+ \to \Delta([0,1])$*,*

2. *a value function* $V : \mathbb{R}_+ \to \mathbb{R}$*,*

3. *citizens' belief updating satisfying Bayes's rule and threshold strategies* (1)

4. *and an institutional response rule* $F(\nu) = f(P(\nu))$*,*

such that for every $s \geq 0$*,*

$$V(s) = \max_{\nu \in \mathcal{N}(s)} \left\{ R(s)P(\nu) - C\left(\int_0^1 \mathrm{KL}(\mu \| \mu_0)\, d\nu(\mu) \right) + \beta V(\delta s + \eta f(P(\nu))) \right\}, \tag{11}$$

and $\nu^*(s)$ *attains the maximum.*

In equilibrium, the protest movement chooses persuasion strategies as a function of current salience, taking into account how institutional responses affect the future evolution of political attention.

Proposition 1. *A MPBP equilibrium exists.*

The existence follows from standard dynamic programming arguments. For each s, the feasible set of posterior distributions $\mathcal{N}(s)$ is compact under weak convergence. The objective in (11) is continuous in ν since both $P(\nu)$ and the KL cost depend continuously on the distribution of posteriors. With $\beta < 1$, the Bellman operator is a contraction on the space of bounded continuous value functions. A fixed point therefore exists, and an optimal stationary policy $\nu^*(s)$ attains the maximum.

3 Optimal Persuasion

After establishing the existence of an MPBP equilibrium, we now characterize the protest movement's equilibrium information design problem within a period, conditional on a given level of salience s. This analysis serves two purposes. First, it clarifies how protest strategically shapes beliefs at a given point in time. Second, it identifies the endogenous limits of persuasion that discipline the dynamic equilibrium and generate short-run versus persistent effects.

Intra-Period Persuasion Problem Throughout this section, we fix a salience level $s \geq 0$ and take the continuation value function $V(\cdot)$ as given. The protest movement chooses an information structure $\nu \in \Delta([0,1])$ to maximize

$$\max_{\nu \in \mathcal{N}(s)} \left\{ \int_0^1 \tilde{v}_s(\mu)\, d\nu(\mu) - C\left(\int_0^1 \mathrm{KL}(\mu \| \mu_0)\, d\nu(\mu) \right) \right\}, \tag{12}$$

where the indirect payoff from inducing posterior belief μ at salience s is

$$\tilde{v}_s(\mu) = R(s)\, G(\mu) + \beta V(\delta s + \eta f(G(\mu)))\,. \tag{13}$$

The feasible set of information structures $\mathcal{N}(s)$ (Eq. 8) imposes Bayes plausibility and a salience-dependent limit on persuasive capacity, so that higher salience allows for greater dispersion in posterior beliefs.

The objective in (12) combines three components. The first term ($R(s)G(\mu)$) captures the contemporaneous political value of persuasion. The second term (($\beta V(\cdot)$) reflects a dynamic incentive: Sufficiently strong persuasion may trigger institutional reinforcement and sustain future salience. The third term represents the strictly convex cost of producing informative and salient persuasion.

For convenience, define information intensity

$$I(\nu) \equiv \int_0^1 \mathrm{KL}(\mu \| \mu_0)\, d\nu(\mu).$$

Problem (12) is not linear in ν because the information cost is a strictly convex function of aggregate information intensity. The next lemma shows that, despite this nonlinearity, optimal persuasion can be characterized in a standard Bayesian persuasion problem with a linear KL penalty.

Lemma 3. *Let ν^* solve (12), and let $I^* \equiv I(\nu^*)$. Then, there exists $\lambda \geq 0$ such that ν^* also solves*

$$\max_{\nu \in \Delta([0,1])} \int_0^1 \left(\tilde{v}_s(\mu) - \lambda\, \mathrm{KL}(\mu \| \mu_0) \right) d\nu(\mu) \quad s.t \quad \int_0^1 \mu\, d\nu(\mu) = \mu_0. \tag{14}$$

Moreover, λ can be chosen as $\lambda = C'(I^*) + \kappa$, where $\kappa \geq 0$ is the Lagrange multiplier on the feasibility constraint $I(\nu) \leq s$, satisfying complementary slackness $\kappa\,(I^* - s) = 0$.

Lemma 3 shows that the strictly convex information cost and the salience-based capacity constraint jointly determine an effective marginal cost of informativeness. Optimal

persuasion can therefore be characterized by the linearly penalized problem (14), with λ summarizing both forces.

Concavification To characterize optimal persuasion, we rely on standard geometric arguments from the Bayesian persuasion literature (Kamenica and Gentzkow, 2011; Gentzkow and Kamenica, 2017; Bergemann and Morris, 2016). Define the effective payoff

$$\phi_{s,\lambda}(\mu) = \tilde{v}_s(\mu) - \lambda \, \mathrm{KL}(\mu \| \mu_0). \tag{15}$$

Problem (14) is linear in ν subject to Bayes plausibility and therefore admits a standard concavification characterization.

Proposition 2. *Fix $s \geq 0$ and $\lambda \geq 0$. Let ν^* solve (14). Then, there exists an affine function $\ell(\mu) = a + b\mu$ such that:*

1. *$\ell(\mu) \geq \phi_{s,\lambda}(\mu)$ for all $\mu \in [0,1]$;*
2. *$\ell(\mu) = \phi_{s,\lambda}(\mu)$ for ν^*-almost every μ.*

Equivalently, induced posterior beliefs lie in the contact set of a common supporting line of the upper concave envelope of $\phi_{s,\lambda}$.

Proposition 2 characterizes the structure of optimal persuasion for a given effective payoff $\phi_{s,\lambda}$. Whenever this payoff is locally nonconcave, optimal information design exploits the nonconcavity by inducing belief dispersion through a common-tangent construction. If $\mathrm{supp}(\nu^*) = \{\mu_L, \mu_H\}$ with $\mu_L < \mu_H$, then the supporting line is tangent at both points, and Bayes plausibility pins down the mixing weights so that $\mathbb{E}_{\nu^*}[\mu] = \mu_0$. Importantly, this characterization is conditional on the feasible set of posterior beliefs.

Salience affects persuasion precisely through this feasibility channel. By relaxing the constraint on information intensity, higher salience enlarges the feasible set $\mathcal{N}(s)$ and lowers the effective marginal cost of informativeness summarized by λ. As a result, nonconcavities of $\phi_{s,\lambda}$ that are irrelevant at low salience may become exploitable at higher salience. The next proposition formalizes this insight by distinguishing low- and high-salience regimes.

Proposition 3. *Suppose that $\tilde{v}_s(\mu)$ is continuous in (s,μ) and twice continuously differentiable in μ.*

1. *(*Low salience*) For every $\epsilon > 0$, there exists $\underline{s}(\epsilon) > 0$ such that for all $s \in [0, \underline{s}(\epsilon)]$ and for every optimal solution $v^*(s)$ to the intra-period persuasion problem,*

$$\int_{\{|\mu-\mu_0|\geq\epsilon\}} dv^*(s)(\mu) \;\leq\; \epsilon.$$

That is, when salience is low, the KL constraint forces optimal posterior beliefs to concentrate arbitrarily tightly around the prior μ_0.

2. *(*High salience*) Assume that there exist $(s^\dagger, \lambda^\dagger)$ such that the effective payoff $\phi_{s^\dagger,\lambda^\dagger}$ is locally nonconcave in a neighborhood of μ_0. Then, there exists $\bar{s} < \infty$ such that for all $s \geq \bar{s}$, the intraperiod persuasion problem admits at least one optimal information structure $v^*(s)$ that assigns positive probability to posterior beliefs both below and above μ_0. Moreover, such a solution can be chosen with two-point support $\{\mu_L, \mu_H\}$, where $\mu_L < \mu_0 < \mu_H$ and $\mathbb{E}_{v^*(s)}[\mu] = \mu_0$.*

Figure 1 provides a stilized illustration of Propositions 2 and 3. The figure plots the effective payoff from posterior beliefs, $\phi_{s,\lambda}(\mu)$, together with its upper concave envelope.

In a low-salience environment **(a)**, the information constraint tightly restricts the dispersion of posterior beliefs around the prior μ_0. The relevant portion of $\phi_{s,\lambda}(\mu)$ is therefore locally concave, and the concave envelope coincides with the payoff function itself. Optimal persuasion induces only small belief shifts, and selective information design does not arise. In a high-salience environment **(b)**, in contrast, the feasible set of posterior beliefs expands. If $\phi_{s,\lambda}(\mu)$ is nonconcave in a neighborhood of μ_0, the upper concave envelope features a linear segment supported by a common tangent. Optimal persuasion then places the probability mass on extreme posterior beliefs μ_L and μ_H, with Bayes plausibility pinning down the mixing weights so that $\mathbb{E}[\mu] = \mu_0$.

These results highlight an endogenous ceiling on persuasion at each level of salience. When salience is low, the feasible set of posterior beliefs is tightly constrained by the movement's informational capacity. Optimal persuasion therefore takes the form of a nearly uniform message: Citizens receive essentially the same signal, posterior beliefs remain close to the prior, and aggregate support can only adjust marginally. In such environments, shifts in public opinion are limited and short-lived. When salience is high, the feasible set of posterior beliefs expands and the protest movement can exploit nonconcavities in the effective payoff through concavification. Optimal persuasion then becomes selective. Rather than moving all beliefs slightly, the movement induces dispersion in posterior beliefs, concentrating strong persuasive signals on citizens who are most responsive while weaker signals

reach others.

This strategy allocates persuasive effort where it has the largest marginal impact on political support. Because institutions respond to aggregate support rather than to belief dispersion itself, selective persuasion can produce discrete increases in political pressure. Persistence arises only if these increases are large enough to trigger institutional reinforcement and sustain salience.

The next section combines these intraperiod results with the law of motion for salience to characterize the global dynamics of persuasion and political persistence.

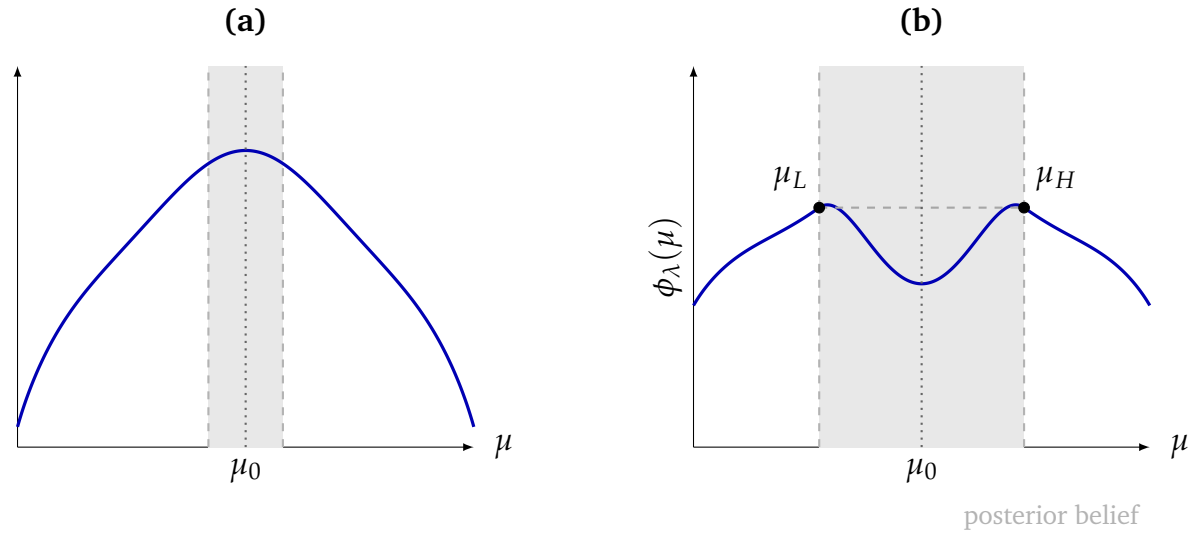


Figure 1: **Salience and concavification in optimal persuasion**
Notes: Panel (a) depicts a low-salience environment in which feasible posterior beliefs are tightly concentrated around the prior μ_0 (shaded band). Panel (b) depicts a high-salience environment with a wider feasible set. When $\phi_\lambda(\mu)$ is locally nonconcave, its upper concave envelope features a linear segment (dashed) over $[\mu_L, \mu_H]$, implying selective persuasion between μ_L and μ_H with mixing weights pinned down by Bayes plausibility, $\mathbb{E}_\nu[\mu] = \mu_0$.

4 Salience Dynamics and Persistence

So far, we have characterized optimal persuasion conditional on a fixed level of salience s. To understand the equilibrium evolution of political attention, this intraperiod problem must now be embedded into the dynamic law of motion for salience.

Recall that equilibrium salience evolves according to

$$s_{t+1} = \Phi(s_t) \equiv \delta s_t + \eta\, f\big(P(v^*(s_t))\big),$$

where $\nu^*(s)$ denotes the optimal persuasion policy at salience s, and

$$P(\nu) \equiv \int_0^1 G(\mu)\, d\nu(\mu)$$

is aggregate support for reform generated by the induced distribution of beliefs.

A key object governing these dynamics is the endogenous *persuasion frontier*:

$$\overline{P}(s) \equiv \sup_{\nu \in \mathcal{N}(s)} P(\nu),$$

which represents the maximal aggregate support attainable at salience s subject to Bayes plausibility and the KL feasibility constraint. By Lemma 1, this supremum is well defined and attained. Moreover, because the feasible set $\mathcal{N}(s)$ expands with salience, the frontier $\overline{P}(s)$ is weakly increasing: Higher salience relaxes informational constraints and allows the movement to generate larger belief dispersion and therefore higher aggregate support.

It is useful to associate to $\overline{P}$ the *upper-bound salience map*

$$\overline{\Phi}(s) \;\equiv\; \delta s + \eta\, f(\overline{P}(s)).$$

Along any equilibrium path, optimal persuasion satisfies $P(\nu^*(s)) \leq \overline{P}(s)$, implying $s_{t+1} = \Phi(s_t) \leq \overline{\Phi}(s_t)$, with equality whenever the equilibrium policy attains the persuasion frontier.

In general, the map Φ is nonlinear and may admit multiple stationary points. This nonlinearity arises from the interaction of two forces.

First, institutional responses are threshold based. By assumption, $f(p) = 0$ for $p \leq \bar{P} - \varepsilon$ and $f(p) = 1$ for $p \geq \bar{P} + \varepsilon$, with a smooth transition in between. Institutions therefore react little to marginal changes in support far from the threshold but respond strongly when aggregate support approaches the politically relevant region.

Second, the persuasion frontier $\overline{P}(s)$ is itself endogenous to optimal information design. As we have seen, when salience is low, the KL constraint tightly limits belief dispersion, forcing posterior beliefs to remain close to the prior. In this region, aggregate support responds only weakly to persuasion, and $\overline{P}(s)$ increases slowly with salience. Once salience becomes sufficiently high for concavification to be exploited, persuasion becomes selective: The movement concentrates persuasive effort on belief realizations that generate the largest political impact. Aggregate support then becomes much more responsive to increases in salience.

These two forces can make the composite map $s \mapsto f(\overline{P}(s))$ locally convex and globally S-shaped, as illustrated in Figure 2.c. Such shapes are the canonical source of multiple stationary salience levels.

Regime Characterization To study the local dynamics, suppose derivatives exist. The slope of the upper-bound map is

$$\overline{\Phi}'(s) = \delta + \underbrace{\eta\, f'(\overline{P}(s))\, \overline{P}'(s)}_{\text{endogenous reinforcement gain}} .$$

The first term, δ, captures the mechanical depreciation of political attention. The second term represents an *endogenous reinforcement effect*: Higher salience expands feasible persuasion (captured by $\overline{P}'(s)$), which raises maximal attainable aggregate support—institutions then convert marginal support into reinforcement at rate $f'(\cdot)$, scaled by η. Nonlinear dynamics arise precisely when this reinforcement effect is strong enough to offset depreciation.

We can now characterize the qualitative regimes of equilibrium salience dynamics.

Proposition 4. *The salience dynamics fall into one of three regimes:*

1. *Transitory regime If* $\sup_{s\geq 0} \overline{P}(s) \leq \bar{P} - \varepsilon$, *then* $s^\star = 0$ *is the unique fixed point of* Φ, *and* $s_t \to 0$ *for all initial conditions.*

2. *Persistent regime If there exists* s *such that* $\overline{P}(s) \geq \bar{P} + \varepsilon$ *and* $\delta s + \eta \geq s$, *then* Φ *admits at least one locally stable fixed point* $s^* > 0$.

3. *Tipping regime (multiplicity) If* $\sup_{s\geq 0} \overline{P}(s) \geq \bar{P} + \varepsilon$ *and there exists* $s^\dagger$ *with* $\overline{P}(s^\dagger) \in (\bar{P} - \varepsilon, \bar{P} + \varepsilon)$ *such that*

$$\eta f'(\overline{P}(s^\dagger))\overline{P}'(s^\dagger) > 1 - \delta,$$

then Φ *admits three fixed points* $0 \leq s_L < s_M < s_H$, *where* s_L *and* s_H *are locally stable and* s_M *is unstable.*

Fixed points of the salience dynamics solve $\Phi(s) = s$. Because Φ is continuous and increasing, the number and stability of steady states depend on whether the salience map is locally steeper than the 45-degree line. Figure 2 illustrates the three regimes characterized in Proposition 4.

When the persuasion frontier never reaches the institutional response window ($\overline{P}(s) \leq \bar{P} - \varepsilon$ for all s), institutional reinforcement is never activated, and $f(\overline{P}(s)) \equiv 0$. The dynamics reduce to $\Phi(s) = \delta s$, so sallience depreciates mechanically. As discussed, in this case, persuasion remains diffuse: Informational constraints keep posterior beliefs close to the prior, and aggregate support reacts only weakly to increases in salience. Even large protest efforts may generate temporary visibility, but these effects are not strong enough to trigger institutional responses. Hence, salience converges monotonically to zero (Figure 2.a).

By contrast, when salience is sufficiently high that the persuasion frontier lies above the institutional response region ($\overline{P}(s) \geq \bar{P} + \varepsilon$), institutional reinforcement becomes saturated, and the dynamics are locally affine, $\Phi(s) = \delta s + \eta$. Since the slope is $\delta < 1$, any intersection with the 45-degree line yields a locally stable steady state $s^* > 0$ (Figure 2.b).

At such a point, salience becomes self-sustaining. High attention relaxes informational constraints enough for persuasion to become selective and politically effective, consistently generating aggregate support above the institutional threshold. Institutional responses feed attention back into the system at rate η, offsetting the natural depreciation of attention. Small deviations from s^* therefore dissipate over time, making the equilibrium locally stable.

Multiplicity arises in the intermediate region where institutional responsiveness is locally steep and persuasion feasibility accelerates with salience. In this regime, the reinforcement term $\eta\, f'(\overline{P}(s))\, \overline{P}'(s)$ can dominate the depreciation term $(1 - \delta)$, so that $\Phi'(s) > 1$. The salience map then crosses the 45-degree line three times, generating a low stable equilibrium s_L, an unstable threshold s_M, and a high stable equilibrium s_H (Figure 2.c). Below the threshold s_M, persuasion remains too diffuse to activate institutional responses, and salience gradually declines toward s_L. Above s_M, however, a marginal increase in attention relaxes informational constraints enough to raise aggregate support into the institutional response region. Institutional reinforcement then amplifies future salience, pushing the system toward the high-attention equilibrium s_H.

The tipping regime highlights a sharp distinction between *reform being on the agenda* (via the initial condition s_0) and *reform being in a self-reinforcing regime*. An issue may be publicly visible and politically contested yet remain below the unstable threshold s_M, in which case salience inevitably fades. By contrast, shocks that push initial salience s_0 above s_M move the system into a region where persuasion becomes highly elastic and institutional responsiveness is active, so that reinforcement offsets depreciation and attention becomes persistent. Initial conditions therefore play a decisive role: Otherwise similar

protest episodes can generate radically different long-run outcomes depending on whether they cross the tipping region.

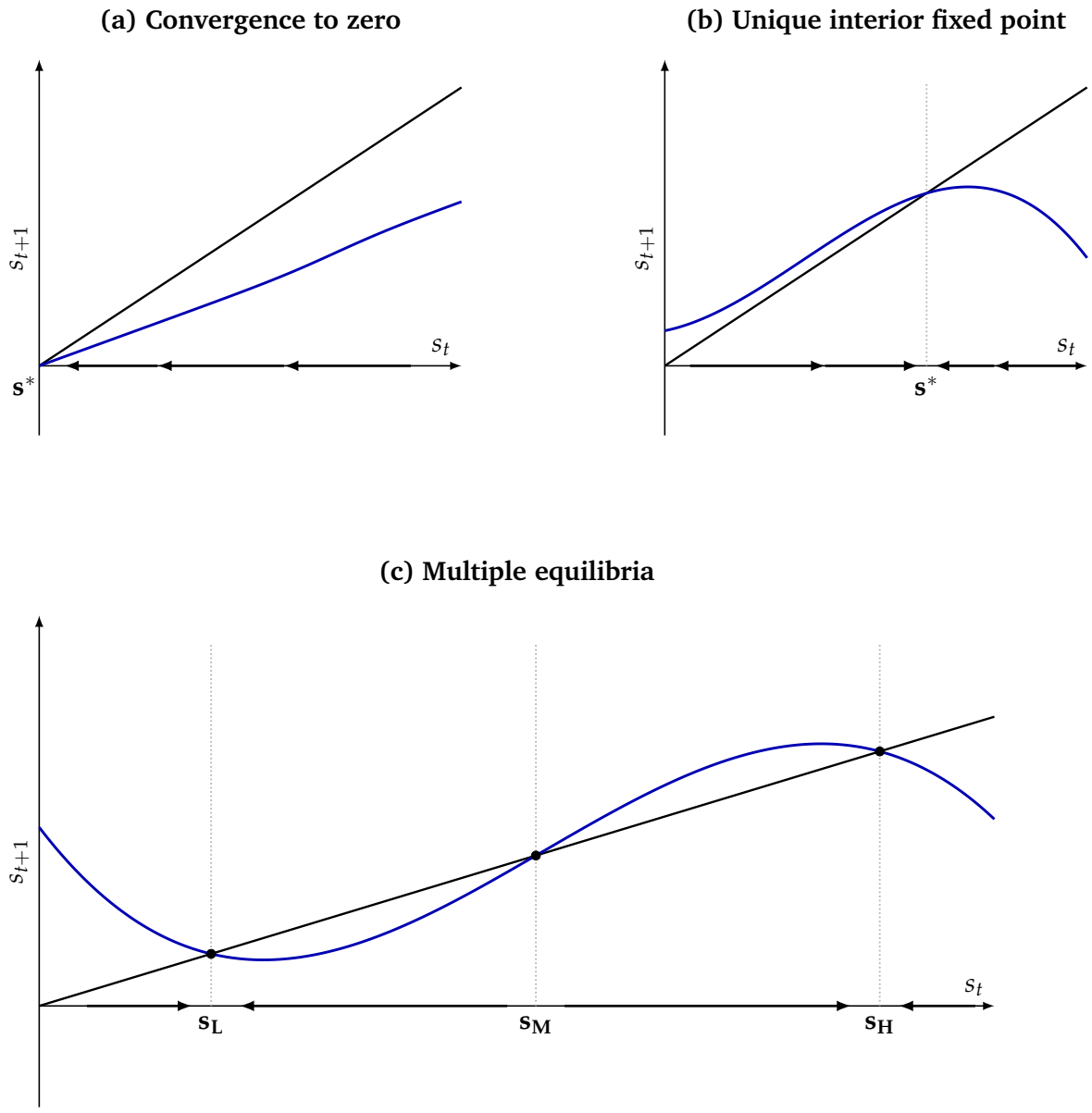


Figure 2: **Salience dynamics** Black: $s_{t+1} = s_t$. Blue: $s_{t+1} = \Phi(s_t)$. Fixed-point labels and dynamics arrows are placed on the x-axis ($y = 0$).

4.1 Discussion

Protest movements differ sharply in their long-run impact. The model highlights three key forces: initial salience s_0, the effectiveness of persuasion at that level of attention (the per-

suasion frontier $\overline{P}(s)$), and the strength of institutional feedback $f(\cdot)$. Their interaction determines whether protest generates a temporary spike in attention or a self-reinforcing political dynamic. When salience remains below the threshold required to trigger institutional feedback, persuasion has limited effects and attention gradually fades. When this threshold is crossed, institutional responses reinforce salience, sustaining persuasion and political change.

This framework yields three main trajectories. First, *transitory persuasion*: protest generates a large but short-lived increase in attention and support, insufficient to trigger sustained institutional response. The protests following George Floyd's murder illustrate this pattern. They produced a sharp rise in attention and support for policing reform, but much of this effect dissipated over time, even as some adjacent policies—such as crisis-response programs—were adopted.

Second, protest may trigger strong political reactions without stabilizing the original issue. The French Yellow Vest movement generated high salience and significant policy concessions, yet the political agenda shifted away from the initial environmental reform toward broader concerns about taxation and purchasing power. In this case, institutional feedback redirected rather than sustained the initial conflict.

Third, *self-reinforcing dynamics* arise when early mobilization is sufficient to sustain salience through institutional feedback. Antiwar protests during the Vietnam War and the civil rights movement both illustrate this regime: sustained mobilization kept these issues central to public debate, allowing persuasion and institutional responses to reinforce each other and ultimately produce major policy change.

The framework also implies the possibility of *multiple equilibria*. When feedback is strong, small differences in initial conditions can generate sharply divergent outcomes. The #MeToo movement, for example, transformed longstanding concerns into widespread institutional change after a large salience shock, while earlier mobilizations had more limited effects. Similarly, the 1989 protests in Eastern Europe show how small initial mobilizations can trigger rapid regime change once attention crosses a critical threshold.

Overall, the model implies that protest success depends not only on generating attention but on converting attention into sustained persuasion and institutional feedback. Below the threshold, protest fades; above it, political change becomes self-reinforcing.

4.2 Extension: Polarization and Asymmetric Protest Effects

So far, citizens differed only in their policy thresholds. We now extend the framework to allow for political polarization. In many policy domains, citizens belong to political constituencies that differ systematically in their receptiveness to reform. We capture this feature by allowing both the distribution of policy thresholds and the initial salience of the issue to differ across political groups.

Consider two political constituencies indexed by $p \in \{D, R\}$, which may be interpreted as Democratic- and Republican-leaning voters. Let $\lambda_D \in (0,1)$ denote the population share of group D and $\lambda_R = 1 - \lambda_D$ the share of group R. Within each group p, policy thresholds are distributed according to a cumulative distribution function G_p with density g_p. Aggregate support for reform becomes

$$G(\mu) = \lambda_D G_D(\mu) + \lambda_R G_R(\mu).$$

We focus on reforms that are, on average, more receptive among Democratic voters—such as policing reform following the Black Lives Matter movement or environmental regulation following large-scale climate mobilizations. This asymmetry is captured by assuming that G_D first-order stochastically dominates G_R.

Polarization affects protest dynamics through two complementary channels: differences in initial salience and differences in policy thresholds.

Initial Salience Political environments may start from different levels of attention for a given issue. Let $s_{0,D}$ and $s_{0,R}$ denote initial salience in Democratic-leaning and Republican-leaning jurisdictions, respectively, since some electorates may already be more attentive to issues such as policing, racial justice, or inequality.

When salience dynamics exhibit multiplicity, as in the tipping regime of Proposition 4, these differences are amplified. If, e.g., $s_{0,D} > s_M$ and $s_{0,R} < s_M$, where s_M denotes the unstable tipping threshold, then salience converges toward a positive steady state in Democratic jurisdictions but decays toward zero in Republican ones—even if the persuasion frontier $\overline{P}(s)$ and institutional response function $f(\cdot)$ are identical across environments. Protest may therefore generate visible and politically meaningful short-run effects in one constituency while remaining transient in another simply because one electorate starts closer to the self-reinforcing region of the salience dynamics.

Polarization and Heterogeneous Returns to Persuasion Polarization also affects the marginal returns to persuasion. Aggregate support for reform depends on

$$G'(\mu) = \lambda_D g_D(\mu) + \lambda_R g_R(\mu),$$

where g_p denotes the density of thresholds within group p. Because G_D first-order stochastically dominates G_R, Democratic voters have systematically lower thresholds for supporting reform. In a neighborhood of the prior μ_0, $g_D(\mu)$ is therefore relatively large while $g_R(\mu)$ is small. As a result, moderate belief shifts generate substantial increases in Democratic support but negligible changes in Republican support. Since the KL information cost grows convexly in the distance from the prior, inducing posterior beliefs large enough to cross Republican thresholds is disproportionately costly. Optimal persuasion thus concentrates belief dispersion where persuasion is both cheaper and politically more effective—among lower-threshold constituencies.

Proposition 5. *Suppose that G_D first-order stochastically dominates G_R and that salience dynamics are either transitory or in the tipping regime of Proposition 4. Then, there exist salience levels s such that the optimal information structure $\nu^*(s)$ induces belief dispersion* $\text{supp}(\nu^*(s)) = \{\mu_L, \mu_H\}$ *with $\mu_L < \mu_0 < \mu_H$, and*

$$G_D(\mu_H) - G_D(\mu_0) > 0 \qquad \textit{while} \qquad G_R(\mu_H) - G_R(\mu_0) \approx 0.$$

Consequently, protest increases support and salience in Democratic-leaning jurisdictions while generating little or no persistent effect in Republican-leaning ones.

These mechanisms provide a unified explanation for asymmetric protest effects under polarization. Differences in initial salience determine whether protest begins inside the self-reinforcing region of the salience dynamics, while differences in policy thresholds determine where persuasion yields the largest political returns.

In practice, issues such as police reform, racial justice, or inequality tend to be more salient in Democratic-leaning environments. Protest shocks therefore start closer to the tipping region and are likelier to generate short-run changes in policy attitudes. Yet because persuasion primarily mobilizes lower-threshold constituencies, aggregate support may still remain below the institutional threshold required for durable reform.

This logic helps rationalize empirical patterns observed during the Black Lives Matter protests of 2020, which reduced support for police funding in Democratic jurisdictions in

the short run while having little effect in Republican ones and produced limited long-run institutional change. More broadly, the framework explains why protest often appears highly effective within sympathetic constituencies yet fails to reshape broader political coalitions: Polarization makes persuasion asymmetric, and persistence depends on initial salience and institutional responsiveness.

4.3 Testable Implications

The framework yields several empirical implications that can be tested in the context of recent protest movements. In particular, the model predicts distinct patterns in the short- and long-run effects of protest on public opinion and policy responses.

Implication 1: Salience Shocks Generate Short-Run Persuasion When a highly visible event increases political attention, the feasible set of persuasion expands. The model therefore predicts that large salience shocks should produce immediate changes in public opinion. However, when aggregate support remains below the institutional threshold required to sustain political feedback, these opinion changes should be temporary. As salience depreciates over time, attitudes gradually revert toward their prior level.

Implication 2: Polarization Generates Asymmetric Persuasion Effects Under political polarization, persuasion primarily affects constituencies with lower policy thresholds. As shown in Proposition 5, belief shifts are therefore more likely to translate into changes in support among Democratic-leaning voters than among Republican-leaning ones. Protest should thus produce stronger opinion shifts in politically receptive environments, while having limited effects in electorates where policy thresholds are systematically higher.

Implication 3: Institutional Responses May Occur Along Adjacent Policy Margins Even when protest does not generate durable opinion shifts for the most radical policy demands, it may still induce institutional responses along related policy dimensions. In the model, such responses arise when persuasion increases aggregate support enough to activate partial political feedback without fully stabilizing the salience dynamics.

5 Empirical Application

The protests following the murder of George Floyd in May 2020 provide a natural setting to test our model. The event generated an unusually large and sudden increase in public attention to issues of police and racial justice, creating a plausibly exogenous shock to political importance.

Our empirical analysis examines three predictions of the model. First, we test whether support for radical reforms such as defunding the police increased in the short run butdeclined as attention faded. Second, we examine whether these opinion changes were concentrated among Democratic-leaning voters, consistent with asymmetric persuasion under polarization. Third, we analyze whether local governments responded by expanding alternative crisis response policies or nonpolice emergency programs, particularly in Democratic jurisdictions.

Together, these tests allow us to evaluate whether protest-induced salience shocks translate into temporary persuasion, asymmetric opinion change, and partial policy responses, as predicted by the model.

5.1 Background: From "Defund the Police" to Police Alternatives

After George Floyd's murder, activists' demand to defund the police faced two constraints that made the durable attitudinal change difficult to sustain. First, cuts in police funding are broadly perceived as threatening public safety (Vaughn et al., 2022), and experimental evidence shows that information on police violence changes attitudes only modestly and primarily among copartisans (Boudreau et al., 2019). Second, historical episodes of Black political mobilization against state violence have followed a similar pattern: initial gains among sympathetic partisans, backlash among opponents, and incremental rather than transformative policy change (Wasow, 2020; Francis and Wright-Rigueur, 2021; Knight and Weaver, 2025). The outcomes of the 2020 movement proved no exception. In Minneapolis, its epicenter, nine city council members pledged to dismantle the police department; less than two years later, voters rejected a ballot initiative to replace it (Phelps, 2024). Nationally, early evidence shows no systematic reductions in municipal police budgets after the protests (Ebbinghaus et al., 2025). Whether the protests moved public opinion toward defunding—even temporarily—is the first question we take to the data.

However, the absence of literal defunding does not imply the absence of institutional change. A concrete component of the reallocation demand was the creation of police

alternative programs: crisis response models that dispatch mental health professionals, social workers, or paramedics to calls—such as mental health crises, substance use, or homelessness—that would otherwise receive a traditional police response. These programs vary along two dimensions: the degree of police involvement (coresponse with officers vs. civilian-only dispatch) and the institutional home (police departments, fire/EMS teams, or community-based organizations). In contrast to direct budget cuts, they are typically financed through municipal general funds, nonprofit contracts, or state and federal grants, lowering the political threshold for adoption. Emerging evidence indicates broad public support for such programs and suggests that they can reduce crime without compromising public safety (Dee and Pyne, 2022; Ba et al., 2025). Whether the 2020 protests accelerated the adoption of these programs—translating the reallocation demand into concrete institutional change—is the second question we examine.

5.2 Data Sources and Sample Construction

5.2.1 Measuring Salient Protests

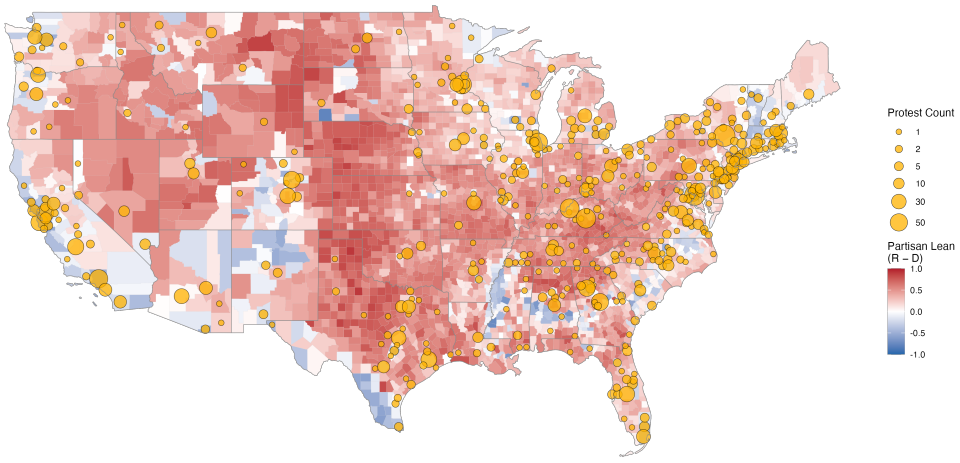
We operationalize exposure to the 2020 protests using GDELT, which codes events from global news sources using the CAMEO framework, focusing on the period May 25–July 31, 2020. We match protests to killings by police using Mapping Police Violence (MPV), searching linked news article URLs for victims' first and last names and assigning each protest two locations: the protest site and the nearest police department.

Figure 3 maps the geographic distribution of protests following the murder of George Floyd. Protests occurred in over XXXX counties, with about XXXX% of them experiencing protests. Notably, protests were not confined to Democratic strongholds—smaller protests took place in Republican-leaning counties throughout the Midwest, South, and Mountain West. However, protest intensity was strongly correlated with urbanization and Democratic partisanship: the highest protest counts clustered in major metropolitan areas along the coasts and in cities such as Minneapolis, Denver, and Chicago. The pattern suggests that while the summer 2020 movement achieved broad geographic reach across partisan lines, sustained large-scale mobilization remained concentrated in more liberal, urban areas.

5.2.2 Attitude Change: Measuring Support for Police Funding

We measure attitudes toward police spending using the Cooperative Election Study (CES), a nationally representative post-election survey conducted biennially from 2014 to 2024

Figure 3: Geography of Summer 2020 Protests



Notes: This map shows counties with at least one protest between May 25 and July 31, 2020. Dots size is proportional to the number of protests in each county. Background shading indicates county-level partisan lean, measured as the Republican minus Democratic two-party vote share in the 2016 presidential election, where darker red indicates more Republican-leaning counties and darker blue indicates more Democratic-leaning counties.

(Kuriwaki, 2023; Schaffner et al., 2023). CES provides detailed information on political views, demographics, and the responses to state legislative decisions on spending in sectors such as law enforcement (Mazumder, 2019). These measures allow us to directly examine whether the "defund the police" movement and the 2020 protests against police brutality changed attitudes towards police funding. We excluded respondents with missing location information.

We supplement the CES data with county-level demographic, economic, political, and crime characteristics. Demographic and socioeconomic measures are constructed from the 2019 American Community Survey (ACS). Crime rates and police staffing are drawn from the FBI Uniform Crime Reports (UCR) and the Law Enforcement Employment (LEE) data. Electoral outcomes come from the MIT Election Data and Science Lab county presidential returns, and measures of police killings are obtained from the Mapping Police Violence database.

To capture exposure to protests, we link CES respondents and election data to the GDELT protest data with residential ZIP codes. ZIP codes are mapped to counties and merged with county-level records of police brutality–related protests occurring between May 25 and July 31, 2020. We drop respondents with missing ZIP codes, spending

preferences, or demographic information. The final analytic sample contains 284,139 respondents, including 161,808 or about 57% residing in counties that experienced related protests.

5.2.3 Policy Response: Measuring Adoption of Police Alternatives

To measure institutional responses to the protests, we construct a county-level panel tracking the adoption of crisis response programs—police alternatives that dispatch mental health professionals, social workers, or paramedics to nonviolent emergency calls that would otherwise receive a traditional police response. We compile program-level data from multiple sources, including the *Crisis Response Services Directory*, media reports, and municipal records, identifying programs operating across the US. For each program, we record the year of adoption, the institutional home (community-based organization, fire/EMS, or police department), and whether police officers are involved in the response. We construct a balanced county$\times$year panel from 2016 to 2023 covering all 2,882 counties in our sample. We link this panel to the same GDELT protest data used in the survey analysis. County-level covariates—demographics, income, crime rates, police staffing, and partisan vote shares—are drawn from the 2019 ACS, FBI UCR, and MIT Election Lab, as described above.

5.3 Short and Long-Term Impacts of Protests

5.3.1 Descriptives

Table 1 presents summary statistics for CES respondents in counties with and without protests following George Floyd's murder. The analytic sample contains 284,139 respondents: 161,808 (57%) reside in counties that experienced at least one protest and 122,331 in counties with none. Panel A reports mean support for spending across three sectors. Support for police funding is lower in protest counties (0.85 vs. 0.88), whereas support for education (0.92 vs. 0.91) and infrastructure funding (0.95 in both groups) is nearly identical. This divergence—a gap for police but not for the placebo sectors—motivates our triple-difference design, which differences out compositional differences between protest and nonprotest counties that affect spending preferences uniformly.

Panel B documents the individual-level composition of each group. Relative to nonprotest counties, respondents in protest counties are more racially diverse, younger, more educated, higher-income, less likely to be married, and substantially more likely to have

voted Democratic (49% vs. 39%). These systematic differences underscore that protest exposure is not random and motivate our use of county and year fixed effects to absorb time-invariant compositional differences across groups.

Table 1: Summary Statistics: CES Respondents

	(1) All	(2) No Protest	(3) Any Protest
Panel A: Outcome			
Support for Police Funding	0.86	0.88	0.85
Support for Education Funding	0.92	0.91	0.92
Support for Infrastructure Funding	0.95	0.95	0.95
Panel B: Individual Characteristics			
Male	0.46	0.44	0.46
Black	0.09	0.07	0.11
Hispanic	0.07	0.05	0.08
Other Race	0.08	0.06	0.09
39–54yo	0.26	0.26	0.26
55–64yo	0.25	0.26	0.24
64yo+ or Missing	0.24	0.25	0.23
Married	0.53	0.56	0.51
Employed	0.49	0.46	0.52
Some College	0.22	0.23	0.22
College Grad	0.37	0.34	0.38
Post-Grad	0.15	0.13	0.17
40K–70K	0.24	0.25	0.23
70K–150K	0.23	0.22	0.24
150K+ or Missing	0.24	0.20	0.26
Last Vote: Democrat	0.45	0.39	0.49
Last Vote: Republican	0.35	0.40	0.31
Observations	284,139	122,331	161,808

Notes: This table presents mean characteristics of CES respondents across the full sample (Column 1), counties with no protests following George Floyd's murder (Column 2), and counties with any protests (Column 3). Support variables equal one if the respondent favors maintaining or increasing funding in the given sector. Individual characteristics are from CES survey responses. The sample pools CES waves from 2014 to 2024.

5.3.2 Research Design

DID Design Our objective is to assess how the 2020 protests influenced attitudes toward police funding and voting behavior. We estimate the following canonical DID model (Roth

et al., 2023; de Chaisemartin and D'Haultfoeuille, 2025):

$$y_{ict} = \sum_t D_t \cdot Protest_c \cdot \beta_t + X_i'\delta + \alpha_c + \gamma_t + \epsilon_{ict}, \tag{16}$$

where y_{ict} denotes the outcome for respondent i in county c and year t. The indicator $Protest_c$ equals one for counties that experienced at least one protest during summer 2020 and zero otherwise. This indicator is interacted with year dummies D_t, with 2018 serving as the reference year. The vector X_i includes respondent-level characteristics in Table 1. County fixed effects α_c absorb time-invariant county characteristics, and year fixed effects γ_t capture common shocks. Standard errors are clustered at the county level.

The key coefficients, β_t, capture the average treatment effect on the treated (ATT) over time and trace how outcomes evolve over time in counties with protests, relative to 2018 and benchmarked against counties with no protest activity. Identification relies on the standard parallel trends assumption: In the absence of the 2020 protests, counties that later experienced protests or no protests would have followed comparable pre-2020 trends in support for police funding. This requires that protest-exposed counties were not already on systematically different trajectories prior to 2020—an assumption we assess using the pretreatment waves from 2014, 2016, and 2018. We also define the average treatment effect post-2020 as $ATT = \sum_{t \geq 2020} \beta_t / 3$.[5]

Triple-Difference Design While equation 16 compares protest and nonprotest counties over time, protest exposure is not random and differs with observables. Counties that experienced protests may also have faced distinct time-varying shocks in 2020–2024—pandemic-related disruption, local fiscal stress, or shifts in public-safety salience—that altered attitudes toward government spending more broadly, biasing the DID estimates.

To address this, we implement a triple-difference (DDD) design exploiting the fact that the CES asks parallel questions about state legislative spending across multiple domains. The DDD compares the protest-induced change in support for police funding to the contemporaneous change in support for "placebo" domains—education and infrastructure—that are not direct targets of the *defund the police* demand. We estimate:

$$y_{ickt} = \sum_t D_t \cdot Protest_c \cdot Police_k \cdot \beta_t + X_i'\delta + \alpha_{ck} + \gamma_{tk} + \theta_{ct} + \epsilon_{ickt}, \tag{17}$$

where $k \in \{\text{police}, \text{education}, \text{infrastructure}\}$ indexes the spending domain, $Police_k$ indi-

[5] Our analysis has three waves post-treatment t: 2020, 2022, and 2024.

cates the police spending outcome, and all other variables are defined as in (16). County–domain fixed effects α_{ck} absorb time-invariant differences in baseline support across domains; year–domain fixed effects γ_{tk} capture national trends that differ by domain (e.g., shifting views on education spending); and county–year fixed effects θ_{ct} absorb any county-specific annual shocks affecting attitudes across all domains (e.g., local pandemic intensity or fiscal conditions). Standard errors are clustered at the county level.

The coefficients β_t capture the differential effect of protests on police funding support relative to education and infrastructure. Identification requires a "relative" parallel trends assumption: that absent the 2020 protests, the police-versus-placebo gap would have evolved comparably in protest and nonprotest counties. We assess this using the pre-2020 waves. We also define the average treatment effect post-2020 as $ATT = \sum_{t \geq 2020} \beta_t / 3$.

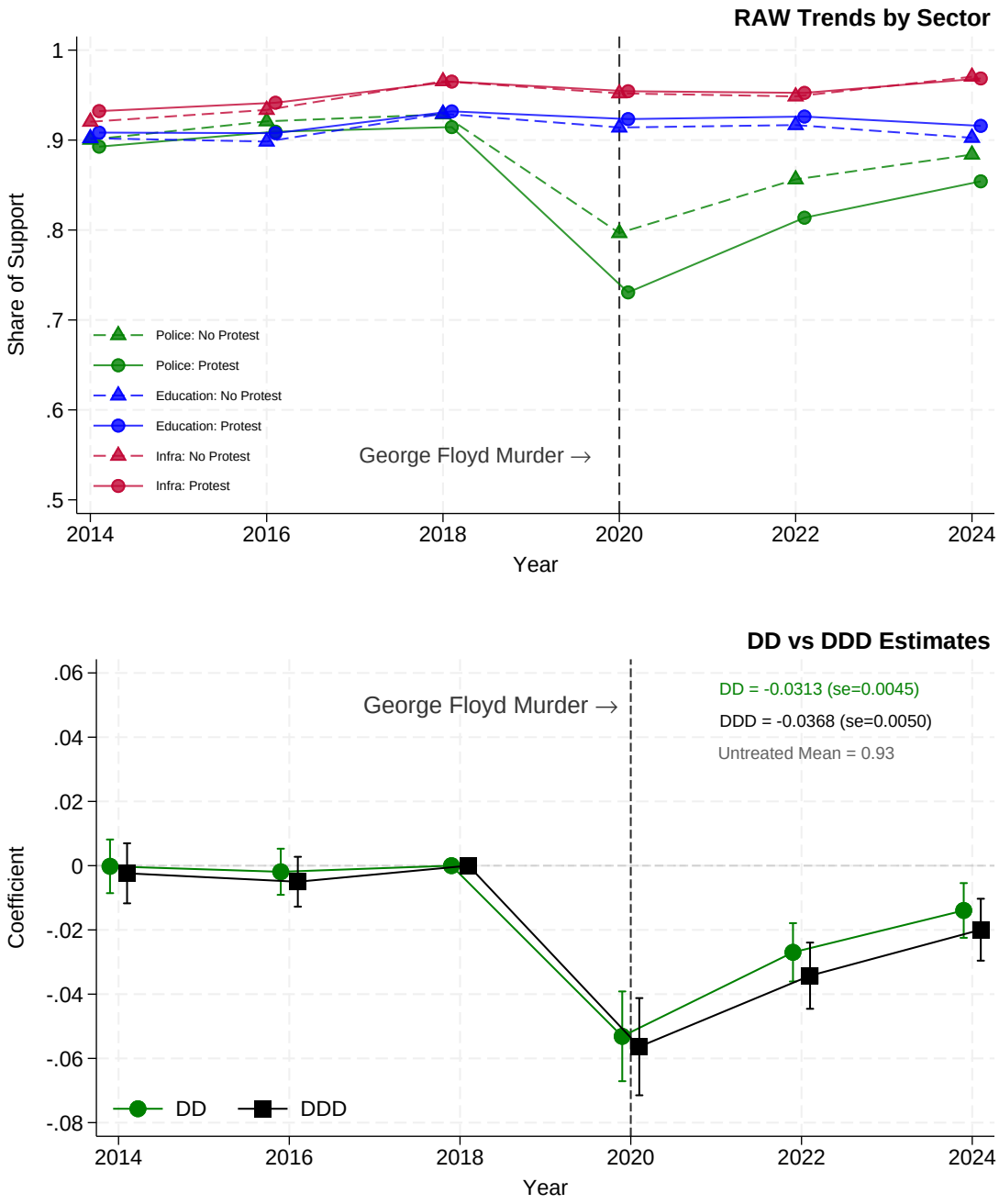
5.3.3 Results: Support for Police Funding

Main Results Figure 4 reveals a sizable but transitory shift on an otherwise stable margin. Average support for maintaining or increasing police funding remains high even in protest counties, yet the 2020 break is clear: The DID and DDD event-study estimates imply a decline of roughly 3–4 pp, relative to a preperiod mean near 0.93. The protests thus moved a nontrivial share of respondents away from "maintain or increase" despite strong baseline favorability toward police budgets. By 2022 and 2024, however, the coefficients shrink appreciably, suggesting that much of the initial persuasion dissipated within one to two survey waves.

This pattern carries two implications. First, the magnitude reflects meaningful short-run persuasion on a highly sticky policy margin—a decline is large enough to register politically when baseline support sits near 0.93. Second, the rapid attenuation indicates little durable preference change; support never approached a new long-run equilibrium. Coupled with the near-zero differential movement in education and infrastructure spending preferences, the results point to a protest-specific shock to attitudes about policing that faded rather than a broad reorientation of fiscal priorities.

Partisan Heterogeneity Figure 5 reveals that the protest-induced attitudinal change was concentrated entirely among Democratic voters. Among Democrats, the raw series show a pronounced national decline in support for maintaining or increasing police funding in 2020, with the drop systematically larger in protest counties than in nonprotest counties. The event-study estimates confirm a statistically significant additional decline of 3–4 pp

Figure 4: Effect of Protests on Support for Police Funding



Notes: This figure presents the impact of the BLM protests following George Floyd's murder on support for maintaining or increasing police funding. The top panel plots raw means by year and protest status. The bottom panel reports DID and DDD event-study coefficients with 95% confidence intervals; standard errors are clustered by county. Vertical line marks May 25, 2020. We report the mean outcome for nonprotest counties in the reference year.

in protest counties at the time of George Floyd's murder, followed by partial reversion in subsequent waves. The average post-treatment effect remains negative and significant, with an untreated mean of 0.90.

Among Republican voters, the protests had no discernible effect. Baseline support for police funding is near a ceiling (untreated mean $= 0.97$), and both the raw trends and event-study coefficients are flat across all post-treatment waves. Neither education nor infrastructure spending exhibits a comparable protest-related divergence for either partisan group, reinforcing the interpretation that the effect operated specifically through attitudes about policing rather than through broad shifts in public spending preferences.

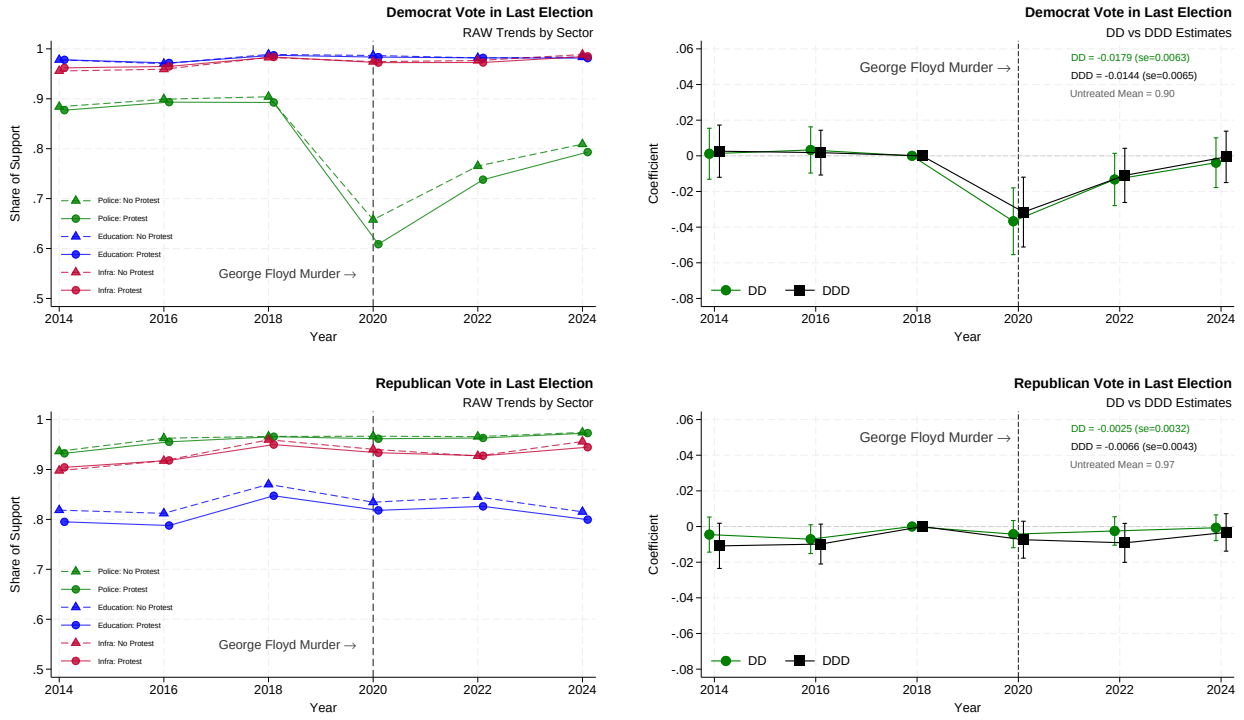
The aggregate effect was thus driven entirely by movement within the Democratic coalition, with no detectable cross-partisan persuasion. This asymmetry is consistent with polarization in underlying policy priors: protests shift attitudes where baseline openness to reform is already relatively high but fail to move voters whose priors strongly favor the status quo.

Racial Heterogeneity Figure 6 explores heterogeneity by respondent race and ethnicity. The protest effect on support for police funding is largest and most precisely estimated among White respondents. The raw series show a clear divergence between protest and nonprotest counties beginning in 2020, and the event-study estimates indicate a significant decline of 4–6 pp at the time of George Floyd's murder, with partial recovery thereafter. The average post-treatment effect is sizable and statistically significant relative to an untreated mean of 0.93.

Among Black respondents, the point estimates are negative but imprecisely estimated, with wide confidence intervals that span zero in most post-treatment waves. The raw trends suggest a modest protest-related divergence, but the smaller sample sizes limit statistical power. Hispanic respondents show no meaningful protest effect: The event-study coefficients fluctuate around zero both before and after 2020, and the average post-treatment estimates are small and insignificant.

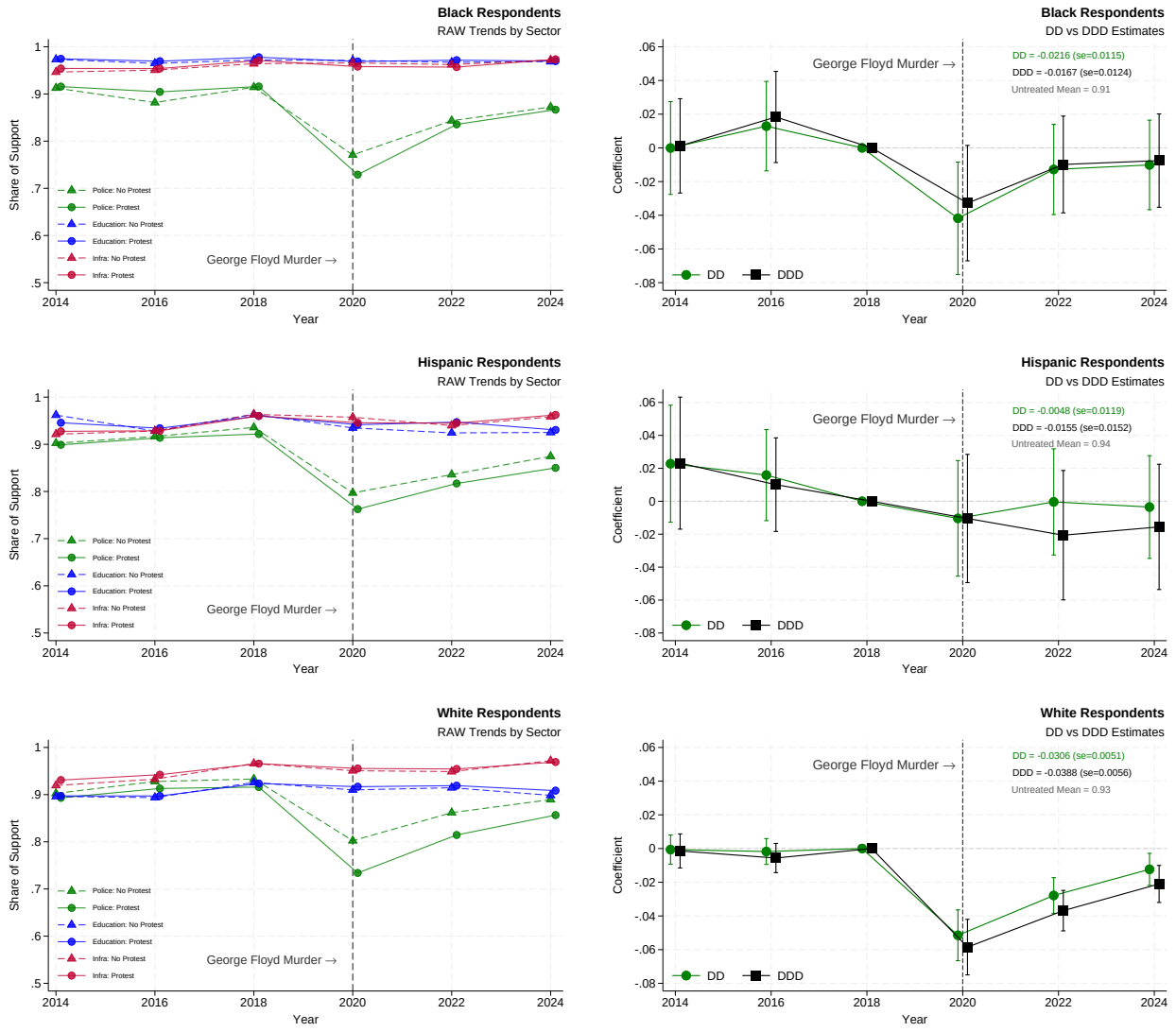
The finding that White respondents drive the aggregate result is consistent with the partisan heterogeneity documented above, given the overlap between White and Democratic subsamples in this context. It also suggests that the protests operated primarily by shifting the attitudes of respondents whose baseline views were closer to the policy margin—supportive of police funding but potentially persuadable—rather than by reinforcing existing preferences among communities most directly affected by policing.

Figure 5: Heterogeneity by Political Affiliation: Support for Police Funding



Notes: This figure presents the impact of the BLM protests following George Floyd's murder on support for maintaining or increasing police funding, split by respondents' self-reported past presidential vote (Democratic vs. Republican). The left column plots raw means by year and protest status. The right column reports DID and DDD event-study coefficients with 95% confidence intervals; standard errors are clustered by county. Vertical line marks May 25, 2020. We report the mean outcome for nonprotest counties in the reference year.

Figure 6: Heterogeneity by Race and Ethnicity: Support for Police Funding



Notes: This figure examines the impact of the BLM protests following George Floyd's murder on support for maintaining or increasing police funding, by respondent race and ethnicity (Black, Hispanic, and White). The left column plots raw means by year and protest status. The right column reports DID and DDD event-study coefficients with 95% confidence intervals; standard errors are clustered by county. Vertical line marks May 25, 2020. We report the mean outcome for nonprotest counties in the reference year.

5.4 Impact of Protests on Policy Adoption

5.4.1 Descriptives

Table 2 presents county-level summary statistics for the adoption sample, which contains 2,882 counties observed annually from 2016 to 2023. Of these, 410 (14%) experienced at least one protest following George Floyd's murder. Panel A reports the outcome: By 2023, 12% of protest counties had adopted at least one crisis response program, compared with only 1% of nonprotest counties. Panel B documents county characteristics. Protest counties are substantially larger (375,000 vs. 53,000 residents), denser, wealthier, more racially diverse, and more Democratic-leaning (52% vs. 66% Republican vote share). They also have higher violent crime rates and more police per capita. As with the survey data, these compositional differences are absorbed by county fixed effects in the estimation.

Table 2: Summary Statistics: County-Level Adoption Sample

	(1) All	(2) No Protest	(3) Any Protest
Panel A: Outcome			
Has Crisis Response Program (2023)	0.02	0.01	0.12
Panel B: County Characteristics			
Population (1,000)	98.46	52.55	375.27
Share White	0.77	0.78	0.70
Share Black	0.09	0.09	0.11
Share Hispanic	0.10	0.09	0.12
Population Density	216.82	125.34	768.40
Share Unemployed	0.05	0.05	0.05
Median HH Income ($1,000)	52.89	51.46	61.53
Share Republican Votes	0.64	0.66	0.52
Share Democrat Votes	0.31	0.29	0.42
Violent Crime Rate (per 1,000)	2.25	2.12	3.03
Police per 1,000	1.85	1.82	2.05
Counties	2,882	2,472	410

Notes: This table presents mean county characteristics across the full sample (Column 1), counties with no protests following George Floyd's murder (Column 2), and counties with any protests (Column 3). The outcome variable equals one if the county has adopted at least one crisis response program by 2023.

Figure 7 plots the share of counties with a crisis response program by year, separately for protest and nonprotest counties. Before 2020, adoption rates were low and nearly identical in the two groups. For years after 2020, the two series diverge sharply: Among

protest counties, the share with a program rose from near zero to approximately 13% by 2023, while adoption in nonprotest counties remains close to zero. This pattern contrasts with the attitudinal effects documented in the previous section, where the protest-induced changes in support for police funding largely dissipated within one to two survey waves. Here, adoption continued to grow through the end of the sample period, suggesting that the institutional response triggered by the protests did not revert.

5.4.2 Research Design

We estimate the effect of protest exposure on the adoption of alternative police programs using a DID design similar to that in the previous section. The unit of observation is a county–year. The outcome y_{ct} equals one if the county c has adopted at least one crisis response program by year t and zero otherwise. The estimation equation is

$$y_{ct} = \sum_t D_t \cdot Protest_c \cdot \beta_t + X'_{ct}\,\delta + \alpha_c + \gamma_t + \epsilon_{ct}, \tag{18}$$

where $Protest_c$ equals one for counties that experienced at least one protest between May 25 and July 31 of 2020, α_c denotes fixed effects for the county, and γ_t denotes fixed effects for the year. The vector X_{ct} includes controls at the county level that vary in time. Standard errors are clustered at the county level. The coefficients β_t indicate the evolution of adoption in protest counties relative to nonprotest counties, with 2019 as the reference year. The average post-treatment effect is defined as $ATT = \sum_{t\geq 2020} \beta_t / 4$, averaging over the four post-treatment years 2020–2023. Identification requires that, absent the 2020 protests, adoption would have followed parallel trends in protest and nonprotest counties. The preperiod estimates from 2016 to 2019 provide a direct test of this assumption. As documented in the bottom panel of Figure 7, adoption rates were low and nearly identical between the two groups before 2020, consistent with the identification assumption.

5.4.3 Results: Adoption of Police Alternative Programs

Main Results Figure 7 presents the results. The pretreatment coefficients are small and statistically insignificant, consistent with parallel pretrends. For the years after 2020, the estimates increase monotonically: Protest exposure raises the probability of adoption of a crisis response program by roughly 1 pp in 2020; this grows to approximately 11 pp for 2023 ($p < 0.01$). The average post-treatment effect is 5.9 pp ($SE = 0.009$), relative to a preperiod baseline near zero in both groups.

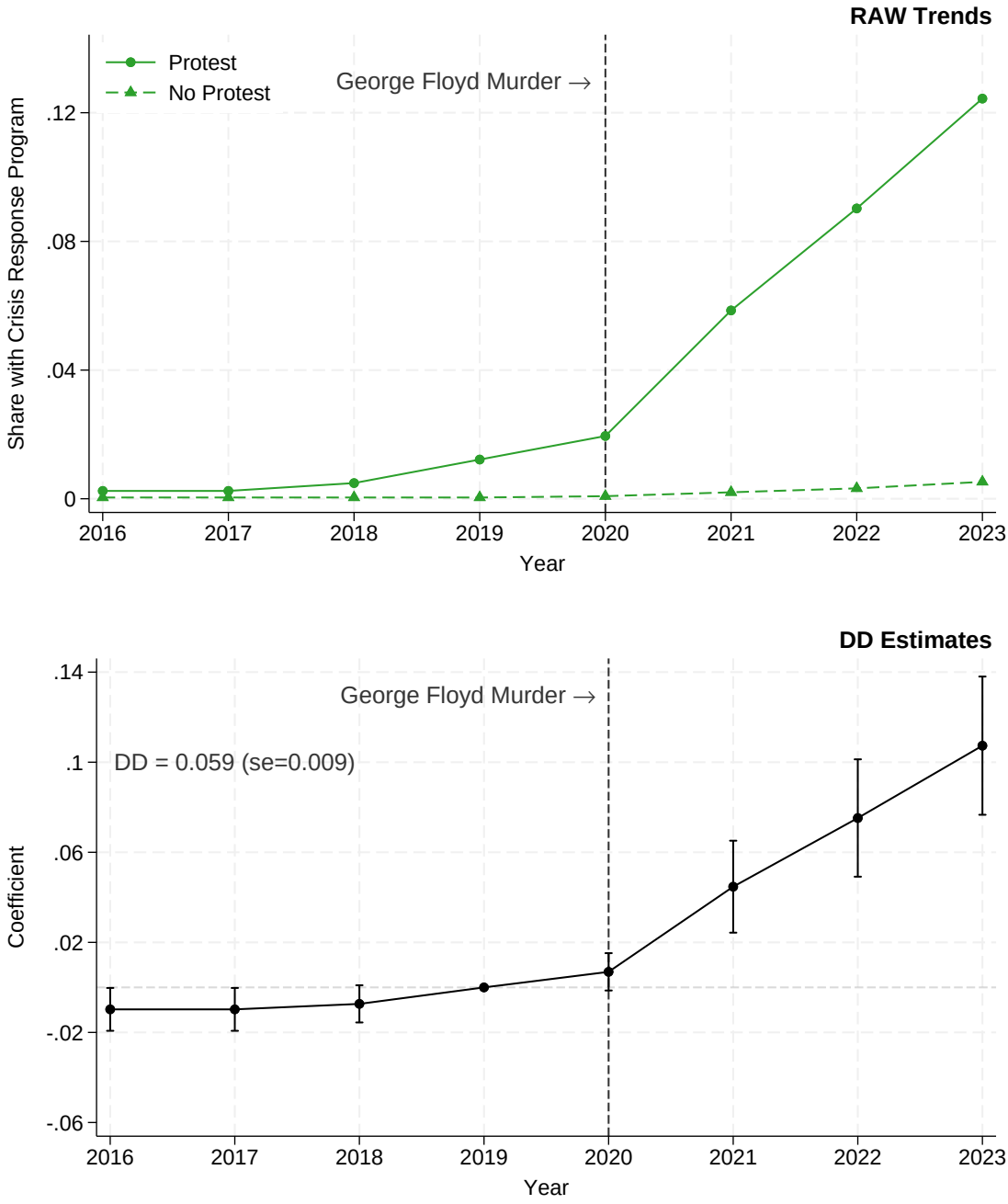
This pattern contrasts sharply with the attitudinal effects documented in the previous section. Whereas the protest-induced changes in support for police funding dissipated within one to two survey waves, institutional adoption has not only persisted but compounded over time. The results suggest that the protests triggered an ongoing process of program creation rather than a one-time adjustment.

Program Home The top-left panel of Figure 8 splits adoption by the institutional home of the program: community-based, fire/EMS team–based, or police department–based. All three types show flat preperiod trends, consistent with parallel trends. After 2020, adoption increases across all categories, but the effects differ in magnitude. The largest average post-treatment effect is for community-based programs, at 3.0 pp ($SE = 0.007$), followed by fire/EMS team–based programs at 2.5 pp ($SE = 0.006$) and police department–based programs at 1.4 pp ($SE = 0.005$). The ordering suggests that the protests disproportionately induced the adoption of programs that operate outside or alongside the police department, rather than of programs housed within it.

Who Responds? The top-right panel of Figure 8 divides the programs by whether they involve police in the response. The contrast is stark. Programs without police involvement account for the bulk of the aggregate effect, with an average post-treatment increase of 5.0 pp ($SE = 0.009$) that continues to grow through 2023. Programs with police involvement show a much smaller and flatter response, at 0.9 pp ($SE = 0.003$). This pattern indicates that protest-induced adoption was concentrated in programs representing a genuine structural departure from the traditional police response, rather than reforms that leave the institutional role of the police intact.

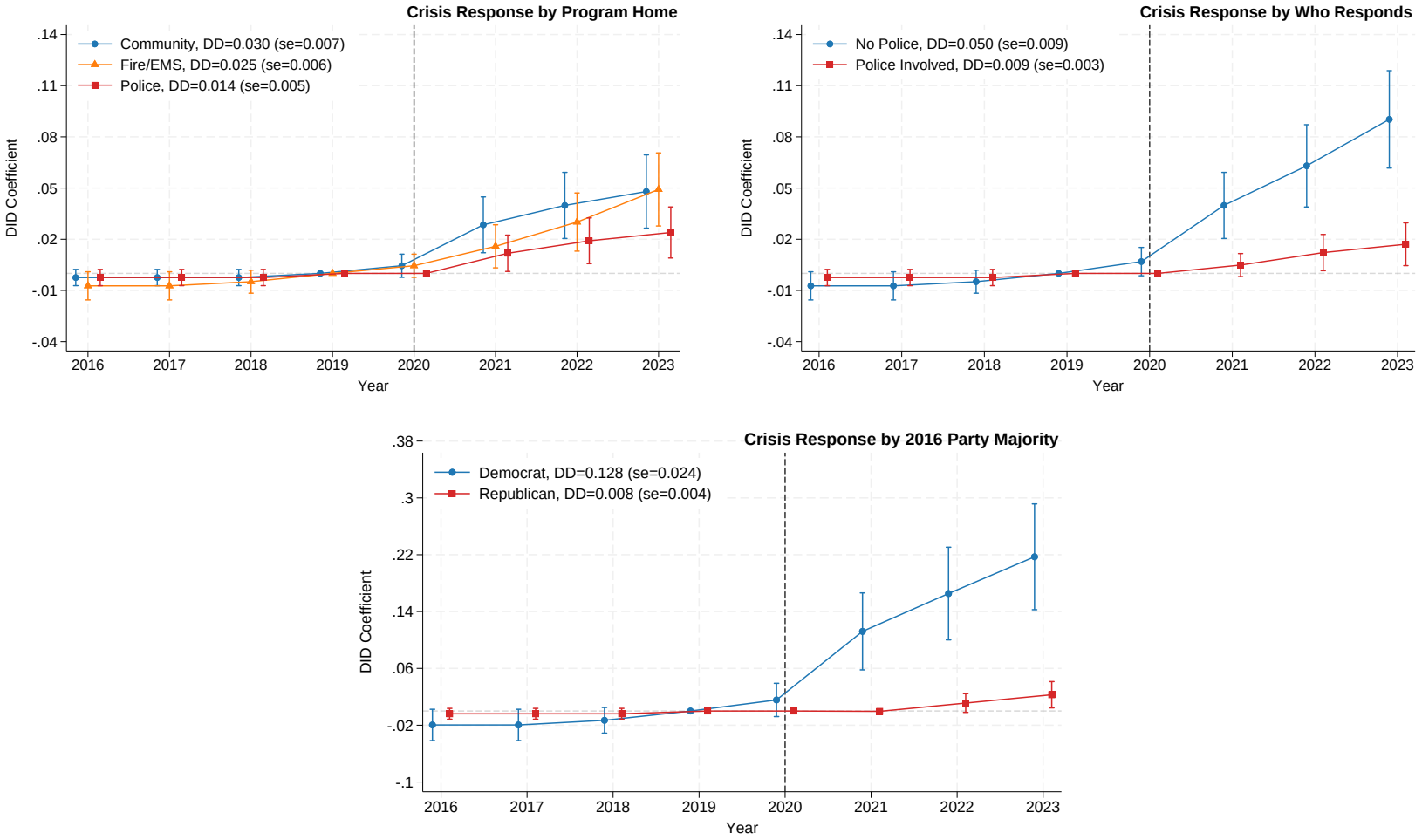
Role of Partisanship The bottom panel of Figure 8 splits counties by their 2016 presidential vote majority. The partisan asymmetry is large. For Democratic-majority counties, protest exposure raises the probability of adoption by 12.8 pp ($SE = 0.024$), with the coefficient rising steeply from near zero for 2020 to approximately 22 pp for 2023. In Republican-majority counties, the effect is small and precisely estimated near zero for all years throughout the post-period, at 0.8 pp ($SE = 0.004$). The contrast mirrors the partisan heterogeneity in the attitudinal effects documented in the previous section: Protest-induced persuasion translated into institutional adoption primarily in political environments where initial openness to reform was already relatively high.

Figure 7: Effect of Protests on the Adoption of Police Alternatives Programs



Notes: This figure presents the impact of the BLM protests following George Floyd's murder on the likelihood of local governments adopting police alternatives programs. The top panel plots raw adoption rates by year and protest exposure status. The bottom panel reports DID event-study coefficients with 95% confidence intervals; standard errors are clustered by county. Vertical line marks 2020, the year of George Floyd's murder. The sample includes 2,882 counties observed from 2016 to 2023.

Figure 8: Who Adopts Police Alternatives?
Heterogeneity by Program Type and County Partisanship



Notes: This figure presents heterogeneity in the effect of BLM protests on the adoption of police alternatives programs. The top-left panel splits programs by institutional home (community-based, fire/EMS team, or police department). The top-right panel splits by whether police are involved in the response. The bottom panel splits counties by majority 2016 presidential vote share. All panels report DID event-study coefficients with 95% confidence intervals; standard errors are clustered by county. Vertical line marks 2020, the year of George Floyd's murder. The sample includes 2,882 counties observed from 2016 to 2023.

6 Conclusion

This paper develops a model of protest-based persuasion in which a salience shock expands the feasible set of belief distributions an activist can induce. The key insight is that persuasion and institutional change operate on different thresholds. Attitude shifts require only that salience remain high enough to sustain attention; institutional adoption requires that aggregate support cross a political threshold at which policymakers act. When salience fades, attitudes revert — but institutions, once created, do not. The model further predicts that persuasion is asymmetric under polarization: protests move opinion among voters whose priors are closer to the reform margin while leaving opponents largely unaffected. Together, these forces generate a distinctive pattern — transitory persuasion on the primary demand, concentrated among sympathetic partisans, alongside durable institutional responses on adjacent margins where the political threshold is lower.

The empirical application confirms each prediction. The 2020 protests following George Floyd's murder reduced support for police funding in exposed counties, but the effect was concentrated among Democrats, had no detectable impact on Republicans, and largely dissipated within two survey waves. In response to the primary movement demand, the protests achieved short-term persuasion without durable attitudinal change. However, on an adjacent policy margin, the response was qualitatively different: exposed counties were significantly more likely to adopt a police alternative program, and unlike the attitudinal effects, this institutional response compounded over time. Adoption was concentrated in civilian-only, community-based models and overwhelmingly in Democratic-leaning jurisdictions.

The broader lesson is that the political legacy of mass protest may depend less on whether it durably changes minds than on whether it pushes support past the threshold at which institutions act. Where that threshold is high—as with defunding the police, which requires sustained majority support on a deeply polarized margin—even the largest protest movement in American history fell short. Where it is low—as with police alternatives, which do not require budget cuts or partisan consensus—transitory persuasionwas sufficient to trigger institutional adoption that persisted and grew long after public attention had moved on.

References

Acemoglu, D., T. A. Hassan, and A. Tahoun (2018). The power of the street: Evidence from egypt's arab spring. *The Review of Financial Studies 31*(1), 1–42.

Adena, M., R. Enikolopov, M. Petrova, V. Santarosa, and E. Zhuravskaya (2015). Radio and the rise of the Nazis in prewar Germany. *Quarterly Journal of Economics 130*(4), 1885–1939.

Alesina, A. and N. Fuchs-Schündeln (2007). Good-bye lenin (or not?): The effect of communism on people's preferences. *American Economic Review 97*(4), 1507–1528.

Allcott, H. and M. Gentzkow (2017). Social media and fake news in the 2016 election. *Journal of economic perspectives 31*(2), 211–236.

Ba, B. A., P. Chen, T. Cheng, M. C. Eies, and J. E. Holz (2025). What is the best response? examining the impact of police and their alternatives. Working Paper 34344, National Bureau of Economic Research. NBER Working Paper 34344.

Bergemann, D. and S. Morris (2019). Information design: A unified perspective. *Journal of Economic Literature 57*(1), 44–95.

Boudreau, C., S. A. MacKenzie, and D. 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.

Buchanan, L., Q. Bui, J. K. Patel, B. Saget, and A. Singhvi (2020, July). Black lives matter may be the largest movement in U.S. history.

Bursztyn, L., D. Cantoni, D. Y. Yang, N. Yuchtman, and Y. J. Zhang (2021). Persistent political engagement: Social interactions and the dynamics of protest movements. *American Economic Review: Insights 3*(2), 181–198.

Cantoni, D., D. Y. Yang, N. Yuchtman, and Y. J. Zhang (2019). Protests as strategic games: Experimental evidence from Hong Kong's antiauthoritarian movement. *The Quarterly Journal of Economics 134*(2), 1021–1077.

de Chaisemartin, C. and X. D'Haultfoeuille (2025). Credible answers to hard questions: Differences-in-differences for natural experiments.

Dee, T. S. and J. Pyne (2022). A community response approach to mental health and substance abuse crises reduced crime. *Science advances 8*(23), eabm2106.

DellaVigna, S. and M. Gentzkow (2010). Persuasion: Empirical evidence. *Annual Review of Economics 2*(1), 643–669.

DellaVigna, S. and E. Kaplan (2007). The Fox News effect: Media bias and voting. *Quarterly Journal of Economics 122*(3), 1187–1234.

Derenoncourt, E. and C. Montialoux (2021). Minimum wages and racial inequality. *The Quarterly Journal of Economics 136*(1), 169–228.

Ebbinghaus, M., N. Bailey, and J. Rubel (2025). The effect of the 2020 Black Lives Matter protests on police budgets: How "defund the police" sparked political backlash. *Social Problems 72*(3), 1198–1215.

Edmond, C. (2013). Information manipulation, coordination, and regime change. *Review of Economic studies 80*(4), 1422–1458.

Enikolopov, R., A. Makarin, and M. Petrova (2020). Social media and protest participation: Evidence from Russia. *Econometrica 88*(4), 1479–1514.

Enikolopov, R., M. Petrova, and E. Zhuravskaya (2011). Media and political persuasion: Evidence from Russia. *American Economic Review 101*(7), 3253–3285.

Francis, M. M. and L. Wright-Rigueur (2021). Black lives matter in historical perspective. *Annual Review of Law and Social Science 17*, 441–458.

Gentzkow, M. and E. Kamenica (2016). A rothschild-stiglitz approach to bayesian persuasion. *American Economic Review 106*(5), 597–601.

Gentzkow, M. and J. M. Shapiro (2011). Ideological segregation online and offline. *The Quarterly Journal of Economics 126*(4), 1799–1839.

Kamenica, E. and M. Gentzkow (2011, October). Bayesian persuasion. *American Economic Review 101*(6), 2590–2615.

Knight, D. J. and V. M. Weaver (2025). Black political mobilization and the US carceral state: How tracing community struggles for safety changes the policing narrative. *Annual Review of Criminology 8*(1), 25–52.

Kuriwaki, S. (2023). Cumulative CES Common Content.

Kuziemko, I. and E. Washington (2018). Why did the Democrats lose the South? Bringing new data to an old debate. *American Economic Review 108*(10), 2830–2867.

Levy, R. and M. Mattsson (2025). The effects of social movements: Evidence from #MeToo. CEPR Discussion Paper No. 20319.

Lipnowski, E. and D. Ravid (2020). Cheap talk with transparent motives. *Econometrica 88*(4), 1631–1660.

Madestam, A., D. Shoag, S. Veuger, and D. Yanagizawa-Drott (2013). Do political protests matter? Evidence from the Tea Party movement. *The Quarterly Journal of Economics 128*(4), 1633–1685.

Malmendier, U. and S. Nagel (2011). Depression babies: do macroeconomic experiences affect risk taking? *The quarterly journal of economics 126*(1), 373–416.

Mazumder, S. (2019, May). Black lives matter for whites' racial prejudice: Assessing the role of social movements in shaping racial attitudes in the united states.

Phelps, M. S. (2024). *The Minneapolis Reckoning: Race, Violence, and the Politics of Policing in America*. Princeton University Press.

Roth, J., P. H. Sant'Anna, A. Bilinski, and J. Poe (2023). What's trending in difference-in-differences? a synthesis of the recent econometrics literature. *Journal of Econometrics 235*(2), 2218–2244.

Schaffner, B., S. Ansolabehere, and M. Shih (2023). Cooperative Election Study Common Content, 2022.

Shadmehr, M. and D. Bernhardt (2019). Vanguards in revolution. *Games and Economic Behavior 115*, 146–166.

Vaughn, P. E., K. Peyton, and G. A. Huber (2022). Mass support for proposals to reshape policing depends on the implications for crime and safety. *Criminology & Public Policy*.

Wasow, O. (2020). Agenda seeding: How 1960s black protests moved elites, public opinion and voting. *American Political Science Review 114*(3), 638–659.

A Mathematical Proofs

Proof of lemma 1. Recall that $\Delta([0,1])$ denotes the set of Borel probability measures on $[0,1]$, endowed with the weak topology, and that

$$\mathcal{N}(s) = \Big\{\nu \in \Delta([0,1]) \,:\, \int_{[0,1]} \mu\, \nu(d\mu) = \mu_0 \text{ and } \int_{[0,1]} \mathrm{KL}(\mu\|\mu_0)\, \nu(d\mu) \leq s\Big\}.$$

(i) Nonemptiness Consider the degenerate probability distribution (Dirac measure) $\nu = \delta_{\mu_0}$. This distribution trivially satisfies Bayes plausibility since $\int_{[0,1]} \mu\, \delta_{\mu_0}(d\mu) = \mu_0$. Moreover, because the Kullback–Leibler divergence vanishes at equality, $\int_{[0,1]} \mathrm{KL}(\mu\|\mu_0)\, \delta_{\mu_0}(d\mu) = \mathrm{KL}(\mu_0\|\mu_0) = 0 \leq s$. Hence $\delta_{\mu_0} \in \mathcal{N}(s)$, and the set is nonempty.

(ii) Convexity Let $\nu_1, \nu_2 \in \mathcal{N}(s)$ and let $\lambda \in [0,1]$. Define the convex combination $\nu = \lambda\nu_1 + (1-\lambda)\nu_2$. Since integration is linear in the underlying measure, both constraints defining $\mathcal{N}(s)$ are preserved under such mixtures. Thus $\nu \in \mathcal{N}(s)$, which shows that $\mathcal{N}(s)$ is convex.

(iii) Weak compactness Since the state space $[0,1]$ is compact and metrizable, Prokhorov's theorem implies that the set $\Delta([0,1])$ of probability measures on $[0,1]$ is compact for the weak topology. It therefore suffices to show that $\mathcal{N}(s)$ is weakly closed. Let $(\nu_n)_{n\geq 1}$ be a sequence in $\mathcal{N}(s)$ that converges weakly to some $\nu \in \Delta([0,1])$.

We first verify that Bayes plausibility is preserved in the limit. The function $\mu \mapsto \mu$ is continuous and bounded on $[0,1]$, hence weak convergence implies $\int \mu\, d\nu_n \longrightarrow \int \mu\, d\nu$. Since $\int \mu\, d\nu_n = \mu_0$ for all n, we obtain $\int \mu\, d\nu = \mu_0$.

Additionally, because $\mu_0 \in (0,1)$, by Eq. 4, the function $\mu \mapsto \mathrm{KL}(\mu\|\mu_0)$ is continuous on $[0,1]$. Moreover, continuity on a compact domain implies boundedness. Weak convergence therefore yields $\int \mathrm{KL}(\mu\|\mu_0)\, d\nu_n \longrightarrow \int \mathrm{KL}(\mu\|\mu_0)\, d\nu$. Since each ν_n satisfies the inequality $\int \mathrm{KL}(\mu\|\mu_0)\, d\nu_n \leq s$, passing to the limit shows that the same inequality holds for ν. Consequently, $\nu \in \mathcal{N}(s)$, and $\mathcal{N}(s)$ is weakly closed.

Combining closedness with compactness of $\Delta([0,1])$ yields weak compactness of $\mathcal{N}(s)$.

(iv) Monotonicity Let $s \leq s'$. The Bayes plausibility constraint does not depend on s, and the inequality $\int \mathrm{KL}(\mu\|\mu_0)\, d\nu \leq s$ implies $\int \mathrm{KL}(\mu\|\mu_0)\, d\nu \leq s'$. Hence $\mathcal{N}(s) \subseteq \mathcal{N}(s')$, and the correspondence is increasing.

(v) Closed graph Finally, consider a sequence (s_n, ν_n) such that $s_n \to s$ and $\nu_n \Rightarrow \nu$, with $\nu_n \in \mathcal{N}(s_n)$. By the same arguments as above, Bayes plausibility passes to the limit:

$\int \mu \, d\nu = \mu_0$. Moreover,

$$\int \mathrm{KL}(\mu \| \mu_0) \, d\nu = \lim_{n\to\infty} \int \mathrm{KL}(\mu \| \mu_0) \, d\nu_n \leq \lim_{n\to\infty} s_n = s.$$

Thus $\nu \in \mathcal{N}(s)$, which shows that the graph of the correspondence $s \mapsto \mathcal{N}(s)$ is closed. □

Proof of Proposition 1. We work throughout with the weak topology on $\Delta([0,1])$. Let $\mathcal{B}$ denote the space of bounded continuous functions on $\mathbb{R}_+$, endowed with the supremum norm $\|\cdot\|_\infty$.

Define the Bellman operator $T : \mathcal{B} \to \mathcal{B}$ by

$$(TW)(s) = \max_{\nu \in \mathcal{N}(s)} \left\{ u(s,\nu) + \beta W\big(\delta s + \eta f(P(\nu))\big) \right\},$$

where the per-period payoff is given by

$$u(s,\nu) = R(s)\, P(\nu) - C\left(\int_0^1 \mathrm{KL}(\mu \| \mu_0) \, d\nu(\mu) \right).$$

We first verify that $TW(s)$ is well defined and finite for every $s \geq 0$. Since R is bounded and $P(\nu) \in [0,1]$, we have $|R(s)P(\nu)| \leq \|R\|_\infty$. Moreover, $C(\cdot) \geq 0$ and, by definition of $\mathcal{N}(s)$, $\int_0^1 \mathrm{KL}(\mu \| \mu_0) \, d\nu(\mu) \leq s$, so the argument of C is bounded on any bounded subset of the state space. Finally, W is bounded by assumption. It follows that the objective function inside the maximization is bounded above uniformly over $\nu \in \mathcal{N}(s)$, and hence $(TW)(s)$ is finite for every s.

Step 1: Existence of a maximizer Fix $s \geq 0$. By Lemma 1, the feasible set $\mathcal{N}(s)$ is nonempty and compact in the weak topology. The mapping $\nu \mapsto P(\nu)$ is continuous under weak convergence, since $P(\nu) = \int_0^1 G(\mu) \, d\nu(\mu)$ and G is continuous and bounded. Similarly, because $\mu_0 \in (0,1)$, the function $\mu \mapsto \mathrm{KL}(\mu \| \mu_0)$ is continuous and bounded on $[0,1]$, which implies continuity of the mapping $\nu \mapsto \int \mathrm{KL} \, d\nu$.

Since W is continuous and f is continuous by assumption, the composite mapping $\nu \longmapsto u(s,\nu) + \beta W\big(\delta s + \eta f(P(\nu))\big)$ is continuous on $\mathcal{N}(s)$. Therefore, by the Weierstrass theorem, the maximum is attained.

Step 2: Continuity of the Bellman operator We now show that T maps $\mathcal{B}$ into itself. Let $W \in \mathcal{B}$. By Step 2, the value $(TW)(s)$ is the maximum of a continuous function over a compact set. Moreover, Lemma 1 implies that the correspondence $s \mapsto \mathcal{N}(s)$ has a

closed graph and is increasing. Berge's maximum theorem therefore implies that TW is continuous in s. Since TW is also bounded, we conclude that $TW \in \mathcal{B}$.

Step 3: Contraction property Let $W_1, W_2 \in \mathcal{B}$ and fix $s \geq 0$. For any $\nu \in \mathcal{N}(s)$,

$$\left|u(s,\nu) + \beta W_1(\delta s + \eta f(P(\nu))) - u(s,\nu) - \beta W_2(\delta s + \eta f(P(\nu)))\right| \leq \beta \|W_1 - W_2\|_\infty.$$

Taking the supremum over $\nu \in \mathcal{N}(s)$ yields $|(TW_1)(s) - (TW_2)(s)| \leq \beta \|W_1 - W_2\|_\infty$. Since this bound holds uniformly in s, we obtain $\|TW_1 - TW_2\|_\infty \leq \beta \|W_1 - W_2\|_\infty$. Thus T is a contraction on $(\mathcal{B}, \|\cdot\|_\infty)$.

Step 4: Existence of equilibrium Because $(\mathcal{B}, \|\cdot\|_\infty)$ is a complete metric space and T is a contraction, Banach's fixed point theorem implies that T admits a unique fixed point $V \in \mathcal{B}$.

Finally, for each s, the argmax correspondence associated with $(TW)(s)$ is nonempty by Step 2. Standard measurable selection results apply here, since the objective function is continuous and the feasible correspondence is compact-valued with a closed graph. Hence, there exists a measurable policy $\nu^*(s) \in \mathcal{N}(s)$ attaining the maximum. The pair (V, ν^*) therefore solves the Bellman equation (11) and defines a stationary Markov equilibrium. □

Proof of Lemma 3. Fix $s \geq 0$ and consider the intra-period problem (12). Write the constraints defining $\mathcal{N}(s)$ as

$$\int_0^1 \mu \, d\nu(\mu) = \mu_0 \qquad \text{and} \qquad I(\nu) \equiv \int_0^1 (\mu \| \mu_0) \, d\nu(\mu) \leq s.$$

Denote the objective functional by

$$J(\nu) \equiv \int_0^1 \tilde{v}_s(\mu) \, d\nu(\mu) - C(I(\nu)).$$

By Lemma 1, $\mathcal{N}(s)$ is nonempty and weakly compact. Since $\tilde{v}_s$ is continuous and bounded on $[0,1]$, the mapping $\nu \mapsto \int \tilde{v}_s \, d\nu$ is continuous for weak convergence. Moreover, because $\mu_0 \in (0,1)$, the mapping $\nu \mapsto I(\nu)$ is also continuous (since $\mathrm{KL}(\cdot \| \mu_0)$ is continuous and bounded on $[0,1]$), and therefore $\nu \mapsto C(I(\nu))$ is continuous as well. Hence J is continuous on the compact set $\mathcal{N}(s)$, and a maximizer ν^* exists.

Finally, note that $\nu \mapsto I(\nu)$ is linear in ν and C is convex, so $-C(I(\nu))$ is concave in ν; hence J is concave over the convex feasible set. This guarantees that first-order (KKT)

conditions are valid and economically interpretable.

Next consider the Lagrangian associated with the inequality constraint $I(\nu) \leq s$:

$$\mathcal{L}(\nu, \kappa) = \int_0^1 \tilde{v}_s(\mu)\, d\nu(\mu) - C(I(\nu)) - \kappa\,(I(\nu) - s),$$

where $\kappa \geq 0$ is the mutiplier, and Bayes plausibility $\int \mu\, d\nu = \mu_0$ is kept as an explicit constraint.

Because $\delta_{\mu_0} \in \mathcal{N}(s)$ and satisfies $I(\delta_{\mu_0}) = 0 < s$ whenever $s > 0$, Slater's condition holds for $s > 0$. Thus there exists a KKT multiplier $\kappa \geq 0$ satisfying complementary slackness $\kappa\,(I^* - s) = 0$, where $I^* \equiv I(\nu^*)$.[6]

Finally, fix the optimal intensity I^* and define the effective marginal cost $\lambda \equiv C'(I^*) + \kappa \geq 0$. We claim that ν^* maximizes the linearized problem (14) with this λ.

To see this, take any ν satisfying Bayes plausibility $\int \mu\, d\nu = \mu_0$. By convexity of C and differentiability at I^*, it follows that

$$C(I(\nu)) \;\geq\; C(I^*) + C'(I^*)\,(I(\nu) - I^*).$$

Rearranging yields the global upper bound

$$-C(I(\nu)) \;\leq\; -C(I^*) - C'(I^*)\,(I(\nu) - I^*).$$

Adding $-\kappa(I(\nu) - s)$ to both sides and using $I^* \leq s$ and complementary slackness gives

$$-C(I(\nu)) - \kappa(I(\nu) - s) \;\leq\; -C(I^*) - \kappa(I^* - s) - (C'(I^*) + \kappa)\,(I(\nu) - I^*) = \text{const} - \lambda\, I(\nu),$$

where the constant does not depend on ν. Therefore, for any Bayes-plausible ν,

$$\mathcal{L}(\nu, \kappa) = \int \tilde{v}_s\, d\nu - C(I(\nu)) - \kappa(I(\nu) - s) \;\leq\; \text{const} + \int \tilde{v}_s\, d\nu - \lambda\, I(\nu).$$

At $\nu = \nu^*$, the convexity bound is tight at I^*, and complementary slackness implies $\mathcal{L}(\nu^*, \kappa) = \max_{\nu \in \mathcal{N}(s)} \mathcal{L}(\nu, \kappa)$. Hence ν^* must maximize $\int \tilde{v}_s\, d\nu - \lambda I(\nu)$ among Bayes-plausible ν, which is exactly (14).

This proves the lemma, with $\lambda = C'(I^*) + \kappa$ and $\kappa\,(I^* - s) = 0$. □

Proof of Proposition 2. Fix (s, λ) and define $\phi \equiv \phi_{s,\lambda}$ as in (15). Consider the (static)

[6] Note that when $s = 0$, feasibility forces $I(\nu) = 0$ and the statement is immediate; one may take any $\kappa \geq 0$

persuasion program with a single Bayes-plausibility constraint:

$$\sup_{\nu \in \Delta([0,1])} \left\{ \int_0^1 \phi(\mu)\, d\nu(\mu) \;:\; \int_0^1 \mu\, d\nu(\mu) = \mu_0 \right\}. \tag{19}$$

We still work on the compact belief space $[0,1]$ and endow $\Delta([0,1])$ with the weak topology.

Step 1: Definition of the concave envelope and an upper bound Define the *upper concave envelope* (a.k.a. concavification) of ϕ by

$$\operatorname{cav}\phi(\mu) \;\equiv\; \inf\left\{ \ell(\mu) \;:\; \ell \text{ affine on } [0,1],\ \ell(\mu') \geq \phi(\mu')\ \forall \mu' \in [0,1] \right\}.$$

Equivalently, $\operatorname{cav}\phi$ is the smallest concave function on $[0,1]$ that dominates ϕ pointwise

Let $\nu \in \Delta([0,1])$ satisfy $\int \mu\, d\nu = \mu_0$. Since $\operatorname{cav}\phi$ is concave, Jensen's inequality implies

$$\int_0^1 \operatorname{cav}\phi(\mu)\, d\nu(\mu) \;\leq\; \operatorname{cav}\phi\left(\int_0^1 \mu\, d\nu(\mu) \right) = \operatorname{cav}\phi(\mu_0).$$

Moreover, since $\operatorname{cav}\phi \geq \phi$ pointwise, we have $\int \phi\, d\nu \leq \int \operatorname{cav}\phi\, d\nu$. Combining yields the universal upper bound,

$$\int_0^1 \phi(\mu)\, d\nu(\mu) \;\leq\; \operatorname{cav}\phi(\mu_0), \tag{20}$$

for every feasible ν.

We now claim that the upper bound in (20) is tight, i.e., the value of (19) equals $\operatorname{cav}\phi(\mu_0)$. This is the standard concavification result in Bayesian persuasion; see Kamenica and Gentzkow (2011, Theorem 1) or Gentzkow and Kamenica (2016/2017, concavification characterization).[7] In particular, there exists an optimal ν^* such that

$$\int_0^1 \phi(\mu)\, d\nu^*(\mu) = \operatorname{cav}\phi(\mu_0). \tag{21}$$

Existence of a supporting line at μ_0 Since $[0,1]$ is compact and $\operatorname{cav}\phi$ is concave and finite-valued, it admits a supporting line at each interior point. Because $\mu_0 \in (0,1)$, there exists

[7] When the belief space is one-dimensional, one may further restrict attention to finite-support solutions by Carathéodory's theorem, as in Kamenica and Gentzkow (2011); we do not need this restriction for the supporting-line argument below

a subgradient $b \in \partial \operatorname{cav} \phi(\mu_0)$ such that

$$\ell(\mu) \equiv \operatorname{cav} \phi(\mu_0) + b(\mu - \mu_0) \quad \text{satisfies} \quad \ell(\mu) \geq \operatorname{cav} \phi(\mu) \ \forall \mu \in [0,1], \text{ and } \ell(\mu_0) = \operatorname{cav} \phi(\mu_0). \tag{22}$$

This is a standard supporting-hyperplane property for concave functions (e.g. Rockafellar, *Convex Analysis*, Corollary 23.4.1). Since $\operatorname{cav} \phi \geq \phi$, (22) implies $\ell(\mu) \geq \phi(\mu)$ for all μ, establishing part (1).

Step 4: Contact on the support of an optimal design Let ν^* be an optimal solution. From $\ell \geq \phi$ we have $\int \ell \, d\nu^* \geq \int \phi \, d\nu^*$. On the other hand, ℓ is affine, and ν^* satisfies Bayes plausibility, hence

$$\int_0^1 \ell(\mu) \, d\nu^*(\mu) = \ell\left(\int_0^1 \mu \, d\nu^*(\mu) \right) = \ell(\mu_0) = \operatorname{cav} \phi(\mu_0).$$

Combining this with (21) yields

$$\int_0^1 \ell(\mu) \, d\nu^*(\mu) = \int_0^1 \phi(\mu) \, d\nu^*(\mu).$$

Therefore,

$$\int_0^1 \big(\ell(\mu) - \phi(\mu)\big) \, d\nu^*(\mu) = 0, \qquad \text{with} \qquad \ell(\mu) - \phi(\mu) \geq 0 \ \ \forall \mu.$$

It follows that $\ell(\mu) = \phi(\mu)$ ν^*-almost surely, i.e., for every μ in the *essential support* of ν^*. In particular, all posteriors induced with strictly positive probability by ν^* belong to the contact set $\{\mu \in [0,1] : \ell(\mu) = \phi(\mu)\}$, establishing part (2).

This is precisely the common-supporting-line characterization of optimal posteriors in the concavification approach to Bayesian persuasion. □

Proof of Proposition 4. Fix $\delta \in (0,1)$ and $\eta > 0$. Define

$$\Phi(s) = \delta s + \eta f(P(\nu^*(s))), \qquad \overline{\Phi}(s) = \delta s + \eta f(\overline{P}(s)),$$

where $\overline{P}$ is continuous and weakly increasing, and f is continuous and weakly increasing with $f(p) = 0$ for $p \leq \bar{P} - \varepsilon$ and $f(p) = 1$ for $p \geq \bar{P} + \varepsilon$. Since $P(\nu^*(s)) \leq \overline{P}(s)$ and f is increasing, we have for all $s \geq 0$,

$$\Phi(s) \leq \overline{\Phi}(s). \tag{23}$$

We prove each regime in turn.

(1) Transitory Regime Assume $\sup_{s\geq 0} \overline{P}(s) \leq \bar{P} - \varepsilon$. Then $\overline{P}(s) \leq \bar{P} - \varepsilon$ for all s, hence $f(\overline{P}(s)) = 0$ for all s and

$$\overline{\Phi}(s) = \delta s.$$

In particular, for every $s > 0$, $\overline{\Phi}(s) = \delta s < s$. Combining with (23) yields $\Phi(s) \leq \delta s < s$ for all $s > 0$. Therefore the only fixed point is $s^\star = 0$.

Moreover, for any initial condition $s_0 \geq 0$, the sequence satisfies

$$0 \leq s_{t+1} = \Phi(s_t) \leq \delta s_t \leq \delta^{t+1} s_0,$$

so $s_t \to 0$ (indeed geometrically). This proves global convergence to zero.

(2) Persistent Regime Assume there exists $\bar{s} > 0$ such that $\overline{P}(\bar{s}) \geq \bar{P} + \varepsilon$ and $\delta\bar{s} + \eta \geq \bar{s}$. Since $f(p) = 1$ for $p \geq \bar{P} + \varepsilon$, we have

$$\overline{\Phi}(\bar{s}) = \delta\bar{s} + \eta \geq \bar{s}.$$

Also $\overline{\Phi}(0) = 0$. Consider the continuous function

$$H(s) \equiv \overline{\Phi}(s) - s.$$

We have $H(0) = 0$ and $H(\bar{s}) \geq 0$. Under the maintained regularity (continuity of $\overline{\Phi}$), there exists at least one fixed point $s^* > 0$ with $H(s^*) = 0$ provided that H is strictly negative for some $s \in (0, \bar{s})$; a sufficient (and economically natural) condition is $\overline{\Phi}'(0) < 1$, which holds whenever either $\overline{P}(0) \leq \bar{P} - \varepsilon$ (so $f' = 0$ near 0) or more generally because the feedback term is inactive far below the threshold.[8]

Local stability in the saturated region is immediate. Whenever $\overline{P}(s) \geq \bar{P} + \varepsilon$, we have $f(\overline{P}(s)) = 1$ and hence $\overline{\Phi}(s) = \delta s + \eta$, which is affine with slope $\delta < 1$. Thus any fixed point satisfying $\overline{P}(s^*) \geq \bar{P} + \varepsilon$ is locally stable.

Finally, by (23), if the equilibrium policy attains (or sufficiently approximates) the frontier at high salience, then there exist equilibrium paths for which reinforcement remains active and salience stays bounded away from zero. This establishes persistence.

[8] If $H(s) \geq 0$ for all $s \in [0, \bar{s}]$, then every $s \in [0, \bar{s}]$ satisfies $\overline{\Phi}(s) \geq s$, and any smallest $s > 0$ with equality is a fixed point

(3) Tipping Regime (Multiplicity) Assume $\sup_{s\geq 0} \overline{P}(s) \geq \bar{P} + \varepsilon$. By continuity and monotonicity of $\overline{P}$, there exists at least one $s^\dagger$ such that $\overline{P}(s^\dagger) \in (\bar{P} - \varepsilon, \bar{P} + \varepsilon)$ (the transition region).

Define again $H(s) = \overline{\Phi}(s) - s$. Then H is continuous. We claim that under the stated slope condition, H must cross zero at least three times.

First, note that for sufficiently small s, the feedback is inactive and $\overline{\Phi}'(s) = \delta < 1$ (because $f'(\cdot) = 0$ off the transition region). Hence for small $s > 0$,

$$H(s) = \overline{\Phi}(s) - s < 0,$$

i.e., the graph of $\overline{\Phi}$ lies below the 45-degree line near the origin.

Second, at $s^\dagger$ we assume $\overline{P}$ is differentiable and

$$\eta f'(\overline{P}(s^\dagger))\overline{P}'(s^\dagger) > 1 - \delta,$$

so $\overline{\Phi}'(s^\dagger) = \delta + \eta f'(\overline{P}(s^\dagger))\overline{P}'(s^\dagger) > 1$, i.e.,

$$H'(s^\dagger) = \overline{\Phi}'(s^\dagger) - 1 > 0.$$

Thus H is locally increasing through the transition region and, generically, must become positive somewhere after having been negative near 0. Concretely, since H is continuous and increases with slope exceeding one at some point, there exists $s_1 > 0$ with $H(s_1) > 0$.

Third, for sufficiently large s, reinforcement is saturated ($f(\overline{P}(s)) = 1$ once $\overline{P}(s) \geq \bar{P} + \varepsilon$), so $\overline{\Phi}(s) = \delta s + \eta$ and hence

$$H(s) = \delta s + \eta - s = -(1 - \delta)s + \eta \to -\infty \quad \text{as } s \to \infty.$$

Therefore there exists s_2 large such that $H(s_2) < 0$.

We have shown: (i) $H(s) < 0$ for s small, (ii) $H(s) > 0$ for some intermediate s_1, and (iii) $H(s) < 0$ for large s_2. By the intermediate value theorem, there exist at least two distinct roots $s_L < s_H$ with $H(s_L) = H(s_H) = 0$ and $s_L < s_1 < s_H$. Moreover, since $H(0) = 0$ and $H(s) < 0$ for small $s > 0$, there is a smallest positive fixed point $s_L > 0$. Because H changes sign from negative to positive and later from positive to negative, H must cross zero at least three times generically; denote the three fixed points by $0 \leq s_L < s_M < s_H$.

Stability follows from the sign of $H'(s) = \overline{\Phi}'(s) - 1$ at each crossing: at the lowest and

highest crossings the graph crosses the 45-degree line from above to below (slope < 1), yielding local stability, whereas at the middle crossing it crosses from below to above (slope > 1), yielding local instability. Therefore s_M acts as a threshold: initial conditions below s_M converge to the low-salience attractor, while initial conditions above converge to the high-salience attractor.

Finally, equilibrium dynamics Φ inherit the same qualitative behavior whenever the equilibrium policy attains (or closely tracks) the frontier in the relevant regions; at minimum, (23) implies that tipping in $\overline{\Phi}$ creates the possibility of tipping in equilibrium dynamics and delivers the comparative-static logic emphasized in the text. □