When Outsider Rhetoric Goes Mainstream: Political Competition and the Credibility Trap

André Quintas*
New York University

Erwin Dekker
George Mason University

August 2026

Abstract

We develop a theory of institutional trust erosion in which anti-establishment rhetoric serves as a signaling device in political competition. Politicians differ in their willingness to sacrifice establishment relationships and use anti-establishment attacks to signal this privately known type. As institutional trust declines, voters place greater value on perceived independence from the establishment. Outsiders initially separate by escalating their rhetoric, but once distrust becomes sufficiently severe, insiders imitate and anti-establishment rhetoric becomes mainstream. Mainstreaming increases the supply of anti-establishment rhetoric while diluting its informational content. Embedding this signaling equilibrium in a dynamic model of institutional trust, we show how rhetoric and distrust can reinforce one another, generating hysteresis and persistent low-trust equilibria. The model identifies when temporary trust shocks produce lasting credibility crises and when institutional reforms can restore trust.

Acknowledgments and disclosures. The authors received no financial support for this research and have no relevant financial relationships or other conflicts of interest.

Keywords: anti-establishment rhetoric; political outsiders; signaling; political trust; party competition; hysteresis

*Corresponding author. Email: andre.quintas@nyu.edu.

1 Introduction

Anti-establishment rhetoric has become a prominent feature of democratic politics. Outsider candidates have long used attacks on "the system," "the elites," or "the political class" to distinguish themselves from established parties and institutions. Donald Trump promised to "drain the swamp"; Javier Milei denounced *la casta* as entrenched in the state and determined to defend its privileges. Yet similar rhetoric is increasingly adopted by politicians who themselves occupy mainstream or incumbent positions. In 2023, after thirteen years of Conservative government and while serving as prime minister, Rishi Sunak described Keir Starmer as "the walking definition of the 30-year political status quo I am here to end."[1]

This diffusion raises a puzzle. If anti-establishment rhetoric is valuable because it communicates distance from the establishment, why would establishment politicians adopt it? And if they do, what happens to the credibility of the signal and to trust in the institutions being attacked?

We develop a signaling model in which politicians differ in their willingness to sacrifice valuable establishment relationships and use anti-establishment rhetoric as a signaling device. Anti-establishment rhetoric is therefore less costly for outsiders than for insiders. The electoral return to anti-establishment rhetoric rises as institutional trust declines because voters place greater value on candidates who appear willing to break with the status quo. Politicians therefore compete over perceptions of outsider status. This competition initially encourages differentiation but eventually induces imitation, until anti-establishment rhetoric loses its capacity to distinguish political types. Distrust therefore first intensifies the outsider signal and then brings about its dilution.

We embed this signaling equilibrium in a dynamic model of institutional trust. We show that political competition can generate a credibility trap: a feedback loop in which anti-establishment rhetoric erodes institutional trust, and declining trust increases the electoral return to anti-establishment rhetoric. The result is hysteresis in institutional trust—temporary shocks can produce persistent distrust.

The model points to a collective-action problem in the production of institutional trust when trust begins to decline: each actor's optimal strategy undermines the credibility environment that all actors depend on. Once both outsiders and insiders attack mainstream institutions, any politician who moderates is immediately perceived as an insider precisely when establishment attachment has become an electoral liability. No individual politician can restore trust through unilateral moderation: the politician bears the electoral cost of restraint, while the resulting improvement in institutional trust is dispersed across the political system. Political competition can therefore sustain an aggregate level of attack that perpetuates distrust even after the rhetoric has lost its capacity to distinguish political types.

We nonetheless identify paths to recovery. Because unilateral moderation is electorally costly, recovery requires an exogenous improvement in the credibility environment—such as visible institutional performance, credible countervailing information, or coordinated restraint that reduces aggregate attack. With a sufficiently positive shock, incentives reverse: the electoral premium on perceived detachment declines, insiders withdraw from anti-establishment rhetoric, outsiders moderate their attacks, and aggregate rhetoric falls. Trust then recovers further, making the high-trust equilibrium self-reinforcing.

Our first contribution is to the political economy of populism (for broader overviews, see Mudde and Rovira Kaltwasser, 2017; Guriev and Papaioannou, 2022). Research on voter demand links

[1] The quotations are from Trump (2016), Milei (2025), and Sunak (2023).

globalization and economic insecurity to support for populist alternatives (Rodrik, 2021; Guiso et al., 2024), with declining trust in established parties operating as one channel (Guiso et al., 2024). Complementary work shows that the supply of populist rhetoric is strategic: outsider candidates intensify it when electorates are economically insecure and elections are competitive (Gennaro et al., 2024), parties use it more extensively when in opposition (Breyer, 2023), and its use diffuses across party systems (Surdea-Hernea, 2025). More broadly, mainstream parties adjust their strategies in response to competition from niche challengers (Meguid, 2005). We add an equilibrium mechanism connecting these patterns by showing how declining institutional trust raises the electoral return to perceived outsider status. Mainstreaming of anti-establishment rhetoric then emerges as a trust-dependent equilibrium response to political competition rather than as an exogenous process of diffusion, and this shift interacts with—and helps shape—the credibility environment in which that competition unfolds.

Second, the paper contributes to research on the dynamics of political trust (Citrin and Stoker, 2018). Existing models show how narrative competition or the interaction between political deception and voter expectations can generate persistent distrust (Herrera and Trombetta, 2024; Glynia, Manalis, and Xefteris, 2025). We identify a complementary source of persistence: the mainstreaming of anti-establishment rhetoric. Although such rhetoric is commonly associated with political outsiders, our model identifies its adoption by insiders as the turning point at which a temporary shock in trust can become a persistent credibility crisis.

Third, the paper contributes to theories of electoral signaling. Existing models show how costly campaign positions can reveal candidates' privately known policy intentions or character (Banks, 1990; Kartik and McAfee, 2007). Research on populism applies this logic to political independence: politicians may adopt populist policies or provocative conduct to signal that they are not captured by powerful elites or attached to the political establishment (Acemoglu, Egorov, and Sonin, 2013; Gonnot and Seabright, 2021). Our key move is to analyze the institutional effects of the signal. Anti-establishment rhetoric conveys information about candidate type, while its aggregate use erodes institutional trust. Because lower trust increases the electoral return to perceived outsider status, the signal creates the conditions for its own diffusion and, ultimately, the loss of its capacity to distinguish types.

The remainder of the paper proceeds as follows. Section 2 presents the signaling model. Sections 3 and 4 derive outsider differentiation and characterize the threshold at which insiders mainstream anti-establishment rhetoric. Section 5 embeds the equilibrium supply of attack in the evolution of institutional trust and establishes the conditions for a persistent low-trust equilibrium. Section 6 develops the model's comparative statics and observable implications. Section 7 discusses its scope conditions and implications for preventing or reversing a credibility crisis. Section 8 concludes.

2 Model

We model anti-establishment rhetoric as a costly signal of a politician's willingness to break with the status quo (Banks, 1990; Kartik and McAfee, 2007). Politicians privately know how willing they are to sacrifice establishment ties. Voters infer this disposition from the intensity of their anti-establishment attacks.

The model has two layers. In the period game, let $\mu_t \in [0, 1]$ denote a normalized measure of institutional trust inherited from the previous period and publicly observed, with higher values

indicating greater trust.[2] Nature draws the candidate's privately observed type and publicly observed policy position. The candidate then chooses the intensity, a, of an anti-establishment attack. Voters observe a, update their beliefs about the candidate's type, and choose a candidate. Across a continuum of contests, these equilibrium attack choices generate the aggregate supply of rhetoric, $A(\mu_t)$, where the politicians' individual choice of attack depends on the degree to which institutional trust has already been eroded.

The dynamic layer maps this aggregate supply into institutional trust in the following period. Because inherited trust changes the electoral value of appearing detached from the establishment, equilibrium attack—and therefore aggregate rhetoric—varies with μ_t:

$$\mu_t \Longrightarrow \text{signaling equilibrium at } \mu_t \Longrightarrow A(\mu_t) \Longrightarrow \mu_{t+1}.$$

The model takes as given an exogenous negative shock to institutional trust, such as a financial crisis, political scandal, or institutional failure. The shock lowers inherited trust and changes the electoral environment in which politicians compete. When trust is maximal, $\mu = 1$, perceived willingness to break with the status quo carries no electoral advantage, so both types choose zero anti-establishment attack.

As trust begins to fall, anti-establishment rhetoric emerges gradually. Outsiders initially choose attacks of low intensity. Further declines in trust increase the electoral value of perceived detachment and make imitation more attractive to insiders. Outsiders therefore intensify their anti-establishment attacks so that voters can continue to distinguish them from insiders, choosing at each level of trust an intensity high enough to make imitation unprofitable. This process continues until outsiders reach the maximum feasible intensity of anti-establishment rhetoric. At that point, adjustment shifts from the intensity of outsider attack to the composition of political speakers: outsiders remain at that maximum, while the share of insiders who imitate them rises continuously. Aggregate rhetoric therefore continues to grow even though outsiders can no longer escalate. As insider participation expands, attack becomes progressively less informative. If the electoral premium from being perceived as an outsider is sufficiently large —as formalized in Proposition 3—this process culminates in full pooling, in which both types attack and voters can no longer infer political type from the signal.

The equilibrium response to declining trust thus unfolds along two margins. Outsiders first increase the intensity of their attacks while insiders remain quiet. Once outsiders exhaust this margin of differentiation, further declines in trust progressively change the composition of political speakers, as summarized in Table 1.

[2]For example, μ_t may be interpreted as the share of citizens expressing trust in political institutions, although the model only requires a normalized aggregate index.

Table 1: How political rhetoric changes as institutional trust declines

Trust environment	Outsiders	Insiders	What voters learn
Maximum trust	Remain quiet	Remain quiet	Anti-establishment attack is absent
Trust begins to decline	Gradually intensify their attacks	Remain quiet	An attack identifies an outsider
Outsiders reach their signaling limit	Continue attacking at maximum intensity	Progressively begin to imitate	Attack becomes progressively less informative
Severe distrust	Continue attacking at maximum intensity	All imitate	Attack no longer distinguishes political types

Note: The table previews the equilibrium sequence. The imitation and full-pooling thresholds are defined formally below.

In the dynamic stage, institutional trust is shaped by two opposing forces: gradual recovery toward a baseline and erosion generated by aggregate anti-establishment rhetoric. Following a negative shock, institutional trust tends to recover toward its baseline. Repeated anti-establishment rhetoric works against this recovery. At the aggregate level, repeated anti-establishment attacks do more than signal candidate type: they also cast the political system as corrupt, captured, or broken. We therefore allow anti-establishment rhetoric produced in the current period to reduce trust in the next. Whether trust recovers depends on whether equilibrium attack remains limited or becomes strong enough to outweigh recovery. Under the conditions derived in Section 5, outsider attack alone cannot sustain a low-trust equilibrium, so trust eventually recovers. Insider entry into anti-establishment rhetoric changes this by increasing both the level of aggregate attack and its responsiveness to further declines in trust. Mainstreaming anti-establishment rhetoric can therefore create a stable low-trust equilibrium. A sufficiently strong exogenous shock can then push trust toward that equilibrium, making the initial decline self-sustaining.

Figure 1 previews this distinction. It compares the recovery path generated by outsider attack alone with the path that becomes possible once insiders enter and amplify aggregate rhetoric.

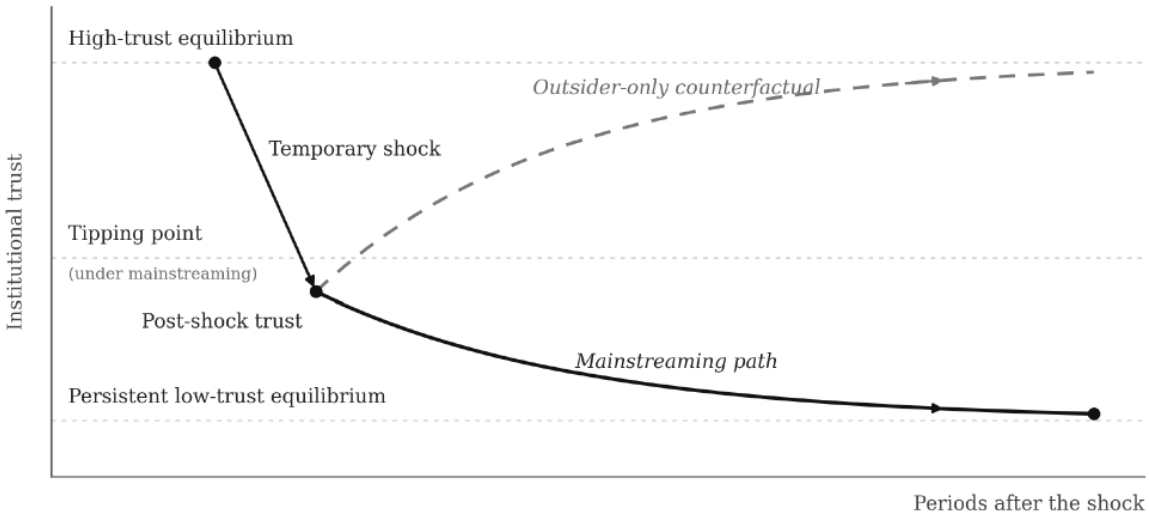


Figure 1. Mainstreaming changes the long-run response to a temporary trust shock.
The same temporary shock produces different long-run outcomes. With outsider attack alone, aggregate attack remains bounded and trust recovers (dashed path). With mainstreaming, insider entry expands aggregate attack, so trust continues to fall after the shock disappears and converges to a persistent low-trust equilibrium (solid path).

2.1 Candidates, types, and anti-establishment attack

Each electoral contest pits a candidate against a fixed benchmark alternative, denoted B.[3] Nature draws the candidate's privately observed type $\theta \in \{O, I\}$. With probability $\lambda \in (0, 1)$, the candidate is an outsider type O; otherwise, the candidate is an insider type I. Types differ in their willingness to sacrifice establishment ties in order to break with the status quo. We call the type more willing to sacrifice those ties an outsider and the type less willing to do so an insider. Observable career histories may inform the prior λ, but they do not fully reveal how willing a candidate will be to break with established arrangements once in office.

Nature also draws a publicly observed policy position x, independently of the candidate's type.[4] The representative voter has ideal point d, making the candidate's policy distance $q = (x - d)^2$. Policy distance determines the candidate's baseline electoral attractiveness, while beliefs about type determine the additional value voters attach to the candidate's perceived willingness to break with the status quo. This separation allows us to isolate the signaling role of anti-establishment rhetoric from ordinary policy competition.

After observing their type and policy position, the candidate chooses the intensity of an observable anti-establishment attack,

$$a \in [0, \bar{a}].$$

The upper bound $\bar{a}$ captures the limit to meaningful intensification of anti-establishment rhetoric. Candidates can escalate by broadening and strengthening their attacks on established political actors

[3] Treating the alternative as fixed isolates the signaling incentives of insider and outsider types for a given electoral benchmark.

[4] This assumption allows us to distinguish anti-establishment positioning from policy extremism. In particular, candidates located farther from the representative voter's ideal point are not assumed to be more likely to be outsiders, nor does policy distance mechanically generate more intense anti-establishment rhetoric. Insiders and outsiders may occupy any position in the policy space; they differ only in their willingness to sacrifice establishment ties.

and institutions. This rhetorical position eventually reaches a point of saturation: once a candidate has characterized the entire political system as corrupt, captured, or illegitimate and called for its dismantling, more offensive, norm-breaking, or hyperbolic language does not convey greater outsider status. We denote this saturation point by $\bar{a}$. As discussed below, as distrust increases, outsiders escalate toward $\bar{a}$ to preserve separation. Once they reach it, further differentiation through rhetorical escalation is impossible; insiders begin to imitate, bringing anti-establishment rhetoric into the political mainstream.

Anti-establishment attack is costly because it can damage relationships with parties, donors, officials, experts, and other actors whose cooperation provides political value. Let $c(a, \theta)$ denote the cost to a type-θ candidate of choosing attack intensity a. We assume

$$c(a, \theta) = \frac{1}{2} k_\theta a^2, \qquad 0 < k_O < k_I,$$

where $k_\theta > 0$ is the type-specific establishment capital; k_O and k_I denote the coefficients for outsiders and insiders, respectively.

A given anti-establishment attack is therefore less costly for outsiders than for insiders. This reflects the assumption that outsiders have less establishment capital at stake: they possess fewer or less valuable existing ties with established political actors and mainstream institutions. This cost difference allows attack intensity to reveal the candidate's otherwise unobserved type.

2.2 Voter beliefs and the value of perceived detachment

After observing the candidate's attack, the voter forms the posterior probability that the candidate is an outsider:

$$p(a) \equiv \Pr(O \mid a),$$

where $p(a)$ is the voter's updated belief that the candidate is an outsider after observing an attack of intensity a.

To translate this belief into an electoral evaluation, recall that μ denotes institutional trust. The benchmark alternative is assigned credibility μ. A candidate believed to be an insider is likewise assigned credibility μ, while the credibility of a candidate believed to be an outsider is normalized to one. Given the posterior $p(a)$, the candidate's expected credibility is therefore

$$(1 - p(a))\mu + p(a) = \mu + (1 - \mu)p(a).$$

The voter evaluates the candidate relative to the benchmark alternative based on perceived credibility and policy proximity to the voter's ideal point d. Let $w(\mu) \in (0, 1)$ denote the weight placed on perceived credibility when institutional trust is μ, with the remaining weight $1 - w(\mu)$ placed on policy proximity. We assume that w is continuously differentiable and satisfies $w'(\mu) < 0$. Thus, as institutional trust falls, voters place greater weight on perceived credibility and less weight on policy proximity. The voter's evaluation of the candidate is therefore

$$-(1 - w(\mu))(x - d)^2 + w(\mu)\big[\mu + (1 - \mu)p(a)\big]. \tag{1}$$

Rewriting to retain terms that determine the selection between the candidate and the benchmark alternative, the candidate's normalized evaluation can be written as

$$u(a, x; \mu) = -(1 - w(\mu))(x - d)^2 + \beta(\mu)p(a), \tag{2}$$

where

$$\beta(\mu) \equiv w(\mu)(1-\mu) \tag{3}$$

is the electoral premium on perceived outsider status. Holding policy fixed, shifting the voter's belief from certain insider status, $p(a) = 0$, to certain outsider status, $p(a) = 1$, increases the candidate's evaluation by $\beta(\mu)$.

The premium is positive for every $\mu \in (0, 1)$ and strictly decreasing in institutional trust:

$$\beta'(\mu) = w'(\mu)(1-\mu) - w(\mu) < 0. \tag{4}$$

Holding the candidate's policy position fixed, anti-establishment attack affects the voter's evaluation only by changing $p(a)$. A decline in trust therefore raises the electoral reward from shifting the voter's belief toward the outsider type. This trust-dependent reward drives the candidate's signaling decision as becomes clear below.

2.3 Electoral competition and the period game

Let q_B denote the benchmark alternative's distance from the voter's ideal point. The policy-distance gap

$$\Delta q \equiv q - q_B$$

measures how the candidate compares with the benchmark alternative on policy. Following from eq. (2), a candidate has electoral margin over the benchmark alternative:

$$G(\mu, p, \Delta q) = \beta(\mu)p - [1 - w(\mu)]\Delta q. \tag{5}$$

Following the standard probabilistic-voting formulation in Persson and Tabellini (2000), assume a linear probabilistic-voting rule under which the candidate's probability of winning is[5]

$$\Pr(\text{win} \mid G) = \frac{1}{2} + \phi G, \qquad \phi > 0. \tag{6}$$

The candidate therefore wins with probability above one half if and only if $G > 0$.

Let $R > 0$ denote the value of holding office. The candidate's expected payoff is

$$R\left[\frac{1}{2} + \phi G(\mu, p(a), \Delta q)\right] - \frac{1}{2}k_\theta a^2.$$

Within each period game, institutional trust and policy positions are fixed and do not vary with rhetoric a. Thus, simplifying and normalizing $R\phi = 1$, which fixes the units in which signaling costs are measured, gives the candidate's attack problem:

$$V(a, \theta; \mu) = \beta(\mu)p(a) - \frac{1}{2}k_\theta a^2. \tag{7}$$

The electoral margin G and the signaling payoff V serve different roles. Policy distance enters G and can determine whether signaling is sufficient to win, but under the linear electoral rule it

[5]To capture electoral uncertainty, suppose that after the voter evaluates the alternatives, an aggregate popularity shock ε in favor of the benchmark is realized, where $\varepsilon \sim U[-1/(2\phi), 1/(2\phi)]$. We restrict attention to primitives and a support of policy positions such that $|G(\mu, p, \Delta q)| < 1/(2\phi)$ for every $\mu \in [0, 1]$, $p \in [0, 1]$, and Δq in the support. Thus, G always lies within the support of the popularity shock. The candidate wins whenever $G \geq \varepsilon$, and hence $\Pr(\text{win} \mid G) = \Pr(\varepsilon \leq G) = 1/2 + \phi G$.

does not affect the marginal return to attack and therefore does not enter V. The intensity of attack is instead governed by the electoral premium $\beta(\mu)$, the beliefs induced by the attack, and the type-specific cost k_θ, as becomes clear below.

An equilibrium requires candidates' attack choices and voters' interpretations of those choices to be mutually consistent. At each level of institutional trust, a Perfect Bayesian Equilibrium specifies a possibly mixed attack strategy for each type and posterior beliefs $p(a)$. Each type's strategy maximizes $V(a, \theta; \mu)$ given the voter's beliefs, while beliefs follow Bayes' rule on the equilibrium path and electoral choice follows the rule above. We restrict off-path beliefs using the D1 criterion, which favors the type for whom a deviation can be profitable under a wider set of posterior responses (Banks and Sobel, 1987; Cho and Kreps, 1987).

Across the continuum of contests, aggregate attack is

$$A(\mu) = \lambda \mathbb{E}[a \mid \theta = O, \mu] + (1 - \lambda)\mathbb{E}[a \mid \theta = I, \mu], \tag{8}$$

where the expectations include any equilibrium mixing. Because candidates are atomistic, each takes institutional trust as given and does not internalize the effect of their rhetoric on the evolution of trust. This creates a dynamic credibility externality: an attack may improve a candidate's current electoral prospects, but aggregate attack lowers future trust and thereby raises the electoral return to anti-establishment rhetoric in subsequent contests. The next two sections characterize the equilibrium strategies that generate $A(\mu)$, beginning with the separating region.

3 Outsider differentiation

We begin with the separating region. In a separating equilibrium, insiders remain quiet and outsiders attack, so that $p(0) = 0$ and $p(a_O) = 1$. An insider who imitates the outsider receives

$$\beta(\mu) - \frac{1}{2} k_I a_O^2,$$

whereas remaining quiet yields zero. Separation therefore requires

$$\beta(\mu) - \frac{1}{2} k_I a_O^2 \leq 0.$$

The least-cost separating signal sets this constraint at equality.

Proposition 1 (Outsider differentiation)**.** *For $\mu \in (0, 1)$,*[6] *a least-cost separating equilibrium exists if and only if*

$$\beta(\mu) \leq \frac{1}{2} k_I \bar{a}^{\,2}.$$

Whenever this condition holds, the equilibrium selected by D1 is

$$a_I^*(\mu) = 0, \qquad a_O^*(\mu) = \sqrt{\frac{2\beta(\mu)}{k_I}}, \tag{9}$$

with beliefs $p(0) = 0$ and $p(a_O^*) = 1$. Moreover,

$$\frac{da_O^*}{d\mu} = \frac{\beta'(\mu)}{k_I a_O^*(\mu)} < 0. \tag{10}$$

[6] At $\mu = 1$, $\beta(1) = 0$ and both types choose $a = 0$, so no inference about type is possible and the on-path posterior is $p(0) = \lambda$. This degenerate boundary case is distinct from the low-trust pooling region characterized below. The aggregate-attack expressions nevertheless extend continuously to $A(1) = 0$.

At the separating attack, the outsider obtains

$$\beta(\mu) - \frac{1}{2}k_O[a_O^*(\mu)]^2 = \beta(\mu)\left(1 - \frac{k_O}{k_I}\right) > 0. \tag{11}$$

and therefore prefers signaling to remaining quiet.

Proof. See Appendix A. □

Proposition 1 explains why a negative trust shock initially intensifies outsider rhetoric. As μ falls, $\beta(\mu)$ rises, increasing both the value of being recognized as an outsider and the insider's return to imitation. The outsider responds by raising $a_O^*(\mu)$ just enough to keep imitation unprofitable to the insider. The threat of imitation therefore produces escalation before imitation occurs: outsiders attack more not because their preferences change in response to the trust shock, but because preserving differentiation becomes more demanding as detachment grows electorally valuable.

This logic is visible in $a_O^*(\mu) = \sqrt{2\beta(\mu)/k_I}$. The separating attack is pinned down by k_I, the insider's cost of sacrificing establishment ties, rather than by k_O, the outsider's cost. A higher k_I makes copying more costly and permits a weaker separating attack; k_O determines only the outsider's signaling surplus in eq. (11), $\beta(\mu)(1 - k_O/k_I)$. Nor does Δq enter $a_O^*(\mu)$: policy distance affects whether differentiation reverses the electoral ranking, not the intensity required to separate.

Whether differentiation changes the electoral ranking is therefore a separate question. In separation, the voter assigns the outsider posterior $p = 1$. From Eq. (5), the outsider's electoral margin is $G_S(\mu) = \beta(\mu) - [1 - w(\mu)]\Delta q$.

Proposition 2 (The electoral value of differentiation)**.** *Let $\Delta q > 0$. An interior trust threshold at which outsider differentiation reverses the electoral ranking exists if and only if*

$$\Delta q < \frac{w(0)}{1 - w(0)}.$$

Whenever this condition holds, there exists a unique electoral threshold $\check{\mu}(\Delta q) \in (0, 1)$ satisfying

$$G_S(\check{\mu}(\Delta q)) = 0.$$

The separated outsider's electoral margin is positive for $\mu < \check{\mu}(\Delta q)$ and negative for $\mu > \check{\mu}(\Delta q)$. Moreover, $\check{\mu}(\Delta q)$ is decreasing in Δq.

Proof. See Appendix A. □

Proposition 2 distinguishes informational success from electoral advantage. An outsider's anti-establishment rhetoric may convince voters that the candidate is genuinely willing to break with the status quo without making the candidate electorally favored: the outsider's policy position may still be too far from the voters' preferred position. The threshold $\check{\mu}$ identifies the level of institutional trust below which the credibility advantage of being perceived as an outsider becomes large enough to overcome this policy distance.

Successful separation can be sustained only until the outsider's attack reaches $\bar{a}$. Once outsiders exhaust this margin of differentiation, further declines in trust increase the insider's return to imitation without giving outsiders a costlier signal with which to restore separation. Section 4 characterizes the insider entry that follows.

4 Insider entry and mainstreaming

Outsider differentiation has a limit. When the separating attack reaches $\bar{a}$, further declines in trust continue to raise the insider's return to imitation, but outsiders can no longer respond with a costlier signal.

Define the mainstreaming threshold $\widehat{\mu}$ as the level of institutional trust at which an insider is just indifferent between remaining quiet and imitating the outsider's strongest feasible attack:

$$\beta(\widehat{\mu}) = \frac{1}{2}k_I\bar{a}^2, \qquad \text{equivalently} \qquad a_O^*(\widehat{\mu}) = \bar{a}. \tag{12}$$

For $\mu > \widehat{\mu}$, outsiders can deter imitation with an attack below $\bar{a}$. At $\widehat{\mu}$, the insider is just indifferent between remaining quiet and copying the strongest feasible signal. For $\mu < \widehat{\mu}$,

$$\beta(\mu) > \frac{1}{2}k_I\bar{a}^2,$$

so an insider would imitate even the maximal attack if it continued to reveal an outsider with certainty.

Once imitation begins, however, attack is no longer fully revealing. Let $m(\mu)$ denote the probability that an insider chooses $\bar{a}$. Because outsiders choose $\bar{a}$, Bayes' rule gives

$$p(\bar{a}) = \frac{\lambda}{\lambda + (1-\lambda)m(\mu)}. \tag{13}$$

Insider entry is self-limiting at any given level of trust. A higher $m(\mu)$ lowers $p(\bar{a})$ and therefore reduces the credibility benefit from copying the outsider. The equilibrium entry rate is the value of $m(\mu)$ that leaves insiders indifferent between attacking and remaining quiet:

$$\beta(\mu)p(\bar{a}) = \frac{1}{2}k_I\bar{a}^2.$$

When trust falls further, the increase in the electoral premium on perceived outsider status, $\beta(\mu)$, makes imitation attractive. Restoring indifference then requires additional insider entry and a further decline in $p(\bar{a})$. Worsening distrust therefore brings progressively more insiders into anti-establishment rhetoric even as their participation makes that rhetoric less informative. This process continues until every insider attacks and the posterior reaches the prior λ.

Formally, $m(\mu)$ is a mixed-strategy probability. Across a large number of otherwise identical contests, it is also the expected share of insider candidates who attack. If insiders differ slightly in the establishment capital they must sacrifice, $m(\mu)$ can instead be interpreted as the realized fraction for whom imitation has become worthwhile. The least attached insiders enter first; further declines in trust bring progressively more establishment politicians across the threshold.

Proposition 3 (Insider entry and mainstreaming)**.** *Suppose*

$$\beta(0) > \frac{k_I\bar{a}^2}{2\lambda}.$$

Define $\mu_{\text{pool}} < \widehat{\mu}$ *by*

$$\beta(\mu_{\text{pool}}) = \frac{k_I\bar{a}^2}{2\lambda}. \tag{14}$$

For $\widehat{\mu} \leq \mu < 1$, as shown in Proposition 1, outsiders separate by choosing

$$a_O^*(\mu) = \sqrt{\frac{2\beta(\mu)}{k_I}},$$

while insiders remain quiet. For $\mu_{\text{pool}} \leq \mu < \widehat{\mu}$, outsiders choose $\bar{a}$, while insiders attack with probability

$$m(\mu) = \frac{\lambda}{1-\lambda}\left(\frac{2\beta(\mu)}{k_I \bar{a}^2} - 1\right). \tag{15}$$

The posterior attached to attack is

$$p(\bar{a}) = \frac{k_I \bar{a}^2}{2\beta(\mu)}.$$

As trust falls across this region, $m(\mu)$ rises from zero to one, while $p(\bar{a})$ falls from one to λ. At and below μ_{pool}, both types choose $\bar{a}$, and $p(\bar{a}) = \lambda$. Whenever insiders mix, an outsider earns $\frac{1}{2}(k_I - k_O)\bar{a}^2 > 0$ and strictly prefers to attack.

Aggregate anti-establishment attack is

$$A(\mu) = \begin{cases} \lambda\sqrt{\dfrac{2\beta(\mu)}{k_I}}, & \mu \geq \widehat{\mu}, \\ \dfrac{2\lambda\beta(\mu)}{k_I \bar{a}}, & \mu_{\text{pool}} \leq \mu < \widehat{\mu}, \\ \bar{a}, & \mu < \mu_{\text{pool}}. \end{cases} \tag{16}$$

The aggregate attack schedule is continuous but kinked at $\widehat{\mu}$. Under the quadratic signaling-cost specification, insider entry doubles the local responsiveness of aggregate attack to a decline in trust:

$$A'(\widehat{\mu}^-) = 2A'(\widehat{\mu}^+).$$

Proof. See Appendix A. □

Proposition 3 identifies how outsider rhetoric enters the political mainstream. At levels of trust above $\widehat{\mu}$, worsening distrust produces stronger attacks by outsiders. Once outsiders reach $\bar{a}$, further escalation is impossible. Additional distrust instead brings insiders into the attack. At the moment insider entry begins, aggregate attack becomes more responsive to a marginal decline in trust. Each entrant increases aggregate rhetoric while reducing the probability that an attacker is genuinely willing to break with the status quo. At μ_{pool}, every political type attacks, and attack conveys no information beyond the prior. The mainstreaming of anti-establishment rhetoric therefore expands its production while weakening its ability to distinguish political types.

Figure 2 illustrates how mainstreaming changes both the composition and credibility of anti-establishment attack. Panel (a) decomposes aggregate attack into its outsider and insider components. Panel (b) plots the posterior $p(a)$, the probability that an attacker is an outsider. This posterior equals one under separation and, once insiders enter at $\bar{a}$, falls according to Eq. (13) until it reaches the prior λ under full pooling.

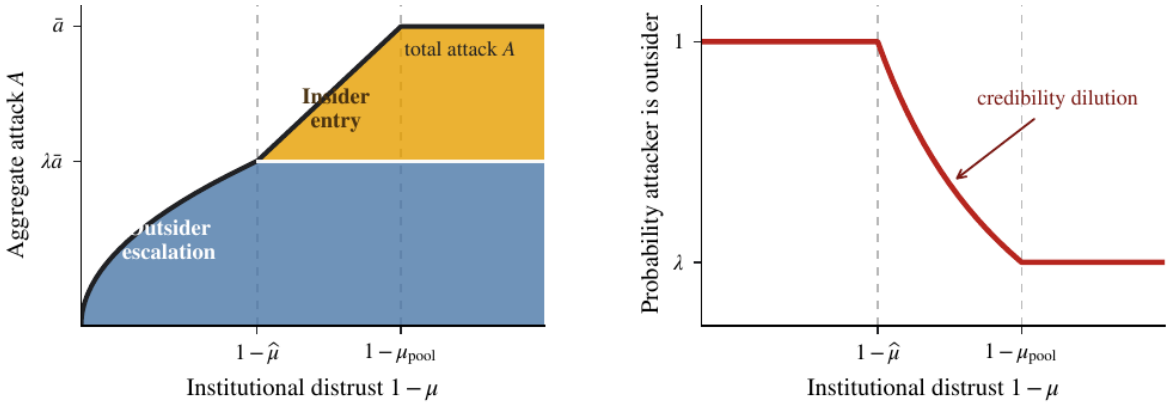


Figure 2. Mainstreaming changes both the composition and credibility of anti-establishment attack.

The decline in the posterior does not mean that voters cease to value detachment from the establishment. Indeed, the electoral premium of being an outsider $\beta(\mu)$ is greatest when trust is lowest. What deteriorates is the information delivered by attack. Under full pooling, observing an attack identifies the candidate as an outsider only with the prior probability λ. The attacking candidate's electoral margin over the benchmark alternative becomes $G_P(\mu) = \lambda\beta(\mu) - [1 - w(\mu)]\Delta q$.

A candidate's policy distance from the voter can therefore again determine the electoral ranking among politicians making the same attack. But this competition occurs in a transformed rhetorical environment: both political types attack, while attack no longer allows voters to distinguish those genuinely willing to break with the status quo from those strategically imitating them.

Mainstreaming need not follow an outsider's electoral victory. Recall from Proposition 2 that a separated outsider becomes electorally favored when trust falls below $\check{\mu}$. When trust is initially high and starts to deteriorate, an interval in which the outsider both separates and wins before insiders enter exists if and only if

$$\hat{\mu} < \check{\mu}.$$

Using Eqs. (5) and (12), this condition is equivalent to

$$\frac{1}{2}k_I\bar{a}^2 > [1 - w(\check{\mu})]\Delta q. \tag{17}$$

When this condition holds, outsider electoral success precedes mainstreaming: outsiders first demonstrate the electoral value of differentiation, and insiders subsequently adopt their rhetoric. When it fails, insiders begin to imitate before a separated outsider overcomes the policy distance. Mainstreaming can therefore be anticipatory rather than reactive. Insiders need not learn from an outsider victory; they respond directly to the same deterioration in trust that makes outsider status electorally valuable.

Section 3 established that distrust initially strengthens outsider differentiation. Section 4 shows how the same force eventually undermines it. Once outsiders can no longer escalate, insider entry

makes anti-establishment rhetoric more prevalent, more responsive to distrust, and less informative about political type. The next section asks what happens when this increase in aggregate attack feeds back into institutional trust.

5 Dynamic trust

The previous sections derive the equilibrium supply of anti-establishment attack at every level of inherited trust. We now carry that political response into the evolution of trust itself. The question is whether insider entry merely increases rhetoric temporarily or changes the set of long-run trust outcomes.

Let $\mu_t \in [0,1]$ denote institutional trust inherited at date t. In the absence of political attack, trust adjusts toward a baseline $\bar{\mu} \in (0,1)$. This baseline summarizes the credibility supported by institutional performance, accumulated experience, and other sources of confidence outside the model. Aggregate anti-establishment attack reduces trust by repeatedly portraying the political system as corrupt, captured, or broken. Trust evolves according to

$$\mu_{t+1} = \left[\mu_t + \rho(\bar{\mu} - \mu_t) - \delta A(\mu_t) + \varepsilon_t\right]_{[0,1]}, \tag{18}$$

where $\rho \in (0,1)$ is the rate of institutional recovery, $\delta > 0$ is the residual effect of aggregate attack on institutional beliefs, and ε_t is an exogenous trust shock.

Equation (18) is a belief dynamic. It does not require anti-establishment rhetoric to damage institutional performance directly, nor does it assume that voters accept every political claim at face value. The coefficient δ captures the part of repeated systemic attack that changes institutional evaluations after voters account for its source and strategic character. This paper's contribution is to endogenize the production of that rhetoric: who attacks, how intensively they attack, and how the composition of speakers changes with inherited trust.

Evidence that elite rhetoric can alter beliefs about democratic institutions and procedures makes such an effect plausible (Clayton et al., 2021; Valentim et al., 2025).

Define the erosion-to-recovery ratio $\kappa \equiv \delta/\rho$ and the credibility target

$$C(\mu) \equiv \bar{\mu} - \kappa A(\mu). \tag{19}$$

With the shock switched off,

$$\mu_{t+1} = (1-\rho)\mu_t + \rho C(\mu_t). \tag{20}$$

An interior resting point satisfies $\mu = C(\mu)$. It is useful to define the resting-point gap

$$D_\kappa(\mu) \equiv C(\mu) - \mu = \bar{\mu} - \mu - \kappa A(\mu). \tag{21}$$

Trust rises when $D_\kappa(\mu) > 0$ and falls when $D_\kappa(\mu) < 0$. At a differentiable resting point,

$$F'(\mu) = 1 + \rho D'_\kappa(\mu) = 1 - \rho + \delta|A'(\mu)|. \tag{22}$$

Because $A'(\mu) \leq 0$, the transition is increasing. A resting point is locally stable when $D'_\kappa(\mu) < 0$, equivalently $\kappa|A'(\mu)| < 1$, and unstable when $D'_\kappa(\mu) > 0$.

To derive rather than assume the shape of the trust dynamic, we impose one regularity condition on the primitive electoral premium $\beta(\mu) = w(\mu)(1-\mu)$. Maintain $\beta(\mu) > 0$ and $\beta'(\mu) < 0$ on $(0,1)$ and $\beta(1) = 0$. Assume that β is twice continuously differentiable on $(0,1)$ and satisfies

$$0 \leq 2\beta(\mu)\beta''(\mu) \leq [\beta'(\mu)]^2. \tag{23}$$

The first inequality makes the insider-entry response convex in distrust. The second makes the square-root separating response sufficiently concave in trust to rule out additional high-trust crossings. This is a restriction on the exogenous voter-weight function, not on equilibrium roots. It is satisfied automatically by the affine family used in Figure 3,

$$w(\mu) = \omega_0 + \omega_1(1-\mu), \qquad \omega_0 > 0, \quad \omega_1 > 0, \quad \omega_0 + \omega_1 < 1. \tag{24}$$

because then $\beta''(\mu) = 2\omega_1 > 0$ and $[\beta']^2 - 2\beta\beta'' = \omega_0^2 > 0$. All inequalities used below are strict in the numerical construction, so the result also holds for an open set of nearby functions.

5.1 Outsider rhetoric without mainstreaming

We first ask whether outsiders alone can sustain a credibility crisis. Suppose insiders remain quiet even after outsiders reach the strongest feasible signal. The maximal outsider-only attack schedule is

$$A^O(\mu) = \begin{cases} \lambda\sqrt{\dfrac{2\beta(\mu)}{k_I}}, & \widehat{\mu} \le \mu < 1, \\ \lambda\bar{a}, & \mu < \widehat{\mu}, \end{cases} \qquad A^O(1) = 0. \tag{25}$$

This is deliberately demanding: below $\widehat{\mu}$, it assigns outsiders their maximal feasible attack at every trust level. Any outsider-only environment produces no more than $\lambda\bar{a}$ aggregate attack.

Let $D_\kappa^O(\mu) = \bar{\mu} - \mu - \kappa A^O(\mu)$. Define

$$\kappa_H \equiv \frac{\bar{\mu} - \widehat{\mu}}{\lambda\bar{a}}. \tag{26}$$

If $\kappa < \kappa_H$, then $D_\kappa^O(\widehat{\mu}) > 0$, while $D_\kappa^O(1) = \bar{\mu} - 1 < 0$. Condition (23) implies that A^O is concave on $(\widehat{\mu}, 1)$ and hence that D_κ^O is convex there. The endpoint signs and convexity imply exactly one crossing $\mu_H \in (\widehat{\mu}, 1)$, and the gap crosses from positive to negative at that point. Thus $D_\kappa^{O\prime}(\mu_H) < 0$: the high-trust resting point is locally stable. Below $\widehat{\mu}$, the gap is bounded below by

$$D_\kappa^O(\mu) \ge \bar{\mu} - \widehat{\mu} - \kappa\lambda\bar{a} > 0,$$

so there is no outsider-only low-trust resting point.

This result does not require a global derivative bound. Indeed, $|(A^O)'(\mu)|$ diverges as $\mu \to 1$ because the separating attack approaches zero with a square-root cusp. Curvature and endpoint signs, rather than a uniform slope restriction, establish the unique stable high-trust equilibrium.

The outsider-only benchmark isolates the source of persistence. When $\kappa < \kappa_H$, even maximal outsider attack cannot hold trust below the mainstreaming threshold. A negative shock may intensify outsider rhetoric, but trust ultimately returns to μ_H if insiders stay out.

5.2 Mainstreaming and the credibility trap

Return now to the equilibrium attack schedule from Proposition 3:

$$A(\mu) = \begin{cases} \lambda\sqrt{\dfrac{2\beta(\mu)}{k_I}}, & \widehat{\mu} \le \mu < 1, \\ \dfrac{2\lambda\beta(\mu)}{k_I\bar{a}}, & \mu_{\text{pool}} \le \mu < \widehat{\mu}, \\ \bar{a}, & \mu < \mu_{\text{pool}}, \end{cases} \qquad A(1) = 0. \tag{27}$$

Mainstreaming changes the dynamic through two margins. First, it raises the maximal level of aggregate attack from $\lambda\bar{a}$ to $\bar{a}$. Second, it doubles the response of aggregate attack to a marginal decline in trust at the entry threshold:

$$|A'(\widehat{\mu}^-)| = 2|A'(\widehat{\mu}^+)|, \qquad |A'(\widehat{\mu}^+)| = \frac{\lambda|\beta'(\widehat{\mu})|}{k_I\bar{a}}. \tag{28}$$

The level effect can support a low-trust state; the responsiveness effect can make the boundary between the high- and low-trust basins unstable.

Define the lower threshold

$$\kappa_L \equiv \frac{\bar{\mu} - \mu_{\text{pool}}}{\bar{a}}. \tag{29}$$

The interval (κ_L, κ_H) is nonempty if and only if

$$\lambda < \frac{\bar{\mu} - \widehat{\mu}}{\bar{\mu} - \mu_{\text{pool}}}. \tag{30}$$

Finally, let

$$S_H \equiv \frac{\lambda|\beta'(\widehat{\mu})|}{k_I\bar{a}}$$

denote the responsiveness of outsider attack immediately above $\widehat{\mu}$. We require the derived kink to reverse the local force at the upper boundary:

$$\kappa_H S_H < 1 < 2\kappa_H S_H. \tag{31}$$

The first inequality says recovery dominates outsider escalation just above the threshold; the second says the doubled response created by insider entry dominates recovery just below it. Unlike a condition on equilibrium roots, Equation (31) is a closed-form comparison of primitive parameters. The factor of two comes from the signaling equilibrium.

Proposition 4 (When mainstreaming creates the credibility trap)**.** *Suppose $\bar{\mu} > \widehat{\mu}$, the regular-curvature condition in equation (23) holds, the interval in equation (30) is nonempty, and the kink-reversal condition in equation (31) holds. If*

$$\kappa_L < \kappa < \kappa_H,$$

then: (i) under the maximal outsider-only schedule A^O, the trust dynamic has exactly one resting point $\mu_H \in (\widehat{\mu}, 1)$, and it is stable; (ii) under the equilibrium schedule A, the trust dynamic has exactly three resting points

$$\mu_H > \mu_T > \mu_L;$$

(iii) μ_H and μ_L are stable, while μ_T is unstable; and (iv)

$$\mu_T \in (\mu_{\text{pool}}, \widehat{\mu}), \qquad \mu_L = \max\{0, \bar{\mu} - \kappa\bar{a}\} < \mu_{\text{pool}}.$$

Proof. See Appendix A. □

The proposition derives the phase portrait from the attack schedule rather than assuming its crossings. At μ_{pool},

$$D_\kappa(\mu_{\text{pool}}) = \bar{a}(\kappa_L - \kappa) < 0,$$

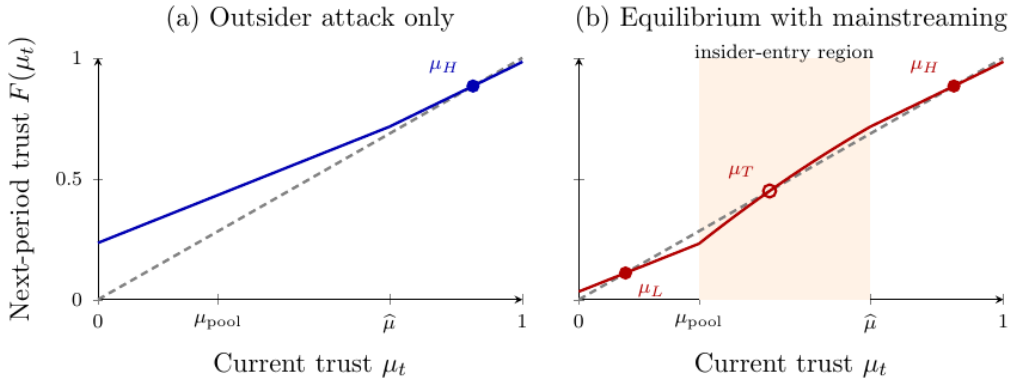


Figure 3: **Mainstreaming changes the long-run dynamics of trust.** The dashed line is the 45-degree line. Filled circles denote stable resting points; the open circle denotes the unstable tipping point. Panel (a) uses the maximal outsider-only schedule and has only the stable high-trust equilibrium. Panel (b) uses the equilibrium attack schedule. Insider entry creates a stable low-trust equilibrium and an unstable basin boundary. The curves use $w(\mu) = 0.02 + 0.75(1-\mu)$, $k_I = 1$, $\bar{a} = 0.4$, $\lambda = 0.2$, $\bar{\mu} = 0.95$, $\rho = 0.30$, and $\delta = 0.63$, so $\kappa = 2.1$. Appendix A gives the construction.

whereas at $\widehat{\mu}$,

$$D_\kappa(\widehat{\mu}) = \lambda\bar{a}(\kappa_H - \kappa) > 0.$$

In the insider-entry region, Equation (23) makes D_κ concave. The two endpoint signs therefore generate exactly one crossing, and it must cross from negative to positive. That crossing is μ_T, and $D'_\kappa(\mu_T) > 0$ makes it unstable. Below μ_{pool}, attack is constant at $\bar{a}$, giving the stable low point μ_L. Above $\widehat{\mu}$, the same curvature condition yields the unique stable high point established in the outsider-only benchmark.

Mainstreaming does not merely intensify an independently existing crisis. The upper bound $\kappa < \kappa_H$ means that even maximal outsider attack cannot sustain trust below $\widehat{\mu}$. The lower bound $\kappa > \kappa_L$ means that the attack produced when both types participate is sufficient to sustain trust below μ_{pool}. Insiders therefore change the set of long-run outcomes.

The two margins of mainstreaming perform different roles. The increase in the level of attack, from at most $\lambda\bar{a}$ to $\bar{a}$, supports the low-trust resting point. The increase in responsiveness creates the unstable boundary between basins. Equation (31) isolates this second role: institutional recovery can dominate the outsider response and yet be overwhelmed after insider entry doubles the response of aggregate rhetoric.

Outsider scarcity is consequently substantive. Condition (30) requires outsiders to be sufficiently scarce that their maximal aggregate attack cannot sustain the low-trust state on its own. Yet Proposition 3 also requires enough outsider presence for full insider entry to occur at an interior trust level. The trap arises in an environment where outsider status is initially scarce and informative but becomes sufficiently valuable for insiders to imitate.

5.3 Temporary shocks and hysteresis

The three resting points divide the state space into two basins. Suppose the polity begins at μ_H and receives a one-period negative shock $\varepsilon < 0$, after which $\varepsilon_t = 0$. Because the unshocked transition leaves μ_H unchanged, the post-shock state is $\mu_1 = [\mu_H + \varepsilon]_{[0,1]}$. Define

$$s^* \equiv \mu_H - \mu_T. \tag{32}$$

Proposition 5 (Persistence and hysteresis)**.** *Maintain the conditions of Proposition 4. Beginning from μ_H, a one-period negative shock satisfies*

$$|\varepsilon| < s^* \Longrightarrow \mu_t \to \mu_H, \qquad |\varepsilon| > s^* \Longrightarrow \mu_t \to \mu_L.$$

Equality places the polity at the unstable resting point μ_T. Moreover, relaxing only the restriction $\kappa_L < \kappa < \kappa_H$ and allowing κ to vary, the equilibrium branch structure is

$$\begin{aligned} \kappa < \kappa_L \quad &: \quad \textit{one stable high-trust resting point,} \\ \kappa_L < \kappa < \kappa_H \quad &: \quad \textit{stable high and low points separated by } \mu_T, \\ \kappa > \kappa_H \quad &: \quad \textit{one stable low-trust resting point.} \end{aligned}$$

The healthy branch disappears at κ_H, while the trapped branch disappears only at $\kappa_L < \kappa_H$.

Proof. See Appendix A. □

Persistence comes from political competition after the shock, not from persistence of the shock itself. A disturbance that leaves trust above μ_T increases attack but does not change the long-run state: the gap remains positive until trust returns to μ_H. A disturbance that crosses μ_T enters a region where the gap is negative. Insider participation expands, aggregate attack rises, and trust continues to fall after the original disturbance has disappeared.

The eventual decline can exceed the disturbance that triggers it. The critical shock threshold separating the two basins is $s^* = \mu_H - \mu_T$, while the resulting long-run decline is $\mu_H - \mu_L$. Since $\mu_L < \mu_T$,

$$\mu_H - \mu_L > \mu_H - \mu_T.$$

The model generates disproportionate collapse: a temporary disturbance just above the threshold can leave a permanent loss of trust greater than the disturbance itself.

The second part of Proposition 5 establishes hysteresis without assuming that the response ordering continues as κ changes. Regular curvature controls the gap for every value of κ. At κ_H, the stable high point and the unstable tipping point meet at $\widehat{\mu}$. The kink-reversal condition ensures that once κ rises above this value, the gap is negative throughout the insider-entry and separating regions, so the polity moves toward μ_L. Starting from the trapped state and reducing κ, the low point persists until it meets the tipping point at μ_{pool} when $\kappa = \kappa_L$. Only below that value is the gap positive throughout the low and insider-entry regions.

The asymmetry follows from the endogenous composition of political speakers. Before collapse, only outsiders supply the strongest attacks, so aggregate erosion is bounded by $\lambda\bar{a}$. After collapse, outsiders and insiders both attack, sustaining aggregate erosion of $\bar{a}$. Returning the erosion-to-recovery ratio merely to the level immediately preceding collapse is therefore insufficient. Recovery requires the larger change from κ_H to κ_L.

Nor can an individual politician generally restore trust through unilateral moderation. In the selected pooling equilibrium, a deviation from maximal attack is attributed to the type with the

stronger incentive to preserve establishment capital: the insider. A politician who moderates is therefore identified with the establishment precisely when establishment attachment carries its greatest electoral penalty. Each politician remains willing to participate in an aggregate practice that sustains distrust, even though the rhetoric no longer distinguishes political types.

The trap is consequently both dynamic and strategic. Trust remains low because mainstreamed attack continues to erode it, while attack remains mainstreamed because low trust makes withdrawal electorally costly. Escaping the trap requires either a temporary intervention large enough to move trust above μ_T or a lasting change in $\bar{\mu}$, ρ, δ, or the political supply of attack. Sections 6 and 7 examine the observable implications of these mechanisms and the institutional changes capable of weakening them.

6 Discussion

The model identifies how institutional trust and political competition over perceived outsider credibility reorganize the supply of anti-establishment rhetoric. Distrust changes the electoral return to claiming outsider status, the intensity required to sustain differentiation, and the set of politicians willing to bear that cost. These results yield observable implications for the order in which politicians adopt anti-establishment rhetoric, the political and communication environments that shape its mainstreaming, the relationship between rhetorical and policy convergence, and the consequences for institutional trust.

First, the model predicts an ordered adoption of anti-establishment rhetoric across political types. Following a negative shock to institutional trust, candidates with clear ex ante outsider profiles should intensify their attacks first. Insider adoption should begin only after distrust has raised the electoral return to perceived detachment sufficiently to offset the higher cost of sacrificing establishment capital. Among insiders, adoption should proceed from the least to the most embedded: opposition politicians, backbenchers, recent entrants, and politicians with weaker party or donor ties should adopt the rhetoric before senior officeholders whose careers depend more heavily on established relationships. More broadly, this ordering may also be shaped by career structures and promotion incentives, which affect politicians' behavior across stages of office (Martinelli, 2022). Crucially, mainstreaming of anti-establishment rhetoric need not follow an outsider victory: it can be anticipatory rather than reactive. The distinctive empirical sequence is outsider escalation followed by delayed and ordered insider entry.

Second, political organization shapes the timing of mainstreaming. In party-centered systems, politicians depend more heavily on party leaders for nominations, campaign resources, and career advancement. This dependence raises k_I, increasing the cost to insiders of attacking established institutions, and delays their adoption of anti-establishment rhetoric. In candidate-centered systems, politicians can build personal reputations, raise funds, and communicate with voters more independently (Carey and Shugart, 1995). Open primaries, personalized competition, weak nomination control, and direct-to-voter fundraising may therefore lower k_I and bring the mainstreaming of anti-establishment rhetoric forward. Holding institutional trust fixed, insider adoption should occur earlier where political careers are more personalized and organizational gatekeepers are weaker. Institutions that broaden political access may thus have an unintended consequence: they make outsider rhetoric cheaper for insiders to copy.

Third, our model allows mainstream parties to converge rhetorically without necessarily converging in policy. Policy distance Δq determines whether perceived outsider status changes who wins,

but not how much anti-establishment rhetoric is needed to preserve separation or when insiders begin to imitate. Mainstream politicians can therefore adopt outsider rhetoric without changing their substantive positions. This mechanism complements accounts in which mainstream parties respond to challenger success by shifting their policy positions (Abou-Chadi and Krause, 2020). Here, declining trust makes visible attachment to the establishment electorally costly, prompting insiders to adjust their rhetorical strategy rather than their policy platform. Empirically, convergence in anti-establishment rhetoric need not coincide with corresponding changes in manifesto positions, issue placements, or legislative behavior.

Fourth, communication environments shape both the cost of adopting anti-establishment rhetoric and the reach of its claims. Research on digital populism emphasizes that social-media platforms permit direct communication, weaken journalistic gatekeeping, and facilitate repeated and fragmented anti-elite appeals (Engesser et al., 2017; Gerbaudo, 2018; Ernst et al., 2019). In terms of the model, these features operate through two channels: the insider's cost of adopting anti-establishment rhetoric, k_I, and the effect of that rhetoric on institutional trust, δ.

First, social-media platforms may reduce k_I. Politicians who can mobilize supporters, raise funds, and construct personal followings outside established organizations sacrifice less by attacking party elites, professional media, or governing institutions. A lower k_I makes imitation profitable at a higher level of trust and brings the mainstreaming of anti-establishment rhetoric forward.

Second, social-media platforms may increase δ, the effect of anti-establishment rhetoric on institutional beliefs. Visibility, repetition, targeting, and algorithmic amplification do not necessarily change what politicians sacrifice by attacking; they change the reach and persistence of the systemic claim. Because $\kappa = \delta/\rho$, a higher δ strengthens rhetorical erosion relative to institutional recovery. Social-media platforms may therefore accelerate the mainstreaming of anti-establishment rhetoric and deepen its consequences for institutional trust.

More broadly, the model identifies a dynamic credibility externality in electoral competition. As institutional trust declines, anti-establishment rhetoric generates a private electoral return by distancing a politician from a distrusted establishment. Yet the same rhetoric further erodes institutional trust, while that cost is dispersed across the political system. The externality is self-reinforcing: as trust falls, the return to appearing detached rises, while moderation increasingly identifies a politician with the establishment voters distrust. This creates a collective-action problem: anti-establishment rhetoric is individually rational but collectively damaging. In full pooling, no politician has an incentive to moderate unilaterally, because the politician who exercises restraint bears an immediate electoral cost. Institutional recovery therefore cannot be expected to emerge from individual moderation; it requires coordination or institutional changes that reduce the private return to anti-establishment rhetoric or weaken its effect on trust. We discuss these mechanisms next.

7 Institutional reform

Our model yields direct implications for institutional reform. Recall that institutional trust evolves according to

$$\mu_{t+1} - \mu_t = \rho(\bar{\mu} - \mu_t) - \delta A(\mu_t).$$

Institutional reform can alter this dynamic through three channels. First, improvements in institutional performance and the visibility of those improvements can strengthen the recovery of trust by raising $\bar{\mu}$ or ρ. Second, credible and independently verifiable information can reduce

δ, making institutional beliefs less responsive to political rhetoric. Third, changes in political organization and coordination can reduce the equilibrium supply of anti-establishment attack, $A(\mu)$. We consider each channel in turn.

The first channel is institutional performance. Competent government, impartial law enforcement, reliable public-service provision, and lower levels of corruption raise $\bar{\mu}$, the level toward which trust recovers in the absence of political attack. Reforms that make these improvements visible and attributable can also increase ρ, the rate at which improved performance translates into renewed trust. Together, these changes strengthen the recovery term $\rho(\bar{\mu}-\mu)$. Research identifies government performance and political processes as important sources of political trust, while also showing that poor performance can erode trust more readily than good performance restores it (Citrin and Stoker, 2018; Hetherington and Rudolph, 2008). This asymmetry becomes more consequential once anti-establishment rhetoric has become mainstream. Institutional improvements must then offset a continuing flow of anti-establishment rhetoric from across the political system. They must therefore be sufficiently large, visible, and durable to generate a self-sustaining recovery.

The second channel is the information environment through which political rhetoric affects trust. Independent fact-checking organizations and widely trusted sources of public information can reduce δ by making claims about institutional misconduct easier to verify. Evidence suggests that independently produced information can strengthen electoral accountability and improve the accuracy of political beliefs, although its effects depend on dissemination and prior attitudes (Ferraz and Finan, 2008; Walter et al., 2020). Such mechanisms need not increase trust mechanically: verification may substantiate political criticism as well as refute it. Their role is to make beliefs about institutions less responsive to rhetoric alone.

The third channel is aggregate political attack, $A(\mu)$. Once anti-establishment rhetoric has become mainstream, preserving institutional trust poses a collective-action problem. A politician who moderates bears the electoral cost of appearing closer to a distrusted establishment, while the resulting gain in institutional credibility is shared across political competitors. Stronger party discipline and more party-centered forms of political organization can reduce aggregate attack by making politicians more dependent on collective party reputations and organizational resources. In terms of the model, these arrangements raise k_I, delaying insider entry and reducing $A(\mu)$. They thereby weaken the feedback between institutional distrust and anti-establishment rhetoric.

Taken together, these three channels show why institutional reform becomes more demanding after anti-establishment rhetoric has become mainstream. Improvements in institutional performance must then overcome not only inherited distrust but also a political response that continually reproduces it. Durable recovery requires reforms that shift the balance between institutional recovery and rhetorical erosion—by raising $\bar{\mu}$ or ρ, reducing δ, or lowering $A(\mu)$—enough to make renewed trust self-sustaining. This is the institutional implication of hysteresis: a reform sufficient to prevent a credibility crisis may be insufficient to reverse one after anti-establishment rhetoric has become mainstream.

8 Conclusion

This paper develops a theory of institutional trust erosion driven by political competition over anti-establishment rhetoric. Because such rhetoric both signals political type and contributes to declining trust, a temporary shock can alter political incentives in ways that make distrust self-entrenching even after the shock has disappeared. The analysis thus offers a new perspective on

credibility crises in democracies: their persistence may arise not only from continuing institutional failures, but also from the political strategies that the crisis itself induces.

Our paper sheds new light on why the adoption of anti-establishment rhetoric by insiders is so consequential for institutional trust. When governing parties, former officeholders, and long-standing political actors begin to campaign against "the system," they do more than borrow a successful outsider strategy. They multiply the voices portraying the institutions they themselves inhabit as corrupt, captured, or broken. Our model identifies this as the turning point at which insiders become the marginal suppliers of anti-establishment rhetoric. Mainstream parties thereby cease merely to be the target of distrust and become a source of its persistence.

More broadly, we hope to contribute to a deeper understanding of the role of institutional trust in democratic life. Institutional trust is not simply a voter attitude or a reflection of institutional performance; it is also shaped by rhetoric deployed by political actors who respond strategically to it. By choosing the intensity of their anti-establishment rhetoric, politicians respond to inherited distrust while altering the credibility environment in which future competition occurs. A costly-signaling framework therefore reveals trust as both a condition and an outcome of political competition. Future research should therefore examine when the interaction between trust and anti-establishment rhetoric becomes self-reinforcing and which features of democratic politics allow such dynamics to be reversed.

References

Abou-Chadi, T., and Krause, W. (2020). The causal effect of radical right success on mainstream parties' policy positions: A regression discontinuity approach. *British Journal of Political Science*, 50(3):829–847.

Acemoglu, D., Egorov, G., and Sonin, K. (2013). A political theory of populism. *Quarterly Journal of Economics*, 128(2):771–805.

Banks, J. S. (1990). A model of electoral competition with incomplete information. *Journal of Economic Theory*, 50(2):309–325.

Banks, J. S. and Sobel, J. (1987). Equilibrium selection in signaling games. *Econometrica*, 55(3):647–661.

Breyer, M. (2023). Populist positions in party competition: Do parties strategically vary their degree of populism in reaction to vote and office loss? *Party Politics*, 29(4):672–684.

Carey, J. M. and Shugart, M. S. (1995). Incentives to cultivate a personal vote: A rank ordering of electoral formulas. *Electoral Studies*, 14(4):417–439.

Cho, I.-K. and Kreps, D. M. (1987). Signaling games and stable equilibria. *Quarterly Journal of Economics*, 102(2):179–221.

Citrin, J. and Stoker, L. (2018). Political trust in a cynical age. *Annual Review of Political Science*, 21:49–70.

Clayton, K., Davis, N. T., Nyhan, B., Porter, E., Ryan, T. J., and Wood, T. J. (2021). Elite rhetoric can undermine democratic norms. *Proceedings of the National Academy of Sciences*, 118(23):e2024125118.

Engesser, S., Ernst, N., Esser, F., and Büchel, F. (2017). Populism and social media: How politicians spread a fragmented ideology. *Information, Communication & Society*, 20(8):1109–1126.

Ernst, N., Blassnig, S., Engesser, S., Büchel, F., and Esser, F. (2019). Populists prefer social media over talk shows: An analysis of populist messages and stylistic elements across six countries. *Social Media + Society*, 5(1):1–14.

Ferraz, C. and Finan, F. (2008). Exposing corrupt politicians: The effects of Brazil's publicly released audits on electoral outcomes. *Quarterly Journal of Economics*, 123(2):703–745.

Gennaro, G., Lecce, G., and Morelli, M. (2024). Intertemporal evidence on the strategy of populism in the united states. *Electoral Studies*, 92:102853.

Gerbaudo, P. (2018). Social media and populism: An elective affinity? *Media, Culture & Society*, 40(5):745–753.

Glynia, N., Manalis, G., and Xefteris, D. (2025). Trust dynamics in electoral competition. *European Economic Review*, 179:105125.

Gonnot, J. and Seabright, P. (2021). Establishment and outsiders: Can political incorrectness and social extremism work as a signal of commitment to populist policies? CEPR Discussion Paper 15971, Centre for Economic Policy Research.

Guiso, L., Herrera, H., Morelli, M., and Sonno, T. (2024). Economic insecurity and the demand for populism in Europe. *Economica*, 91(362):588–620.

Guriev, S., and Papaioannou, E. (2022). The political economy of populism. *Journal of Economic Literature*, 60(3):753–832.

Hetherington, M. J. and Rudolph, T. J. (2008). Priming, performance, and the dynamics of political trust. *Journal of Politics*, 70(2):498–512.

Herrera, H. and Trombetta, F. (2024). Alternative worldviews, distrust and populism. CEPR Discussion Paper 19674, Centre for Economic Policy Research.

Kartik, N. and McAfee, R. P. (2007). Signaling character in electoral competition. *American Economic Review*, 97(3):852–870.

Martinelli, C. (2022). Accountability and grand corruption. *American Economic Journal: Microeconomics*, 14(4):645–679.

Meguid, B. M. (2005). Competition between unequals: The role of mainstream party strategy in niche party success. *American Political Science Review*, 99(3):347–359.

Mudde, C., and Rovira Kaltwasser, C. (2017). *Populism: A Very Short Introduction*. Oxford University Press.

Milei, J. (2025). Milei en el CICyP: "La casta hará lo posible por defender sus privilegios." Argentina.gob.ar, August 28.

Persson, T., and Tabellini, G. (2000). *Political Economics: Explaining Economic Policy*. Cambridge, MA: MIT Press.

Rodrik, D. (2021). Why does globalization fuel populism? Economics, culture, and the rise of right-wing populism. *Annual Review of Economics*, 13:133–170.

Sunak, R. (2023). "Be in no doubt it is time for a change and we are it": Conservative Party Conference speech. October 4.

Surdea-Hernea, V. (2025). The populist spiral: How populist rhetoric spreads within party systems. *Party Politics*.

Trump, D. J. (2016). A video message from President-Elect Donald J. Trump. The American Presidency Project, November 21.

Valentim, V., Dinas, E., and Ziblatt, D. (2025). How mainstream politicians erode norms. *British Journal of Political Science*, 55:e105.

Walter, N., Cohen, J., Holbert, R. L., and Morag, Y. (2020). Fact-checking: A meta-analysis of what works and for whom. *Political Communication*, 37(3):350–375.

Appendix for "When Outsider Rhetoric Goes Mainstream: Political Competition and the Credibility Trap"

A Proofs

Throughout, $\beta(\mu) = w(\mu)(1-\mu)$ is the electoral premium on perceived outsider status, $\beta'(\mu) < 0$, and $0 < k_O < k_I$. The attack space is $[0, \bar{a}]$. The mainstreaming and full-pooling thresholds satisfy

$$\beta(\widehat{\mu}) = \frac{1}{2} k_I \bar{a}^2, \qquad \beta(\mu_{\text{pool}}) = \frac{k_I \bar{a}^2}{2\lambda}.$$

A.1 Preliminary lemmas

Lemma 1 (The outsider premium falls with trust)**.** *On* $(0,1)$*,* $\beta'(\mu) < 0$*.*

Proof. Differentiation gives

$$\beta'(\mu) = w'(\mu)(1-\mu) - w(\mu).$$

Because $w'(\mu) < 0$, $w(\mu) > 0$, and $1 - \mu > 0$, both terms are negative. □

Lemma 2 (Feasibility of separation)**.** *The least-cost separating attack* $a_O^*(\mu) = \sqrt{2\beta(\mu)/k_I}$ *is feasible if and only if* $\mu \geq \widehat{\mu}$*.*

Proof. Since both quantities are nonnegative,

$$a_O^*(\mu) \leq \bar{a} \iff \frac{2\beta(\mu)}{k_I} \leq \bar{a}^2 \iff \beta(\mu) \leq \frac{1}{2} k_I \bar{a}^2.$$

The last inequality is equivalent to $\mu \geq \widehat{\mu}$ because β is strictly decreasing. □

Lemma 3 (The insider entry probability)**.** *The expression*

$$m(\mu) = \frac{\lambda}{1-\lambda} \left(\frac{2\beta(\mu)}{k_I \bar{a}^2} - 1 \right)$$

belongs to $[0,1]$ *if and only if* $\mu \in [\mu_{\text{pool}}, \widehat{\mu}]$*. It rises from zero to one as trust falls across this interval.*

Proof. At $\widehat{\mu}$, $2\beta/(k_I \bar{a}^2) = 1$, so $m = 0$. At μ_{pool}, $2\beta/(k_I \bar{a}^2) = 1/\lambda$, so $m = 1$. Strict monotonicity follows from $\beta'(\mu) < 0$. □

A.2 Proof of Proposition 1

Fix $\mu \in (0,1)$. In the proposed separating equilibrium, the insider chooses 0 and the outsider chooses a_O^*. Bayes' rule therefore gives $p(0) = 0$ and $p(a_O^*) = 1$. An insider who imitates obtains

$$\beta(\mu) - \frac{1}{2} k_I (a_O^*)^2,$$

whereas remaining quiet yields zero. Deterring imitation requires

$$(a_O^*)^2 \geq \frac{2\beta(\mu)}{k_I}.$$

The least-cost separating signal makes this constraint bind, yielding

$$a_I^*(\mu) = 0, \qquad a_O^*(\mu) = \sqrt{\frac{2\beta(\mu)}{k_I}}.$$

At this signal, the outsider's payoff is

$$\beta(\mu) - \frac{1}{2}k_O(a_O^*)^2 = \beta(\mu)\left(1 - \frac{k_O}{k_I}\right) > 0,$$

so the outsider strictly prefers signaling to remaining quiet. Within the maintained domain, feasibility is equivalent to $\widehat{\mu} \leq \mu < 1$ by Lemma 2.

For equilibrium selection, consider the family of separating profiles supported by off-path beliefs. Any separating signal strictly above $\sqrt{2\beta/k_I}$ admits a downward deviation that is unprofitable for the insider but profitable for the outsider under an outsider posterior. The D1 criterion therefore eliminates higher-cost separating profiles and selects the least-cost signal. For actions not assigned an outsider posterior by D1, set $p(a) = 0$; neither type then benefits from such a deviation.

Finally,

$$\frac{da_O^*}{d\mu} = \frac{\beta'(\mu)}{k_I a_O^*(\mu)} < 0.$$

This proves Proposition 1.

A.3 Proof of Proposition 2

In separation, the outsider is identified with probability one. Its electoral margin is

$$G_S(\mu) = \beta(\mu) - [1 - w(\mu)]\Delta q.$$

For $\Delta q > 0$,

$$G_S'(\mu) = \beta'(\mu) + w'(\mu)\Delta q < 0.$$

At the endpoints,

$$G_S(1) = -[1 - w(1)]\Delta q < 0, \qquad G_S(0) = w(0) - [1 - w(0)]\Delta q.$$

The assumed bound $0 < \Delta q < w(0)/[1 - w(0)]$ makes $G_S(0) > 0$. Continuity and strict monotonicity therefore imply a unique $\check{\mu} \in (0, 1)$ satisfying $G_S(\check{\mu}) = 0$, with $G_S(\mu) > 0$ below the threshold and $G_S(\mu) < 0$ above it.

Implicit differentiation gives

$$\frac{d\check{\mu}}{d\Delta q} = -\frac{\partial G_S/\partial \Delta q}{\partial G_S/\partial \mu} = \frac{1 - w(\check{\mu})}{G_S'(\check{\mu})} < 0.$$

This proves Proposition 2.

A.4 Proof of Proposition 3

For $\mu \in [\widehat{\mu}, 1)$, Proposition 1 applies. Consider next $\mu \in [\mu_{\text{pool}}, \widehat{\mu})$. Outsiders choose $\bar{a}$. Insiders choose $\bar{a}$ with probability m and 0 with probability $1 - m$. Bayes' rule gives

$$p(\bar{a}) = \frac{\lambda}{\lambda + (1 - \lambda)m}, \qquad p(0) = 0.$$

An insider mixes only if

$$\beta(\mu)p(\bar{a}) - \frac{1}{2}k_I\bar{a}^2 = 0,$$

so $p(\bar{a}) = k_I\bar{a}^2/[2\beta(\mu)]$. Equating the Bayesian posterior to this indifference posterior yields

$$m(\mu) = \frac{\lambda}{1-\lambda}\left(\frac{2\beta(\mu)}{k_I\bar{a}^2} - 1\right).$$

Lemma 3 establishes that this is a probability precisely in the stated interval and that it rises from zero to one as trust falls. The corresponding posterior falls from one to λ.

At the mixing posterior, the outsider's payoff is

$$\beta(\mu)p(\bar{a}) - \frac{1}{2}k_O\bar{a}^2 = \frac{1}{2}(k_I - k_O)\bar{a}^2 > 0,$$

so outsiders strictly prefer attacking. For any off-path $a' \in (0, \bar{a})$, equilibrium payoffs in the partial-pooling region are

$$U_I^* = 0, \qquad U_O^* = \frac{1}{2}(k_I - k_O)\bar{a}^2.$$

Given a candidate posterior π, type θ gains from a' if π exceeds its willingness threshold. Those thresholds are

$$\pi_I(a') = \frac{k_I(a')^2}{2\beta}, \qquad \pi_O(a') = \frac{k_O(a')^2 + (k_I - k_O)\bar{a}^2}{2\beta}.$$

Their difference is

$$\pi_O(a') - \pi_I(a') = \frac{(k_I - k_O)[\bar{a}^2 - (a')^2]}{2\beta} > 0.$$

The insider is therefore more willing to make every lower deviation. D1 assigns such deviations to the insider, so $p(a') = 0$, under which neither type gains.

For $\mu \leq \mu_{\text{pool}}$, both types choose $\bar{a}$ and Bayes' rule gives $p(\bar{a}) = \lambda$. The insider's equilibrium payoff is nonnegative because

$$\beta(\mu)\lambda \geq \frac{1}{2}k_I\bar{a}^2.$$

For a deviation $a' < \bar{a}$, type θ gains only if

$$\pi > \lambda - \frac{k_\theta[\bar{a}^2 - (a')^2]}{2\beta}.$$

The insider's threshold is strictly lower because $k_I > k_O$. D1 consequently assigns every lower deviation to the insider. With $p(a') = 0$, deviations yield only their signaling costs and are unprofitable. This also establishes the off-path belief used for unilateral moderation in Section 5.3.

Taking expectations over types and insider mixing gives

$$A(\mu) = \begin{cases} \lambda\sqrt{2\beta(\mu)/k_I}, & \mu \geq \widehat{\mu}, \\ \dfrac{2\lambda\beta(\mu)}{k_I\bar{a}}, & \mu_{\text{pool}} \leq \mu < \widehat{\mu}, \\ \bar{a}, & \mu < \mu_{\text{pool}}. \end{cases}$$

The schedule equals $\lambda\bar{a}$ at $\widehat{\mu}$ and $\bar{a}$ at μ_{pool}, hence is continuous. Its one-sided derivatives at $\widehat{\mu}$ are

$$A'(\widehat{\mu}^+) = \frac{\lambda\beta'(\widehat{\mu})}{k_I\bar{a}}, \qquad A'(\widehat{\mu}^-) = \frac{2\lambda\beta'(\widehat{\mu})}{k_I\bar{a}}$$

which proves $A'(\widehat{\mu}^-) = 2A'(\widehat{\mu}^+)$. This proves Proposition 3.

A.5 Stability reversal and a primitive-parameter construction

In the separating and insider-entry regions,

$$|A'(\mu)| = \begin{cases} \dfrac{\lambda|\beta'(\mu)|}{\sqrt{2k_I\beta(\mu)}}, & \mu > \widehat{\mu}, \\ \dfrac{2\lambda|\beta'(\mu)|}{k_I\bar{a}}, & \mu_{\text{pool}} < \mu < \widehat{\mu}. \end{cases}$$

Because $\beta(\widehat{\mu}) = k_I\bar{a}^2/2$, the interval

$$\frac{k_I\bar{a}}{2\lambda|\beta'(\widehat{\mu})|} < \kappa < \frac{k_I\bar{a}}{\lambda|\beta'(\widehat{\mu})|}$$

is nonempty and implies

$$\kappa|A'(\widehat{\mu}^+)| < 1 < \kappa|A'(\widehat{\mu}^-)|.$$

Thus insider entry can reverse local stability without any change in institutional recovery or in the effect of an individual attack.

The maintained response conditions hold on a nonempty open set of primitives. Let

$$w(\mu) = \omega_0 + \omega_1(1-\mu), \qquad \omega_0 > 0, \qquad \omega_1 > 0, \qquad \omega_0 + \omega_1 < 1.$$

For the construction used in Figure 3, set

$$\omega_0 = 0.02, \quad \omega_1 = 0.75, \quad k_I = 1, \quad \bar{a} = 0.4, \quad \lambda = 0.2, \quad \bar{\mu} = 0.95, \quad \kappa = 2.1.$$

Then

$$\widehat{\mu} \simeq 0.6865, \quad \mu_{\text{pool}} \simeq 0.2829, \quad \kappa_L \simeq 1.6677 < 2.1 < 3.2942 \simeq \kappa_H.$$

The equilibrium transition has fixed points

$$\mu_L \simeq 0.1100, \qquad \mu_T \simeq 0.4495, \qquad \mu_H \simeq 0.8837.$$

whereas the maximal outsider-only transition has only μ_H. At the equilibrium roots, $\kappa|A'|$ equals, respectively, 0, 1.78, and 0.52. All relevant inequalities are strict and continuous in the primitives, so they continue to hold on an open neighborhood of this example.

A.6 Proof of Proposition 4

Define the resting-point gap

$$D_\kappa(\mu) = \bar{\mu} - \mu - \kappa A(\mu).$$

An interior resting point is a zero of D_κ. At the boundaries, projection implies that $\mu = 0$ is a resting point if $D_\kappa(0) \le 0$, while $\mu = 1$ is a resting point if $D_\kappa(1) \ge 0$. At a differentiable interior point,

$$D'_\kappa(\mu) = \kappa|A'(\mu)| - 1.$$

First consider the maximal outsider-only schedule A^O. For every $\mu \le \widehat{\mu}$, $A^O(\mu) = \lambda\bar{a}$, and $\kappa < \kappa_H$ gives

$$D^O_\kappa(\mu) = \bar{\mu} - \mu - \kappa\lambda\bar{a} \ge \bar{\mu} - \widehat{\mu} - \kappa\lambda\bar{a} = \lambda\bar{a}(\kappa_H - \kappa) > 0.$$

Hence no outsider-only resting point lies at or below $\widehat{\mu}$. On $(\widehat{\mu}, 1)$, $A^O(\mu) = \lambda\sqrt{2\beta(\mu)/k_I}$. The curvature condition $2\beta\beta'' \leq (\beta')^2$ implies

$$\frac{d^2}{d\mu^2}\sqrt{\beta(\mu)} = \frac{2\beta(\mu)\beta''(\mu) - [\beta'(\mu)]^2}{4[\beta(\mu)]^{3/2}} \leq 0.$$

Thus A^O is concave and D_κ^O is convex on the separating region. Moreover,

$$D_\kappa^O(\widehat{\mu}) > 0, \qquad D_\kappa^O(1) = \bar{\mu} - 1 < 0,$$

For a convex function, the sublevel set is an interval. These endpoint signs therefore imply exactly one zero $\mu_H \in (\widehat{\mu}, 1)$. At that zero, $D_\kappa^{O\prime}(\mu_H) < 0$; otherwise convexity would imply $D_\kappa^O(1) \geq 0$. Hence

$$0 < F'(\mu_H) = 1 + \rho D_\kappa^{O\prime}(\mu_H) < 1,$$

so the high-trust resting point is locally stable.

Now use the equilibrium schedule. At the bottom of the insider-entry region,

$$D_\kappa(\mu_{\text{pool}}) = \bar{\mu} - \mu_{\text{pool}} - \kappa\bar{a} = \bar{a}(\kappa_L - \kappa) < 0,$$

whereas at its top,

$$D_\kappa(\widehat{\mu}) = \lambda\bar{a}(\kappa_H - \kappa) > 0.$$

In this region $A(\mu) = 2\lambda\beta(\mu)/(k_I\bar{a})$. The curvature condition $\beta''(\mu) \geq 0$ therefore implies

$$D_\kappa''(\mu) = -\frac{2\kappa\lambda}{k_I\bar{a}}\beta''(\mu) \leq 0.$$

Thus D_κ is concave. For a concave function, the superlevel set is an interval. The two endpoint signs therefore imply exactly one zero $\mu_T \in (\mu_{\text{pool}}, \widehat{\mu})$, where the gap crosses from negative to positive. Necessarily $D_\kappa'(\mu_T) > 0$; otherwise concavity would prevent the gap from becoming positive at $\widehat{\mu}$. Hence

$$F'(\mu_T) = 1 + \rho D_\kappa'(\mu_T) > 1,$$

so μ_T is unstable. The separating piece is the same as in the outsider-only benchmark and contributes the stable root μ_H.

Below μ_{pool}, attack is constant at $\bar{a}$. The unconstrained fixed point is $\bar{\mu} - \kappa\bar{a}$. The lower bound $\kappa > \kappa_L$ places it below μ_{pool}. Projection onto the state space therefore gives

$$\mu_L = \max\{0, \bar{\mu} - \kappa\bar{a}\} < \mu_{\text{pool}}.$$

Because $A' = 0$ in this region, $F'(\mu_L) = 1 - \rho \in (0, 1)$ for an interior root; the corresponding one-sided condition holds at the boundary. Thus μ_L is stable.

Finally, (κ_L, κ_H) is nonempty exactly when

$$\frac{\bar{\mu} - \mu_{\text{pool}}}{\bar{a}} < \frac{\bar{\mu} - \widehat{\mu}}{\lambda\bar{a}},$$

which rearranges to

$$\lambda < \frac{\bar{\mu} - \widehat{\mu}}{\bar{\mu} - \mu_{\text{pool}}}.$$

This proves Proposition 4.

A.7 Proof of Proposition 5

Begin at μ_H and apply a one-period shock $\varepsilon < 0$. Since the undisturbed transition leaves μ_H unchanged, the post-shock state is $\mu_1 = \mu_H + \varepsilon$, subject to the bounds of the state space. It remains in the healthy basin if and only if

$$\mu_H + \varepsilon > \mu_T \iff |\varepsilon| < \mu_H - \mu_T = s^*.$$

It enters the trapped basin if and only if $|\varepsilon| > s^*$. Equality places it at the unstable fixed point μ_T. The projected transition map is continuous and increasing because, wherever it is differentiable before projection,

$$F'(\mu) = 1 - \rho + \delta|A'(\mu)| > 0.$$

Together with the sign of D_κ on either side of the three resting points established in Proposition 4, monotonicity implies convergence to μ_H in the first case and to μ_L in the second. Because $\mu_L < \mu_T$,

$$\mu_H - \mu_L > \mu_H - \mu_T,$$

so the eventual decline exceeds the critical shock threshold s^* and therefore also exceeds shocks sufficiently close to and above that threshold.

It remains to establish the branch structure as κ varies. If $\kappa < \kappa_L$, then $D_\kappa(\mu_{\text{pool}}) > 0$ and, because $\kappa_L < \kappa_H$, $D_\kappa(\widehat{\mu}) > 0$. Below μ_{pool}, the gap is linear and is minimized at μ_{pool}, so it is positive throughout the full-pooling region. In the insider-entry region, concavity places the gap weakly above the chord joining its two positive endpoint values, so it is positive there as well. On the separating region, convexity together with $D_\kappa(\widehat{\mu}) > 0$ and $D_\kappa(1) < 0$ gives exactly one stable high-trust root. Thus, for $\kappa < \kappa_L$, only the high-trust resting point exists.

For $\kappa_L < \kappa < \kappa_H$, Proposition 4 gives the stable high and low points separated by μ_T.

Finally, suppose $\kappa > \kappa_H$. Then $D_\kappa(\widehat{\mu}) < 0$ and $D_\kappa(\mu_{\text{pool}}) < 0$. The kink-reversal condition gives

$$D'_{\kappa_H}(\widehat{\mu}^-) = 2\kappa_H S_H - 1 > 0.$$

Because $\beta'' \geq 0$ implies $|\beta'(\mu)| \geq |\beta'(\widehat{\mu})|$ for $\mu \in (\mu_{\text{pool}}, \widehat{\mu})$, it follows that

$$D'_\kappa(\mu) = \frac{2\kappa\lambda|\beta'(\mu)|}{k_I\bar{a}} - 1 \geq 2\kappa_H S_H - 1 > 0$$

throughout the insider-entry region whenever $\kappa \geq \kappa_H$. Hence, when $\kappa > \kappa_H$, the gap is increasing there and remains strictly below its negative value at $\widehat{\mu}$. In the separating region, D_κ is convex and is negative at both endpoints $\widehat{\mu}$ and 1; convexity places it below the chord between those endpoints, so it is negative throughout. No middle or high root remains. Since $\kappa > \kappa_H > \kappa_L$, the unique projected fixed point is

$$\mu_L = \max\{0, \bar{\mu} - \kappa\bar{a}\} < \mu_{\text{pool}},$$

and it is stable.

The healthy branch therefore reaches the mainstreaming threshold at κ_H, whereas the trapped branch reaches the bottom of the insider-entry region at $\kappa_L < \kappa_H$. Reversing the change that immediately preceded collapse is insufficient for recovery. This proves Proposition 5.