

# A Theory of Contact and In-Group Bias\*

Michael Sacks<sup>†</sup>

## Abstract

Why do some contact interventions produce lasting reductions in in-group bias while others fade once exposure ends? I develop a dynamic model of bias evolution in which temporary increases in out-group contact have permanent effects only if they push the population past an endogenous tipping point. The model highlights a central economic distinction between policies that increase contact and policies that raise the value of successful interaction. Because higher interaction stakes induce greater effort in out-group encounters, they reduce steady-state bias but also attenuate the societal feedback that sustains multiplicity. As a result, temporary contact has lasting effects only within a limited tipping region, whereas value-enhancing interventions can generate permanent change from a broader range of initial conditions, including settings where contact alone would fail. The framework yields testable implications for when temporary contact should have lasting effects and for how interventions should be targeted, timed, and sequenced.

March 27, 2026

*JEL Classification:* C73, D83, J15, Z13

**Keywords:** contact hypothesis, evolutionary dynamics, in-group bias, integration policy, tipping points

---

\*I am grateful for comments from participants at the 2026 ASREC conference. Holly Wagner provided excellent research assistance.

<sup>†</sup>David D. Reh School of Business, Clarkson University, Potsdam, New York 13699, [msacks@clarkson.edu](mailto:msacks@clarkson.edu).

*How can someone hate me, if they don't even know me?*

- Daryl Davis

## 1 Introduction

Contact interventions designed to reduce in-group bias exhibit a striking empirical regularity: programs with similar structures often differ sharply in whether their effects persist once the intervention ends. Lowe (2021) finds that cooperative cross-caste contact through a cricket program in India generates persistent and broad reductions in discrimination that extend beyond the contact setting. By contrast, Mousa (2020) studies a closely analogous soccer program pairing Christians and Muslims in post-ISIS Iraq and finds effects that remain confined to the contact setting. More broadly, some structured contact interventions show weak or null long-run effects (Scacco and Warren, 2018; Clochard et al., 2026), and meta-analyses conclude that treatment effects vary substantially across settings (Pettigrew and Tropp, 2006; Paluck et al., 2021; Lowe, 2025). A natural response is to attribute this variation to differences in implementation, populations, or historical context. But that response is largely descriptive: it does not explain when temporary contact should have lasting effects, nor does it provide a framework for designing interventions that make those effects more likely.

This paper offers a different explanation. I develop a dynamic model of in-group bias evolution in which the long-run effect of a contact intervention depends on where the population stands relative to an endogenous tipping point. When bias is sufficiently self-reinforcing, the dynamics admit two stable steady states, a low-bias regime and a high-bias regime, separated by an unstable threshold. Temporary contact generates permanent change only if it moves the population across that threshold; otherwise, bias declines during the intervention and reverts once contact returns to baseline. This threshold logic explains why the same program can succeed in one setting and fail in another, even when the intervention itself is broadly similar.

There is a population divided into two groups whose members are repeatedly matched for social interaction. The degree of assortative matching governs the frequency of out-group contact and is itself an object of policy. After matching, the quality of interaction depends on three forces: the effort individuals exert, their existing biases, and the broader social environment created by aggregate bias. Positive out-group interactions reduce bias, while negative interactions and repeated in-group contact reinforce it. These forces jointly determine the long-run evolution of bias in the

two groups.

A central feature of the model is that individuals choose how much effort to exert in out-group interactions, and that effort depends on the value of successful interaction. This creates a sharp distinction between policies that increase contact and policies that increase the returns to making contact succeed. Higher interaction stakes induce greater effort and therefore reduce steady-state bias, but they also weaken the societal feedback that sustains multiplicity. As a result, temporary contact can produce lasting change only in a limited tipping region, whereas value-enhancing interventions can generate permanent change from a broader range of initial conditions.

The societal feedback channel captures the idea that widespread bias shapes the quality of social interactions beyond the direct effect of the interacting individuals' own biases. When bias is pervasive, social norms and expectations can make positive out-group interaction difficult to sustain even for unbiased individuals; when bias is rare, surrounding norms of cooperation can temper the effect of individual bias. This link between aggregate bias and individual interaction outcomes is the source of the model's tipping dynamics and is consistent with theories of social norm enforcement (Bowles and Gintis, 2002) and preference falsification (Kuran, 1995).

This threshold logic is consistent with evidence that short-lived or weakly institutionalized contact interventions often produce weak or null long-run effects, while more sustained or structured contact can generate lasting change (Paluck et al., 2019; Scacco and Warren, 2018; Enos, 2014; Bazzi et al., 2019; Lowe, 2021; Ghosh et al., 2026). It also differs from canonical discrimination models, in which contact does not alter long-run outcomes, and from belief-correction models, in which positive contact reduces bias without threshold effects (Becker, 1957; Arrow, 1973; Bordalo et al., 2016; Bohren et al., 2025).

Where temporary contact fails, increasing the value of successful interaction can succeed. In the model, this is captured by an increase in  $V$ , the payoff from a successful interaction.  $V$  should therefore be interpreted not as contact itself, but as the material or institutional stake attached to making out-group interaction succeed. Policies that raise  $V$  induce greater effort in out-group interactions and can weaken, or even eliminate, the forces sustaining the high-bias regime. They can therefore produce permanent change from initial conditions under which temporary increases in contact would fail.

This interpretation is consistent with evidence from settings in which successful out-group inter-

action carries meaningful stakes. Rao (2019) documents persistent reductions in discriminatory behavior following integration in elite Delhi schools, with the largest effects in economic exchange tasks where interaction stakes are highest. Boisjoly et al. (2006) and Beaman et al. (2009) likewise study interventions that embed out-group interaction in environments organized around repeated cooperation rather than contact alone. In the model, such settings are naturally interpreted as increasing the value of successful interaction, not merely the frequency of contact.

The asymmetry between the two instruments also implies a sequencing result. A temporary value-enhancing intervention can first move the population into the region where temporary contact becomes permanently effective, after which a temporary increase in out-group contact can complete a transition that contact alone could not achieve. The two instruments are therefore complements rather than substitutes: policies that raise the value of successful interaction can prepare the conditions under which temporary contact has lasting effects.

The effort margin creates a central tension in the model. By making successful out-group interaction more likely, effort reduces steady-state bias. But because effort also dampens the link between aggregate bias and interaction outcomes, it weakens the feedback needed to sustain multiplicity and tipping dynamics. The same force that lowers bias in levels therefore makes permanent regime change through temporary contact harder to achieve. This yields a sharp empirical prediction: environments with higher stakes for successful interaction should exhibit lower steady-state bias, but a smaller region in which temporary increases in contact generate lasting change. In this sense, endogenous effort does not merely reduce bias; it changes which policies can produce permanent change.

The model also yields an asymmetry across groups. Because majority members experience more in-group interaction than minority members, they accumulate more in-group reinforcement and are therefore more biased in any interior steady state. This prediction provides an additional empirical signature of the mechanism and implies that increases in out-group contact reduce not only average bias but also cross-group differences in bias.

Relative to existing models of social interactions with aggregate feedback (Schelling, 1971; Kandori et al., 1993; Young, 1993), the paper's contribution is not simply to show that nonlinear feedback can generate tipping. Rather, it shows how tipping dynamics are reshaped when out-group interaction is mediated by endogenous effort. This yields two main implications: endogenous effort narrows the set of environments in which temporary contact can generate permanent change, and policies that

increase contact are not equivalent to policies that raise the value of successful interaction. The framework also implies a sequencing result and a cross-group asymmetry in steady-state bias.

The remainder of the paper is structured as follows. Section 2 develops the model. Sections 3 and 4 characterize the dynamics and establish the main results. Section 5 studies welfare and policy choice. Section 6 discusses policy implications, and Section 7 concludes.

## 2 The Model

The model combines repeated out-group interaction, aggregate social feedback, and endogenous effort. Out-group interaction can change future bias through its realized outcomes, while aggregate bias feeds back into those outcomes and creates the possibility of self-reinforcing dynamics. Endogenous effort then determines whether improved interaction quality and lasting change move together or in tension. This section introduces the environment and shows how these ingredients generate the law of motion for group-level bias.

Time is discrete and indexed by  $t = 0, 1, 2, \dots$ . There is a fixed population of individuals  $I = [0, 1]$  with Lebesgue measure  $\lambda$ . Each individual  $i \in I$  has a type  $\tau \in \{1, 2\}$ , where type 1 forms the majority. If  $I_\tau \subset I$  denotes the set of type- $\tau$  individuals, then  $q = \lambda(I_1) > \frac{1}{2}$ , and more generally  $q_\tau = \lambda(I_\tau)$ . Each individual also has an in-group bias indicator  $b_i^t \in \{0, 1\}$ , where  $b_i^t = 1$  denotes a biased individual and  $b_i^t = 0$  an unbiased one. These biases evolve over time. Let  $\pi_\tau^t = \frac{1}{\lambda(I_\tau)} \int_{I_\tau} b_i^t d\lambda \in [0, 1]$  denote the share of biased type- $\tau$  individuals at time  $t$ , and let  $\pi^t = (\pi_1^t, \pi_2^t)$  summarize the state of the economy. The initial state  $\pi^0$  is exogenously given. Aggregate bias is  $\bar{\pi}^t = q\pi_1^t + (1 - q)\pi_2^t$ , the overall prevalence of in-group bias in the population. The state of the economy is therefore summarized by the prevalence of bias within each group and in the population as a whole. When there is no confusion, I suppress time superscripts.

*Timing.* Each period has four stages: matching, effort choice, interaction realization, and bias updating.

- (1) Each individual  $i$  is matched with an individual  $j$ , at which point  $(\tau_i, b_i)$  is observed by  $j$  and  $(\tau_j, b_j)$  is observed by  $i$ . Let

$$\eta_\tau = \lambda(I_\tau) - \mu(1 - \lambda(I_\tau))$$

denote the probability that a type  $\tau$  individual is matched with another type  $\tau$  individual.<sup>1</sup>

---

<sup>1</sup>This specification satisfies market clearing:  $q(1 - \eta_1) = (1 - q)(1 - \eta_2)$ ; i.e., the share of type 1 individuals

The *match parameter*  $\mu \in [-1, \frac{1-q}{q}]$  guarantees that  $\eta_\tau \in [0, 1]$  for  $\tau = 1, 2$ . Random matching occurs when  $\mu = 0$ , assortative matching occurs when  $\mu < 0$ , and disassortative matching occurs when  $\mu > 0$ .

- (2) Upon individuals  $i$  and  $j$  being matched, each chooses an effort level  $e \in [0, 1]$  at cost  $\frac{1}{2}e^2$  to maximize current utility, described below.<sup>2</sup> Their efforts  $e_i$  and  $e_j$  are aggregated according to the function

$$s(e_i, e_j) = e_i + e_j - e_i e_j, \tag{1}$$

which determines the probability that effort alone yields a positive interaction, described next.<sup>3</sup>

- (3) The social interaction between  $i$  and  $j$  occurs. Three channels determine whether this interaction is positive or negative.
- (i) *Effort.* With probability  $s(e_i, e_j)$ , the interaction is positive.

If effort is not sufficient to yield a positive interaction, then the biases of the matched individuals and the broader social environment influence the outcome.

- (ii) *Individual feedback.*

If effort fails to yield a positive interaction, occurring with probability  $1 - s(e_i, e_j)$ , then the matched individuals' biases determine the probability that the interaction is resolved positively at the next stage. Given biases  $b_i$  and  $b_j$ , let

$$\beta(b_i, b_j) = \begin{cases} \beta_0 & \text{if } b_i + b_j = 0 \\ \beta_1 & \text{if } b_i + b_j = 1 \\ \beta_2 = 0 & \text{if } b_i + b_j = 2, \end{cases} \tag{2}$$

where  $0 = \beta_2 < \beta_1 < \beta_0 \leq 1$ . This ordering captures the idea that bias reduces the probability of successful interaction by weakening trust and coordination in out-group encounters (Arrow, 1973; Becker, 1957; Akerlof and Kranton, 2000). Mutually

---

matched with type 2 individuals must equal the share of type 2 individuals matched with type 1 individuals.

<sup>2</sup>The qualitative results are invariant to whether effort is chosen before or after matching. Under pre-match timing, individuals form expectations over the type-bias distribution of their match, yielding a Bayesian Nash equilibrium characterized by a  $4 \times 4$  linear system with the same properties as Proposition 1.

<sup>3</sup>The functional form is chosen for tractability rather than necessity. The qualitative results extend to more general aggregation functions that are increasing and concave in own effort and imply strategic substitutability in effort choices, as in standard team-production settings (Holmström, 1982).

biased pairs are therefore least likely to sustain interaction. The normalization  $\beta_2 = 0$  is adopted for convenience; allowing  $\beta_2 > 0$  but sufficiently small leaves the qualitative results unchanged.

For in-group matches, I suppress the bias notation and normalize the pair as  $(b_i, b_j) = (0, 0)$ , since the individual-bias channel is intended to capture frictions specific to out-group interaction.

- (iii) *Societal feedback.* If neither effort nor overcoming individual biases are sufficient to yield a positive interaction, occurring with probability  $(1 - s(e_i, e_j))(1 - \beta(b_i, b_j))$ , the outcome of the interaction is determined by the broader social environment. Let  $\bar{\pi}$  denote the average level of bias in the population. Conditional on reaching this stage, the probability that the interaction fails is given by  $\theta(\bar{\pi})$ , where  $\theta : [0, 1] \rightarrow [0, 1]$  is increasing and twice continuously differentiable with  $\theta(0) = 0$  and  $\theta(1) = 1$ . I interpret  $\theta(\bar{\pi})$  as a reduced form model of societal feedback that captures the idea that when bias is widespread, social norms and expectations make successful interaction less likely even after conditioning on the characteristics of the interacting pair. Section 2.1 offers two microfoundations that yield this reduced form specification.

All orderings of (i)–(iii) are equivalent, since only the composite probability of a positive interaction matters.

- (4) Biases are updated as follows. If  $\tau_i = \tau_j$  at time  $t$ , then with probability  $\varepsilon > 0$ ,  $b_i^{t+1} = 1$  and with complementary probability,  $b_i^{t+1} = b_i^t$ .<sup>4</sup> If  $\tau_i \neq \tau_j$  at time  $t$ , then a positive interaction yields  $b_i^{t+1} = 0$  and a negative interaction yields  $b_i^{t+1} = 1$ .

Hence, conditional on a match between  $i$  and  $j$ , the reduced-form probability of a positive interaction is

$$\rho_{b_i, b_j}(e_i, e_j) = s(e_i, e_j) + (1 - s(e_i, e_j))[\beta(b_i, b_j) + (1 - \beta(b_i, b_j))(1 - \theta(\bar{\pi}))]. \quad (3)$$

This process is illustrated in Figure 1.

*Payoffs.* Conditional on being matched with an individual  $j$  with type-bias pair  $(\tau_j, b_j)$ , individual

---

<sup>4</sup>Examples of in-group interactions reinforcing in-group bias include Tajfel et al. (1971), Tajfel and Turner (1979), Goette et al. (2006), Bernhard et al. (2006), and Chen and Li (2009).

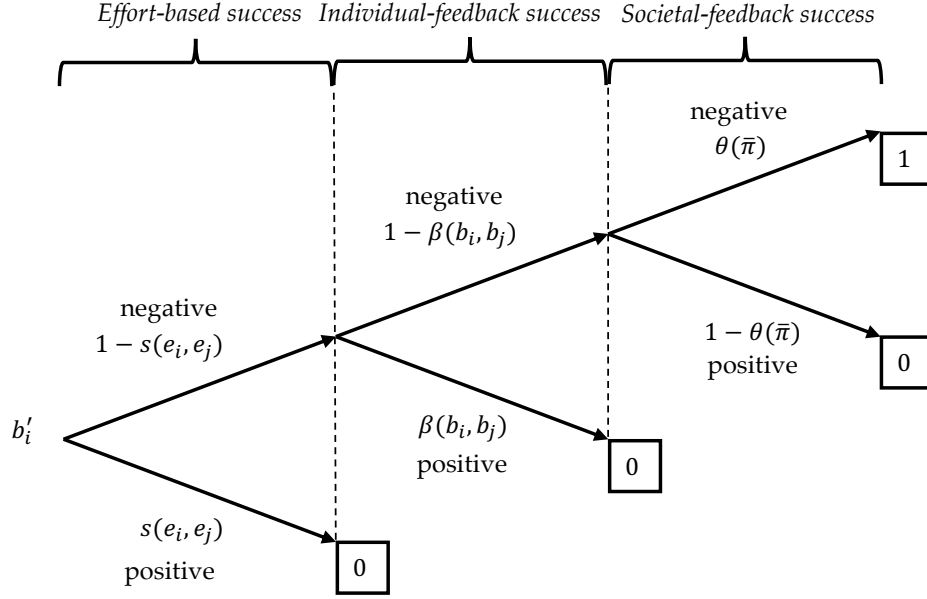


Figure 1: Diagrammatic representation of the probability of a positive interaction and future bias  $b'_i$  given current bias  $b_i$  and the matched individual's bias  $b_j$ . *Positive* corresponds to a positive interaction and *negative* to a negative interaction.

$i$ 's utility is

$$u(e_i, e_j; (\tau_i, b_i), (\tau_j, b_j)) = \rho_{b_i, b_j}(e_i, e_j)V - \frac{1}{2}e_i^2, \quad (4)$$

where  $V > 0$  is the value of a successful interaction. A higher  $V$  therefore corresponds to environments in which successful out-group interaction carries greater material or institutional stakes.

*Equilibrium Definitions.* At each date, individuals choose effort as a function of their own type-bias pair and that of their match. A Nash equilibrium of the within-period interaction game is a profile of such effort choices such that each individual's effort maximizes (4) given the effort choice of the matched individual. Across periods, the distribution of bias evolves according to the updating process above.

Let  $\pi^\infty = (\pi_1^\infty, \pi_2^\infty)$  denote the steady-state shares of biased individuals in each group, and let  $\bar{\pi}^\infty$  denote steady-state aggregate bias. Two benchmark steady states are of particular interest: the fully biased steady state, in which  $\bar{\pi}^\infty = 1$ , and the fully unbiased steady state, in which  $\bar{\pi}^\infty = 0$ .

## 2.1 Microfoundations of the Societal-Feedback Channel

While the model treats  $\theta(\bar{\pi})$  as a reduced-form object, it can be backed out from structural models of societal interactions. One microfoundation is community norm enforcement. Conditional on

reaching the societal-feedback stage, the probability of a successful interaction is mediated by the biases of a finite reference group drawn from the population.

Consider a model of community norm enforcement. Conditional on reaching the societal-feedback stage, the interaction is mediated by a reference group of  $k$  members drawn independently from the population. If a majority of the  $k$  members drawn are biased, then the social interaction is deemed a failure and if a majority are unbiased, then the social interaction is successful. Thus,

$$\theta(\bar{\pi}) = \Pr\left(\text{Binomial}(k, \bar{\pi}) > \frac{k}{2}\right) = \sum_{j=\lfloor k/2 \rfloor + 1}^k \frac{k!}{j!(k-j)!} \bar{\pi}^j (1-\bar{\pi})^{k-j}.$$

For all  $k$ ,  $\theta(0) = 0$  and  $\theta(1) = 1$ . At  $k = 1$ ,  $\theta(\bar{\pi}) = \bar{\pi}$  and at  $k = 2$ ,  $\theta(\bar{\pi}) = \bar{\pi}^2$ .<sup>5</sup>

For all  $k \geq 3$ ,  $\theta$  is sigmoidal, and  $\sup_{\bar{\pi}} \frac{d\theta(\bar{\pi})}{d\bar{\pi}}$  is strictly increasing in  $k$ . Thus, larger reference groups generate sharper threshold behavior in societal feedback. Moreover, as  $k \rightarrow \infty$ ,  $\theta(\bar{\pi})$  converges pointwise to a step function at  $\bar{\pi} = \frac{1}{2}$ , so the size of the relevant social reference group governs the strength of societal feedback. The formal link between this curvature and the existence of multiple steady states is developed below.<sup>6</sup>

A preference-falsification environment in the spirit of Kuran (1995) can also generate the reduced form specification above. These microfoundations capture the broader idea, emphasized in the literature on trust and social capital, that the social environment shapes the expressed quality of bilateral interactions (Bowles and Gintis, 2002) and are illustrative rather than essential. They offer structural environments in which the reduced-form societal-feedback channel can arise, but none of the analytical results depends on any specific functional form or underlying mechanism of social interaction. The proofs are stated for a general reduced-form  $\theta(\bar{\pi})$ .

## 2.2 Preliminary Results

Before characterizing the full dynamics, I establish three preliminary results. I first solve for the effort levels that constitute a Nash equilibrium of the social interaction stage game. I then characterize the implied average effort gap between minority and majority individuals. Lastly, I highlight the role of randomness and in-group reinforcement in the evolution of bias.

---

<sup>5</sup>For even  $k$ , ties arise with positive probability. Any tie-breaking rule preserves the qualitative conclusions emphasized here.  $\theta(\bar{\pi})$  remains increasing,  $k = 1$  yields linear feedback,  $k = 2$  yields convex feedback, and larger  $k$  produces increasingly sharp threshold behavior. The exact location of the inflection point may depend on how ties are resolved, but none of the analytical results relies on that detail.

<sup>6</sup>The conditions for multiple steady-states are outlined in Proposition 2. When the remaining conditions of Proposition 2 are satisfied, there always exists a  $k$  large enough to generate multiple steady states and tipping dynamics.

Recall that, conditional on being matched with a type  $\tau_j$  individual with bias  $b_j$ , individual  $i$  chooses the effort level that maximizes (4).

**Lemma 1.** *Suppose that pairs  $(\tau_i, b_i)$  and  $(\tau_j, b_j)$  are matched. The effort levels constituting a Nash equilibrium of the social interaction stage game are given by*

$$e_i^*(b_i, b_j) = \frac{(1 - \beta(b_i, b_j))\theta(\bar{\pi})V}{1 + (1 - \beta(b_i, b_j))\theta(\bar{\pi})V}. \quad (5)$$

*Thus, effort levels are increasing in the value of a successful interaction, the match-specific probability of failure, and the societal-feedback function  $\theta(\bar{\pi})$ .*

Lemma 1 implies that equilibrium effort is increasing in the match-specific probability of a negative interaction,  $1 - \beta(b_i, b_j)$ , the societal-feedback term  $\theta(\bar{\pi})$ , and the value of a successful interaction  $V$ . Interpreting effort as costly preventive investment, individuals invest more when baseline success is less likely. This logic is consistent with evidence that cooperation rises when punishment or exclusion makes opportunism more consequential (Fehr and Gächter, 2000; Cinyabuguma et al., 2005; Gürer et al., 2006). Related field evidence shows that incentive design, including team-based pay, can mitigate productivity losses associated with inter-group divisions (Hjort, 2014).

Lemma 1 also allows a comparison of average effort between minority (type 2) and majority (type 1) individuals. Given (5), the average effort put forth by a type  $\tau$  is

$$(1 - \eta_\tau)[\pi_1\pi_2e^*(1, 1) + ((1 - \pi_1)\pi_2 + (1 - \pi_2)\pi_1)e^*(1, 0) + (1 - \pi_1)(1 - \pi_2)e^*(0, 0)] + \eta_\tau e^*(0, 0),$$

where  $e_i^*(b_i, b_j) = e^*(b_i, b_j)$  as  $i$  enters only through  $b_i$  and  $e^*(1, 0) = e^*(0, 1)$ .

**Corollary 1.** *Average effort is weakly higher among minority individuals than majority individuals, strictly so when  $\mu > -1$  and in-group bias is present in the population ( $\bar{\pi} > 0$ ).*

### 2.2.1 The Role of Randomness and In-Group Reinforcement

Let  $s^*(b_i, b_j) \equiv s(e_i^*(b_i, b_j), e_j^*(b_i, b_j))$  denote the aggregate effort in the NE of the social interaction stage game. Now, (3) can be written as a function only of biases  $b_i, b_j$ :

$$\rho_{b_i, b_j} = s^*(b_i, b_j) + (1 - s^*(b_i, b_j))[\beta(b_i, b_j) + (1 - \beta(b_i, b_j))(1 - \theta(\bar{\pi}))].$$

For a randomly drawn type  $\tau$  individual, the expected probability of a positive interaction is

$$\bar{\rho}_\tau(\pi) = (1 - \pi_\tau) [(1 - \pi_{-\tau})\rho_{0,0} + \pi_{-\tau}\rho_{0,1}] + \pi_\tau [(1 - \pi_{-\tau})\rho_{1,0} + \pi_{-\tau}\rho_{1,1}],$$

which implies that

$$\pi_\tau^{t+1} = (1 - \eta_\tau) (1 - \bar{\rho}_\tau(\pi^t)) + \eta_\tau (\pi_\tau^t + (1 - \pi_\tau^t)\varepsilon)$$

Hence,

$$\Delta\pi_\tau = (1 - \eta_\tau) [1 - \bar{\rho}_\tau(\pi) - \pi_\tau] + \eta_\tau(1 - \pi_\tau)\varepsilon. \quad (6)$$

The role of randomness ( $\varepsilon > 0$ ) is immediately apparent. It is necessary for out-group contact to affect steady-state bias in the population. When  $\varepsilon = 0$ , out-group interaction, and thus changes in the frequency of out-group contact, affect only the speed of adjustment, not the long-run level of bias.

I interpret  $\eta_\tau(1 - \pi_\tau)\varepsilon$  as a reduced-form socialization or identity-reinforcement mechanism. Conditional on in-group interaction, a fraction  $\varepsilon$  of initially unbiased individuals adopt biased beliefs or norms through exposure to in-group narratives, conformity incentives, and local transmission. This interpretation is consistent with economic models of cultural transmission in which in-group socialization and assortative matching yield persistent group-specific attitudes in the population (Bisin and Verdier, 2000, 2001; Carvalho and Sacks, 2024; Carvalho et al., 2024a).<sup>7</sup> Identity-based frameworks similarly imply that social categories and moral identity shape behavior and beliefs, and that in-group interactions can sustain and reinforce group-consistent worldviews (Akerlof and Kranton, 2000; Bénabou and Tirole, 2011). Because homophily and segregation raise the frequency of in-group interactions, even small in-group reinforcement rates can have quantitatively important implications for long-run bias (Currarini et al., 2009; Gentzkow and Shapiro, 2011).

### 3 Mechanism Decomposition

Before analyzing the full model, I decompose it into its three constituent mechanisms: effort, individual feedback, and societal feedback, and study the long-run behavior that each generates in isolation. This decomposition clarifies which aspects of the dynamics operate through each channel.

#### 3.1 Effort

I begin with the effort channel in isolation. Can effort alone generate persistent effects from temporary increases in out-group contact? It cannot. Although effort responds to and shapes the probability of a positive interaction, the effort-only benchmark admits a unique globally stable

---

<sup>7</sup>See Shayo (2020) for a survey of empirical and experimental evidence on in-group bias and in-group conformity.

steady state. Temporary changes in  $\mu$  or  $V$  therefore affect the transition path but not the long-run outcome.

Fix  $\beta(b_i, b_j) = 0$  and  $\theta(\bar{\pi}) = 1$  for all  $b_i, b_j$ , and  $\bar{\pi}$ . In this case, the probability of a positive interaction is  $\rho_{b_i, b_j}(e_i, e_j) = s(e_i, e_j)$ . By (5),

$$e_i^* = \frac{V}{1+V} \equiv e^*.$$

Evaluating the probability of a positive interaction at  $(e^*, e^*)$  yields

$$s(e^*, e^*) = \frac{V(2+V)}{(1+V)^2} \equiv s^* \in (0, 1).$$

Thus, effort alone generates a constant probability of successful interaction, increasing in  $V$ , but independent of the distribution of bias in the population.

The evolution of each  $\pi_\tau$  is then

$$\Delta\pi_\tau = (1 - \eta_\tau) [(1 - \pi_\tau)(1 - s^*) - \pi_\tau s^*] + \eta_\tau(1 - \pi_\tau)\varepsilon, \quad (7)$$

which yields the unique steady-state values

$$\pi_\tau^\infty = \begin{cases} 1 - \frac{(1-\eta_\tau)s^*}{1-\eta_\tau(1-\varepsilon)} & \text{if } \varepsilon > 0 \text{ or } \mu > -1 \\ \pi_\tau^0 & \text{if } \varepsilon = 0 \text{ and } \mu = -1. \end{cases} \quad (8)$$

If  $\varepsilon = 0$  and  $\mu = -1$ , matching is purely assortative and biases remain constant. Otherwise,  $\Delta\pi_\tau$  is linear and strictly decreasing in  $\pi_\tau$ , with  $\Delta\pi_\tau(0) > 0 > \Delta\pi_\tau(1)$ , so the dynamics converge globally to (8) from any initial condition. The effort-only benchmark therefore admits a unique globally stable steady state.

This benchmark yields two simple comparative statics. First, when  $\varepsilon > 0$ , increasing out-group contact (higher  $\mu$ ) strictly reduces steady-state bias by shifting weight away from in-group reinforcement and toward out-group interaction. Second, when  $\mu > -1$ , increasing  $V$  also reduces steady-state bias by raising equilibrium effort and therefore the probability of successful interaction. Moreover, when  $\varepsilon > 0$  and  $\mu > -1$ , the majority is more biased in steady state than the minority, since  $\eta_1 > \eta_2$  implies greater in-group reinforcement for the majority.

The key limitation of the effort-only benchmark is that temporary interventions do not have lasting effects. A temporary increase in  $\mu$  or  $V$  lowers bias while it is in place, but once the parameter returns to baseline, the dynamics converge back to the original steady state. This benchmark

therefore captures environments in which contact or higher interaction stakes improve outcomes on impact, but cannot generate persistent regime change. That contrast will matter below. Persistence requires not only that interaction outcomes affect future bias, but also that aggregate bias feed back into those outcomes strongly enough to create threshold dynamics.

### 3.2 Individual Feedback

I next isolate the individual-bias channel by shutting down effort and societal feedback. Unlike the effort-only benchmark, this environment can admit biased, unbiased, or interior steady states depending on parameter values. But for any fixed parameter configuration and  $\mu > -1$ , the dynamics still converge to a unique steady state. Thus, the individual-bias channel can change the location of the unique steady state, but it does not generate the threshold dynamics required for temporary interventions to have lasting effects. Temporary changes in out-group contact affect only the transition path, and  $V$  has no effect because it operates only through effort.

Fix  $s(e_i, e_j) = 0$  and  $\theta(\bar{\pi}) = 1$  for all  $e_i, e_j$ , and  $\bar{\pi}$ . In this case, the probability of a positive interaction is given by  $\rho_{b_i, b_j}(e_i, e_j) = \beta(b_i, b_j)$  and the evolution of  $\pi_\tau$  is given by

$$\Delta\pi_\tau = (1 - \eta_\tau) \left[ (1 - \pi_\tau) \left( (1 - \pi_{-\tau})(1 - \beta_0) + \pi_{-\tau}(1 - \beta_1) \right) - \pi_\tau(1 - \pi_{-\tau})\beta_1 \right] + \eta_\tau(1 - \pi_\tau)\varepsilon. \quad (9)$$

Setting  $\Delta\pi_\tau = 0$  and solving for  $\pi_\tau$  yields the steady-state mapping

$$F_\tau(\pi_{-\tau}; \varepsilon) = \frac{(1 - \eta_\tau)[(1 - \pi_{-\tau})(1 - \beta_0) + \pi_{-\tau}(1 - \beta_1)] + \eta_\tau\varepsilon}{(1 - \eta_\tau)[(1 - \pi_{-\tau})(1 - \beta_0 + \beta_1) + \pi_{-\tau}(1 - \beta_1)] + \eta_\tau\varepsilon}, \quad (10)$$

so

$$\frac{\partial F_\tau(\pi_{-\tau}; \varepsilon)}{\partial \pi_{-\tau}} = \frac{(1 - \eta_\tau)\beta_1[(1 - \eta_\tau)(1 - \beta_1) + \eta_\tau\varepsilon]}{((1 - \eta_\tau)[(1 - \pi_{-\tau})(1 - \beta_0 + \beta_1) + \pi_{-\tau}(1 - \beta_1)] + \eta_\tau\varepsilon)^2}. \quad (11)$$

The biased steady state always exists as  $\rho_{1,1}(0, 0) = 0$  implies no endogenous force reduces bias when it is universal. When  $\varepsilon = 0$  and  $\mu = -1$ , steady states are symmetric and determined by the  $\beta_0/\beta_1$  partition. If  $\beta_1 > \frac{1}{2}$  and  $\beta_0 < 1$ , then there is a unique interior steady state

$$\pi_1^\infty = \pi_2^\infty = \frac{1 - \beta_0}{2\beta_1 - \beta_0}, \quad (12)$$

to which almost all initial conditions converge. When  $\beta_1 \leq \frac{1}{2}$ , out-group interactions are insufficiently bias-reducing and the biased steady state is globally attracting. Introducing  $\varepsilon > 0$  eliminates the unbiased steady state and shifts any interior steady state upward. Full bias becomes globally

stable whenever out-group updating near  $\pi = (1, 1)$  is locally too weak to offset in-group reinforcement. The full characterization is in the Appendix.

Except under full assortativity ( $\mu = -1$ ), steady states are independent of  $\mu$ . Varying  $\mu$  affects the speed of convergence but not the steady state. When  $\beta_1 \leq \frac{1}{2}$ , out-group interactions either reduce bias only weakly or reinforce it, and the population converges to full bias. In this region, increasing out-group contact increases the speed at which bias accumulates rather than mitigating it. This feature is consistent with empirical evidence showing that contact reduces prejudice primarily when interactions are cooperative and generate positive informational signals, and may instead entrench bias when interactions are conflict-oriented or unstructured (Allport, 1954; Pettigrew and Tropp, 2006; Lowe, 2021; Enos, 2014).

When  $\varepsilon > 0$ , the dynamics incorporate in-group reinforcement in addition to out-group updating. Even when out-group interactions are mildly positive, in-group contact generates drift toward bias. The unbiased steady state therefore disappears except under full assortativity, and any interior steady state shifts upward. Long-run outcomes are determined by the balance between positive bias-reducing out-group interactions and bias reinforcement near the fully biased state.<sup>8</sup> Full bias becomes globally stable whenever out-group updating is locally too weak to offset in-group reinforcement.

This mechanism resonates with models of cultural transmission and identity formation in which in-group socialization sustains persistent group-specific beliefs (Bisin and Verdier, 2000, 2001; Akerlof and Kranton, 2000; Bénabou and Tirole, 2011; Carvalho and Sacks, 2024). Recent empirical evidence further suggests that contact alone is insufficient to reduce polarization when informational environments amplify in-group narratives. Studies of media consumption and online platforms document how selective exposure and social reinforcement can entrench partisan attitudes even amid out-group interaction (Gentzkow and Shapiro, 2011; Allcott and Gentzkow, 2017; Boxell et al., 2017). Experimental evidence likewise shows that out-group interactions may have limited or even backfiring effects when interactions trigger defensive identity responses (Bail et al., 2018).<sup>9</sup>

Because the in-group interaction rates  $\eta_1$  and  $\eta_2$  generally differ (except under full assortativity,  $\mu = -1$ ), the interior steady state typically features asymmetric bias across types when  $\varepsilon > 0$ . This raises a natural comparative question: when there is an interior steady state, is the majority more

---

<sup>8</sup>The full characterization of this trade-off is in Proposition A1 in the Appendix.

<sup>9</sup>Additional theoretical evidence includes Carvalho, Koyama and Sacks (2017) and Carvalho et al. (2024b).

or less biased than the minority in the long run?

**Corollary 2.** *Suppose  $\mu > -1$ ,  $\varepsilon > 0$  and an interior steady state exists. Then,  $\pi_1^\infty > \pi_2^\infty$ .*

For any  $\mu > -1$ , the majority's in-group interaction rate exceeds that of the minority's:  $\eta_1 - \eta_2 = (2q - 1)(1 + \mu) > 0$ , so the majority experiences less out-group contact than the minority. When  $\varepsilon > 0$ , within-type interactions generate drift toward bias, so a higher  $\eta_\tau$  shifts type  $\tau$ 's steady-state response upward. As a consequence, whenever the limiting state is interior, the majority ends up strictly more biased than the minority.

Relatedly, when  $\varepsilon > 0$ , increasing  $\mu$  generally lowers steady-state bias. Because in-group reinforcement operates through  $\eta_\tau$ , a rise in  $\mu$  weakens the drift toward bias and strengthens the relative force of out-group contact-based bias reduction. As a result, parameters for which the biased steady state is locally stable can cross into a region in which an interior steady state exists. However, this change is not path-dependent. The steady states are pinned down by the contemporaneous value of  $\mu$ . If  $\mu$  reverts to its original level and the parameter configuration again implies that the biased steady state is the unique steady state, the dynamics converge back to that steady state.

The absence of path dependence in the individual-bias channel helps rationalize a recurring empirical finding. Many contact interventions generate short-run improvements in attitudes that dissipate once the intervention ends (Scacco and Warren, 2018; Paluck et al., 2019, 2021). Temporary increases in  $\mu$  alter the transition path but not the steady state unless they permanently change reinforcement. Bias reverts once interaction patterns return to baseline. Durable effects arise only when contact reshapes the underlying reinforcement environment itself.

### 3.3 Societal Feedback

I now isolate the societal-feedback channel by shutting down effort and individual feedback. Can aggregate social feedback alone generate persistent effects from increases in out-group contact? The answer now depends on the shape of  $\theta(\bar{\pi})$ . When  $\theta(\bar{\pi})$  is linear or convex, the dynamics do not generate threshold-based path dependence from temporary changes in out-group contact. When  $\theta(\bar{\pi})$  is sufficiently sigmoidal, threshold effects can emerge, giving rise to multiple steady states and distinct basins of attraction.

Fix  $\beta(b_i, b_j) = s(e_i, e_j) = 0$  for all  $e_i, e_j, b_i$  and  $b_j$ . In this case, the probability of a positive

interaction is given by  $\rho_{b_i, b_j}(e_i, e_j) = 1 - \theta(\bar{\pi})$ , yielding the evolutionary dynamic

$$\Delta\pi_\tau = (1 - \eta_\tau)(\theta(\bar{\pi}) - \pi_\tau) + \eta_\tau(1 - \pi_\tau)\varepsilon. \quad (13)$$

Hence, steady states are given by the solutions to the system

$$\pi_\tau = \gamma(\eta_\tau, \varepsilon)\theta(\bar{\pi}) + [1 - \gamma(\eta_\tau, \varepsilon)] \times 1, \quad (14)$$

where

$$\gamma(\eta_\tau, \varepsilon) = \frac{1 - \eta_\tau}{1 - \eta_\tau(1 - \varepsilon)} \equiv \gamma_\tau \in (0, 1),$$

which is strictly decreasing in both  $\eta_\tau$  and  $\varepsilon$  (and thus increasing in  $\mu$ ). Define  $\bar{\gamma} = q\gamma_1 + (1 - q)\gamma_2$ .

This system can be reduced by studying the aggregate prevailing bias  $\bar{\pi}$ . Using (14) to compute  $\bar{\pi} = q\pi_1 + (1 - q)\pi_2$  yields

$$\bar{\pi} = \bar{\gamma}(\theta(\bar{\pi}) - 1) + 1. \quad (15)$$

Thus, the steady states directly depend on the shape of  $\theta(\bar{\pi})$ . I consider three possibilities for  $\theta$ : linear, convex, and sigmoidal.<sup>10</sup> When linear,  $\theta(\bar{\pi}) = \bar{\pi}$ . When sigmoidal, I assume

$$\lim_{\bar{\pi} \rightarrow 0^+} \frac{d\theta(\bar{\pi})}{d\bar{\pi}} < 1, \quad \lim_{\bar{\pi} \rightarrow 1^-} \frac{d\theta(\bar{\pi})}{d\bar{\pi}} < 1.$$

These assumptions are consistent with Section 2.1 for  $k = 1$  (linear) and  $k$  large (sigmoidal).

Thus, the steady states directly depend on the shape of  $\theta(\bar{\pi})$ . When  $\theta$  is linear or convex, the dynamics do not generate threshold-based path dependence from temporary changes in out-group contact; steady states are either unique or, in the knife-edge linear case, pinned down by initial conditions.<sup>11</sup> The novelty arises when feedback is sufficiently sigmoidal. In that case, small differences in aggregate bias can be amplified rather than dampened, generating tipping points and multiple steady states.<sup>12</sup>

**Proposition 1.** *Fix  $s(e_i, e_j) = 0$  and  $\beta(b_i, b_j) = 0$  for all  $e_i, e_j, b_i,$  and  $b_j$ . Suppose  $\varepsilon > 0$  and  $\mu > -1$ .*

<sup>10</sup>The community norm enforcement foundation in Section 2.1 provides a concrete case,  $k = 1$ ,  $k = 2$ , and  $k \geq 3$ , respectively, for these three shapes.

<sup>11</sup>The full characterization is in Proposition A1 in the Appendix.

<sup>12</sup>Allowing  $\theta$  to be concave does not substantially add anything to the results. When  $\varepsilon = 0$ , there is convergence to the biased steady state from almost all initial conditions and when  $\varepsilon > 0$ , the results are identical to the linear case. See the proof for details

- (i) If  $\theta(\bar{\pi})$  is sigmoidal and  $\frac{d\theta(\bar{x})}{d\bar{\pi}} > \bar{\gamma}^{-1}$  for some  $x \in (0, 1)$ , then there are three steady states: a stable low-bias interior steady state  $\bar{\pi}^*$ , an unstable tipping point  $\bar{\pi}_U^* > \bar{\pi}^*$ , and the biased steady state at  $(1, 1)$ .
- (ii) Under the conditions of (i), if  $\bar{\pi}^0 \in (\bar{\pi}_U^*(\mu^0), \bar{\pi}_U^*(\frac{1-q}{q}))$ , then there exists a  $\mu' > \mu^0$  such that temporarily increasing  $\mu^0$  to  $\mu'$  yields a persistent reduction in aggregate bias.

The proposition delivers the minimal mechanism through which temporary contact can generate persistent change. Sigmoidal societal feedback establishes a tipping point, and contact need only push society across it. The intervention need not be permanent; it need only last long enough to move the economy across the tipping point into the low bias basin of attraction. This rationalizes why some contact interventions produce durable change while others do not. Contact has lasting effects only when it shifts the society across the relevant tipping point. Figure 2 illustrates this mechanism.

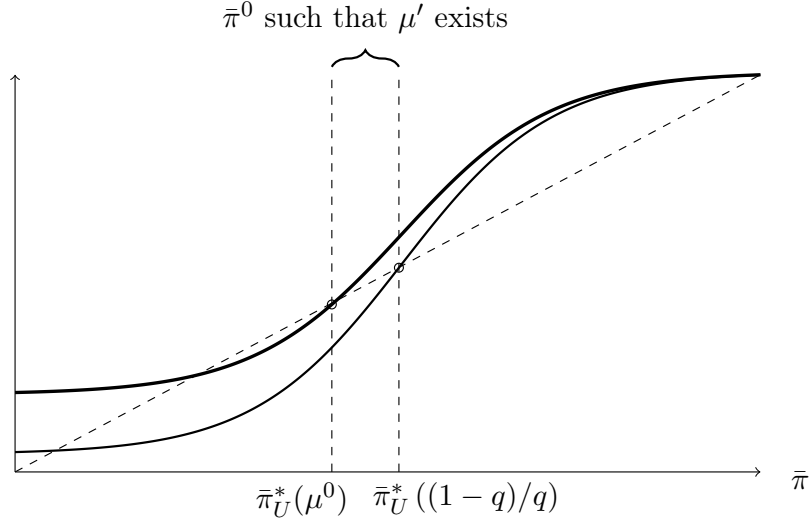
The analysis thus far isolates the minimal mechanism required for temporary contact to generate permanent change. In practice, however, effort, individual feedback, and societal feedback operate simultaneously. I now reintroduce all three channels and characterize the full dynamics of the model, showing how their interaction determines when contact reduces bias and when it instead reinforces high-bias outcomes.

## 4 Dynamics of the Full Model

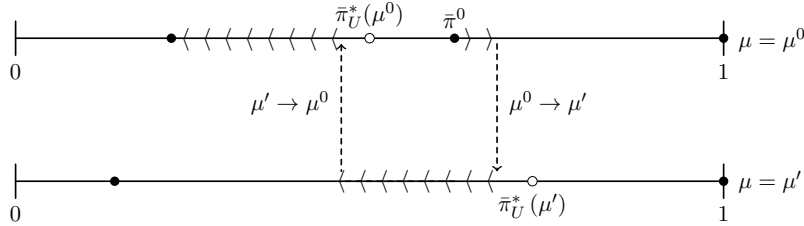
I now reintroduce all three channels: effort, individual feedback, and societal feedback. The full model nests each benchmark in Section 3, but their interaction generates new forces absent in isolation. Effort counteracts in-group bias. As a result, the evolution of bias may be either globally convergent to a unique steady state or threshold-driven, depending on parameters. This section characterizes steady states of the full system and identifies when temporary increases in out-group contact can generate permanent change.

**Lemma 2.** *Conditional on individuals  $i$  and  $j$  matching with biases  $b_i$  and  $b_j$ , the probability of a positive interaction is monotonic in  $\bar{\pi}$  iff*

$$V \leq \hat{V}(b_i, b_j) \equiv \frac{1}{1 - \beta(b_i, b_j)}.$$



(a) Evolution of  $\bar{\pi}$  under  $\mu = \mu^0$  and  $\mu = \frac{1-q}{q}$ .



(b) Transition path given initial biases  $(\pi_1^0, \pi_2^0)$  and initial average bias  $\bar{\pi}^0 = q\pi_1^0 + (1-q)\pi_2^0$  and temporary increase in  $\mu$  from  $\mu^0$  to  $\mu'$ .

Figure 2: Graphical illustration of Proposition 1.

Lemma 2 highlights the key interaction between effort, individual feedback, and societal feedback. In the societal-feedback benchmark of Section 3.3, higher aggregate bias mechanically reduced the probability of a positive interaction. Incorporating effort, individuals respond to this bias. When  $V$  is small, effort is weak and the probability of success declines monotonically in  $\bar{\pi}$ . When  $V$  is sufficiently large, effort responds strongly enough that increases in  $\bar{\pi}$  induce additional investment, dampening the effect of aggregate bias on interaction outcomes and potentially rendering success probabilities non-monotonic in  $\bar{\pi}$ . High-value environments therefore attenuate the propagation of bias. This has a direct policy implication.  $V$ -based interventions may be most effective precisely in highly biased societies, where the incentive to invest in effort is greatest. This reverses the intuition from the benchmark, where high-bias environments are exactly the ones most resistant to intervention.

Effort also eliminates the fully biased steady state.

**Lemma 3.** *Suppose that  $\mu > -1$ . The biased steady state does not exist. Additionally, the unbiased steady state exists iff  $\varepsilon = 0$ .*

In the individual feedback and societal feedback benchmarks, the fully biased steady state was self-sustaining because no endogenous force reduced bias once it was universal. With effort, however, biased individuals still invest to secure positive interactions. As long as  $\mu > -1$ , out-group contact occurs with positive probability, and effort generates positive interactions that reduce bias. The biased steady state therefore disappears.

To characterize the multiplicity of steady states, define the *slope envelope* as

$$\tilde{p}'(\bar{\pi}) = \frac{d\theta(\bar{\pi})}{d\bar{\pi}} \times \max_{\beta \in [0,1]} \frac{(1-\beta)|1 - (1-\beta)\theta(\bar{\pi})V|}{(1 + (1-\beta)\theta(\bar{\pi})V)^3},$$

and set  $\tilde{P}' := \sup_{\bar{\pi} \in [0,1]} \tilde{p}'(\bar{\pi})$ , which characterizes the maximal steepness of the inflection point when societal feedback is self-reinforcing (sigmoidal). Additionally, define

$$\bar{\Delta} = \sup_{\bar{\pi} \in [0,1]} \max_{\beta, \beta' \in \{0, \beta_1, \beta_0\}} |m_\beta(\bar{\pi}) - m_{\beta'}(\bar{\pi})| \geq 0,$$

where  $m_\beta(\bar{\pi}) = 1 - \rho_{b_i, b_j}$  for  $\beta = \beta(b_i, b_j)$ .

The term  $\bar{\Delta}$  is the largest gap in interaction failure probabilities across all levels of aggregate bias and bias configurations.

**Proposition 2.** *Suppose  $\mu > -1$ . The following statements hold.*

- (i) *If  $2\bar{\Delta} + \tilde{P}' < \gamma_2^{-1}$ , then there is a unique steady state  $(\pi_1^\infty, \pi_2^\infty) \in [0, 1]^2$ .*
- (ii) *Suppose  $\varepsilon > 0$  and  $\frac{dm_{\beta_0}(\bar{\pi})}{d\bar{\pi}}$  is single-peaked on  $(0, 1)$  with  $\bar{\pi}_M = \arg \max_{\bar{\pi}} \frac{dm_{\beta_0}(\bar{\pi})}{d\bar{\pi}} \in (0, 1)$ . There exists a  $\hat{\Delta}$  such that if  $\tilde{p}'(0), \tilde{p}'(1) < \bar{\gamma}^{-1} < \frac{dm_{\beta_0}(\bar{\pi}_M)}{d\bar{\pi}}$  and  $\bar{\Delta} < \hat{\Delta}$ , then there are at least three interior steady states.*

Let  $\bar{\pi}_1^*$ ,  $\bar{\pi}_2^*$ , and  $\bar{\pi}_3^*$  denote the steady state aggregate bias levels. Here,  $\bar{\pi}_2^*$  is the tipping point, analogous to  $\bar{\pi}_U^*$  from Section 3.3. Because  $\tilde{P}'$  is decreasing in  $V$ , uniqueness in statement (i) is more easily satisfied when  $V$  is large, individuals then respond aggressively to rising bias, dampening the aggregate feedback required for multiplicity. Effort therefore attenuates the nonlinearities necessary

for tipping dynamics. Invoking the community norm enforcement interpretation of  $\theta(\bar{\pi})$  from Section 2.1, for any  $V > 0$  and maintaining  $\varepsilon >$  and  $\bar{\Delta} < \hat{\Delta}$ , there always exists  $k$  sufficiently large to satisfy the slope conditions of (ii). The latter condition is least restrictive precisely when tipping dynamics are most pronounced (see footnote 21 in the Appendix).

Let  $\Omega = (\theta, \beta_0, V, q, \mu, \varepsilon)$  denote the set of parameters under which the full model admits three steady states and let  $\Omega_S$  denote the analogous set under the societal feedback benchmark. Similarly, let  $\tilde{\Omega}$  and  $\tilde{\Omega}_S$  denote the respective sets such that a temporary increase in  $\mu$  can generate a persistent reduction in long run bias under the full model and societal-feedback benchmark.

The basin-switching logic of Proposition 1 extends to the full model. Under the conditions of Proposition 2, if  $\bar{\pi}^0 \in (\bar{\pi}_2^*(\mu^0), \bar{\pi}_2^*(\frac{1-q}{q}))$ , then there exists a  $\mu' > \mu^0$  such that temporarily increasing  $\mu^0$  to  $\mu'$  yields a persistent reduction in steady-state bias with convergence to  $\bar{\pi}_1^*(\mu^0)$ .<sup>13</sup>

However, because individuals increase effort when  $\bar{\pi}$  rises, the feedback from aggregate bias to contact outcomes is dampened, and the parameter region supporting multiplicity shrinks. The following proposition shows that this shrinkage is monotone in  $V$ . As the value of successful interaction rises, the set of environments admitting  $\mu$ -based tipping contracts strictly.

**Proposition 3.** *Suppose  $\mu > -1$ ,  $\varepsilon > 0$ , and  $\beta_0 \in [0, 1)$ . For each  $\bar{\pi} > 0$ , the pointwise slope envelope  $\tilde{p}'(\bar{\pi}; V)$  is strictly decreasing in  $V$ . Consequently:*

- (i)  $\tilde{P}'(V) = \sup_{\bar{\pi} \in [0, 1]} \tilde{p}'(\bar{\pi}; V)$  and  $\bar{\Delta}(V)$  are both strictly decreasing and continuous in  $V$ .
- (ii) For any  $V' > V > 0$ ,  $\Omega(V') \subset \Omega(V)$ . That is, the parameter region supporting multiple steady states shrinks strictly as  $V$  rises.

Proposition 3 delivers a monotone comparative static that strengthens the cross-model comparison with the societal-feedback benchmark. Setting  $V = 0$  recovers the benchmark (effort is absent and  $m_{\beta_0}(\bar{\pi}; 0) = (1 - \beta_0)\theta(\bar{\pi})$ ), so  $\Omega(V) \subset \Omega_S$  for any  $V > 0$  with strict inclusion when  $\beta_0 \in [0, 1)$ . The within-model result goes further. As  $V$  rises from any baseline, the multiplicity region contracts strictly, and the basin-switching region  $\tilde{\Omega}$  contracts with it.

Taken together, the full model reveals a central tradeoff introduced by effort. Higher-value contact reduces steady-state in-group bias by inducing preventive investment in effort, thereby stabilizing

---

<sup>13</sup>The proof, which follows the benchmark argument with  $\Psi$  replacing  $H$ , is in the Appendix.

the dynamics. At the same time, this stabilization weakens the sensitivity of contact outcomes to aggregate bias. Because individuals respond to rising  $\bar{\pi}$  by increasing effort, the societal feedback loop is dampened and the nonlinearity required for tipping dynamics is harder to sustain. High-value contact environments therefore lower steady-state in-group bias while simultaneously limiting the scope for temporary exposure to generate permanent regime shifts. In this sense, effort changes both the level and the shape of the dynamics. It lowers long-run bias, but it also shrinks the set of environments in which  $\mu$ -based policy succeeds.

If the economy is deeply entrenched in high bias, so that  $\bar{\pi}^0$  lies above the window  $(\bar{\pi}_2^*(\mu^0), \bar{\pi}_2^*(\frac{1-q}{q}))$ , it cannot be rescued by any temporary increase in  $\mu$ . The following proposition shows that, under the maintained assumptions, increasing  $V$  provides an alternative route to permanent change that does not require the initial condition to lie inside the  $\mu$ -window. Let  $\bar{\pi}_\ell^*(V)$  denote the steady states as functions of  $V$ .

**Proposition 4.** *Suppose  $\varepsilon > 0$ ,  $\mu > -1$ , and parameters are contained in  $\Omega$ .*

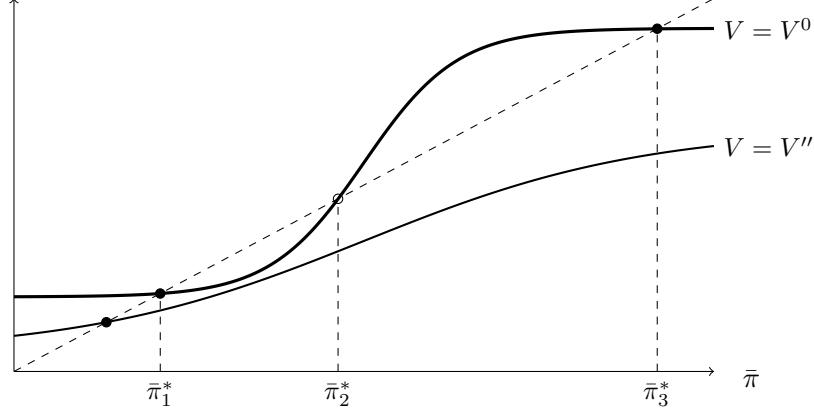
- (i) *There exists  $V' > V^0$  such that for all  $V \geq V'$ , the full model admits a unique globally attracting steady state  $\bar{\pi}_1^*(V) \in (0, 1)$ , with  $\bar{\pi}_1^*(V) < \bar{\pi}_2^*(V^0)$ .*
- (ii) *For any  $\bar{\pi}^0 \in (\bar{\pi}_2^*(V^0), 1)$ , temporarily increasing  $V$  from  $V^0$  to any  $V'' > V'$  satisfying (i) yields a persistent reduction in long-run bias, with the economy converging to  $\bar{\pi}_1^*(V^0)$ .*

Proposition 4 complements the  $\mu$ -intervention but works through a fundamentally different mechanism.<sup>14</sup> A temporary increase in  $\mu$  works by shifting the tipping point  $\bar{\pi}_2^*(\mu)$  upward, expanding the basin of attraction of the low-bias steady state until the current state falls inside it; this requires  $\bar{\pi}^0$  to lie within a specific window of initial conditions. A temporary increase in  $V$  works by raising the return to effort. Individuals then invest heavily in positive contact even in high-bias environments, eroding the self-reinforcing dynamic that sustains the high-bias trap and leaving a unique low-bias steady state. During the intervention, aggregate bias reduces towards the unique low-bias steady state. The restriction on initial conditions near the tipping point disappears entirely. Figure 3 illustrates this mechanism.

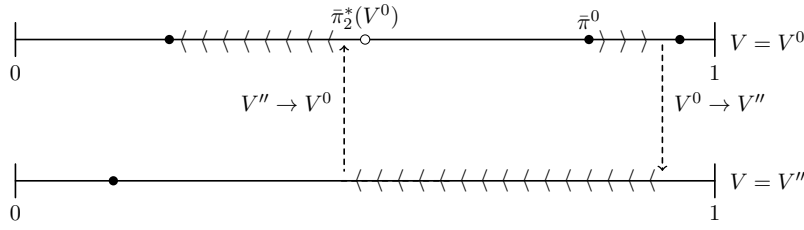
Proposition 3 shows that endogenous effort narrows the set of environments in which temporary contact can work through basin switching. Proposition 4 turns to the complementary margin:

---

<sup>14</sup>Proposition A3 in the Appendix establishes the existence of the  $\mu$ -intervention in the full model.



(a) Evolution of  $\bar{\pi}$  under  $V = V^0$  and  $V = V''$ .



(b) Transition path given initial biases  $(\pi_1^0, \pi_2^0)$  and initial average bias  $\bar{\pi}^0 = q\pi_1^0 + (1 - q)\pi_2^0$  and temporary increase in  $V$  from  $V^0$  to  $V''$ .

Figure 3: Graphical illustration of Proposition 4.

whether raising  $V$  can eliminate the high-bias regime altogether and generate permanent change even when the initial condition lies outside the  $\mu$ -window.

The two instruments therefore differ in robustness.  $\mu$ -based interventions are sufficient only when the economy is already near the tipping point, whereas  $V$ -based interventions admit no such constraint. Proposition 4 shows that a sufficiently strong  $V$ -intervention generates permanent change from any initial condition (provided that aggregate bias was trending towards the high-bias steady state), while Proposition A3 shows that a  $\mu$ -intervention succeeds only when prevailing bias lies within a window around the tipping point. This asymmetry implies that the two instruments are complements rather than substitutes. A temporary  $V$ -intervention can move a deeply entrenched economy into the window from which a subsequent  $\mu$ -intervention completes the transition to the low-bias steady state, allowing the  $V$ -intervention to be withdrawn earlier than under a pure  $V$ -policy. Formalizing this sequencing logic requires two preliminary results.

The first derives a bound on how fast aggregate bias can fall in a single period under the baseline

matching structure.

**Lemma 4.** *For any  $\mu > -1$ ,  $\bar{\pi}^t - \bar{\pi}^{t+1} \leq 2q(1-q)(1+\mu) \equiv C^*(\mu, q)$  for all  $t \geq 0$ .*

Second, the unstable tipping point moves monotonically in both policy instruments. Under  $\Omega$  with  $\mu > -1$  and  $\varepsilon > 0$ , the unstable tipping point  $\bar{\pi}_2^*(\mu, V)$  is strictly increasing in both  $\mu$  and  $V$ .<sup>15</sup>

Fix a candidate sequential matching level  $\tilde{\mu} \in (\mu^0, \frac{1-q}{q}]$  and a  $V$ -intervention level  $V''$  satisfying Proposition 4(i). Define crossing times along the trajectory induced by the constant policy  $(V'', \mu^0)$ :

$$\begin{aligned} T_V &\equiv \min \{t \geq 0 : \bar{\pi}^t < \bar{\pi}_2^*(\mu^0, V^0)\}, \\ \tilde{T}_V &\equiv \min \{t \geq 0 : \bar{\pi}^t < \bar{\pi}_2^*(\tilde{\mu}, V^0)\}, \end{aligned}$$

and the  $\mu$ -phase duration along the trajectory induced by  $(\tilde{\mu}, V^0)$  initiated at  $\bar{\pi}^{\tilde{T}_V}$ :

$$T_\mu \equiv \min \{t \geq 0 : \bar{\pi}^{\tilde{T}_V+t} < \bar{\pi}_2^*(\mu^0, V^0)\}.$$

To obtain a nontrivial sequencing result, assume that a  $\mu$ -intervention alone fails; i.e., for every feasible  $\mu' \in (\mu^0, \frac{1-q}{q}]$ ,  $\bar{\pi}^0 \geq \bar{\pi}_2^*(\mu', V^0)$ . Additionally, assume that there exists  $\tilde{\mu} \in (\mu^0, \frac{1-q}{q}]$  such that  $C^*(\mu^0, q) < \bar{\pi}_2^*(\tilde{\mu}, V^0) - \bar{\pi}_2^*(\mu^0, V^0)$ .

**Proposition 5.** *Suppose  $\varepsilon > 0$ ,  $\mu^0 > -1$ , and parameters are contained in  $\Omega$  for all  $\mu \in [\mu^0, \frac{1-q}{q}]$ , with stable aggregate bias levels  $\bar{\pi}_1^*(\mu, V) < \bar{\pi}_3^*(\mu, V)$  and unstable tipping point  $\bar{\pi}_2^*(\mu, V)$ . Fix  $(\mu^0, V^0)$  and  $\bar{\pi}^0 \in (\bar{\pi}_2^*(\mu^0, V^0), 1)$ . Then  $T_V, \tilde{T}_V, T_\mu < \infty$  and the combined policy that raises  $V^0$  to  $V''$  on  $t \in [0, \tilde{T}_V)$ ,  $\mu^0$  to  $\tilde{\mu}$  on  $t \in [\tilde{T}_V, \tilde{T}_V + T_\mu)$ , then ends the intervention yields a persistent reduction in steady state aggregate bias. Additionally,  $\tilde{T}_V < T_V$ , so the  $V$ -intervention is withdrawn strictly earlier under sequencing than under the  $V$ -only policy.*

Figure 4 illustrates the transition path of the sequenced policy. Proposition 5 shows that sequencing can shorten the required  $V$ -phase, though not necessarily total intervention duration. A  $\mu$ -intervention need only last long enough to push the trajectory below the baseline tipping point  $\bar{\pi}_2^*(\mu_0, V^0)$ , whereas a pure  $V$ -intervention must do so without the aid of a temporary increase in  $\mu$ . Whether sequencing reduces total duration depends on the speed of the  $\mu$ -induced dynamics relative to the remaining  $V$ -phase. A more aggressive  $\tilde{\mu}$  shortens the  $V$ -phase by raising the threshold the

<sup>15</sup>This follows from the implicit function theorem. See the Appendix for a proof.

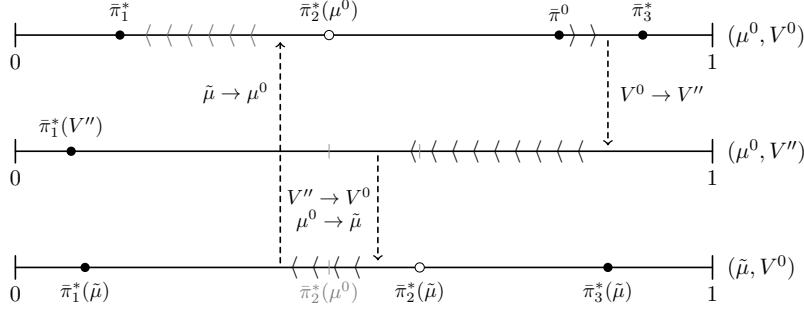


Figure 4: Transition path given initial biases  $(\pi_1^0, \pi_2^0)$  and initial average bias  $\bar{\pi}^0 = q\pi_1^0 + (1-q)\pi_2^0$  and temporary increase in  $V$  from  $V^0$  to  $V''$ , followed by returning  $V''$  to  $V^0$  and instituting a temporary increase in  $\mu$  from  $\mu^0$  to  $\tilde{\mu}$ .

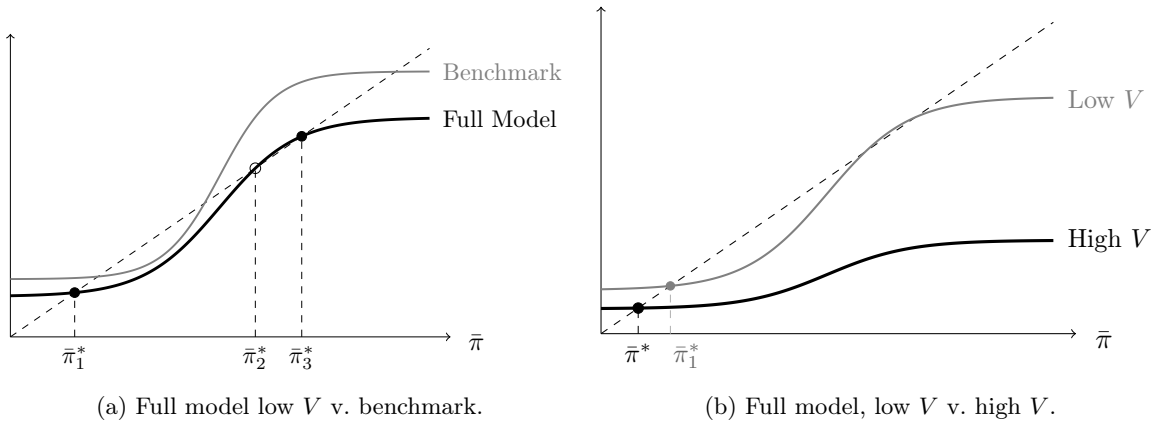


Figure 5: Endogenous effort attenuates the societal feedback sustaining tipping dynamics (Panel (a), Proposition 3). As  $V$  increases, this attenuation eliminates the high-bias steady state entirely (Panel (b), Proposition 4)

trajectory must cross before the  $\mu$ -phase begins, but may lengthen the  $\mu$ -phase if the state enters the tipping window close to its upper boundary. The preferred value of  $\tilde{\mu}$  therefore trades off the duration of the two phases and need not equal the maximum feasible matching rate.

Sections 3 and 4 together clarify the conditions under which contact alters the steady-state in-group bias. When feedback is linear or convex, exposure affects long-run bias but does not generate threshold-driven regime shifts. Figure 5 illustrates the two central mechanisms. Endogenous effort shrinks the parameter region in which  $\mu$ -based tipping is possible, but simultaneously creates the conditions under which  $V$ -based regime elimination becomes effective. The contact hypothesis therefore holds when social feedback is sigmoidal, and the instrument best suited to exploiting that feedback depends on where the population stands relative to the tipping point.

## 5 Welfare Effects of Contact- and Value-Interventions

The preceding section characterizes when each instrument can generate permanent regime change. This section asks which instrument does so at lowest welfare cost. The model's effort structure delivers a closed-form welfare function in which aggregate welfare equals the interaction prize minus a population-averaged effort tax that is increasing in aggregate bias. The low-bias steady state strictly welfare-dominates the high-bias steady state.  $V$ -interventions raise welfare on impact, while  $\mu$ -interventions impose a front-loaded welfare penalty before eventually generating gains. Accounting for these transition-path effects can reverse the fiscal ranking of instruments.

Throughout, I maintain two assumptions: (i) the economy begins in the high-bias basin under  $(V^0, \mu^0)$ , and (ii)  $\bar{\pi}^0 \in (\bar{\pi}_2^*(\mu^0), \bar{\pi}_2^*(\frac{1-q}{q}))$ , so that all three candidate policies,  $\mu$ -only,  $V$ -only, and sequencing, are dynamically feasible and can be compared on equal footing. When (ii) fails,  $\mu$ -only is ruled out by the dynamics of Section 4 and the comparison reduces to  $V$ -only versus sequencing.

### 5.1 The Welfare Function and Steady-State Ranking

Recall from (4) that  $i$ 's utility, when matched with  $j$  can be written as

$$u_i = \rho_{b_i, b_j} V - \frac{1}{2} e_i^2.$$

In equilibrium, the probability of success is  $\rho = 1 - (1 - s^*)(1 - \beta)\theta$ . Using  $e^* = \frac{z}{1+z}$ , where  $z = (1 - \beta)\theta V$  (by Lemma 1),  $\rho V = V - \frac{z}{(1+z)^2}$ , and  $\frac{(e^*)^2}{2} = \frac{z^2}{2(1+z)^2}$ . Therefore,

$$u^* = V - \frac{z(2+z)}{2(1+z)^2} = V - \frac{1}{2} s^*,$$

as  $s^* = f(z) = \frac{z(2+z)}{(1+z)^2}$  and thus  $1 - s^* = \frac{1}{(1+z)^2}$ . Every match yields the full prize  $V$  minus an “effort tax” of  $\frac{s^*}{2}$  that accounts for environmental adversity induced by the tension between biases and a positive interaction.

For any state  $\pi$ , let  $s_\beta^*$  denote the equilibrium aggregate effort given the bias pair  $\beta_k$ ,  $k \in \{0, 1, 2\}$  and let

$$\begin{aligned} \bar{S}(\pi) &= [q\eta_1 + (1-q)\eta_2] s_0^* + [q(1-\eta_1) + (1-q)(1-\eta_2)] \\ &\quad \times \left[ (1-\pi_1)(1-\pi_2) s_0^* + ((1-\pi_1)\pi_2 + \pi_1(1-\pi_2)) s_1^* + \pi_1\pi_2 s_2^* \right] \end{aligned}$$

denote the average aggregated effort of the population in the Nash equilibrium of the social interaction stage game. Then, the welfare function can be written as

$$W(\pi) = V - \frac{1}{2}\bar{S}(\pi). \quad (16)$$

This expression shows that welfare is determined entirely by  $V$  and population-average effort. In-group matches contribute  $s_0^*$  (effort against the no-bias environment), while out-group matches contribute a weighted average of  $s_0^*$ ,  $s_1^*$ , and  $s_2^*$ , where the weights depend on the bias composition.

Observe from above that  $f(z)$  is strictly increasing in  $z$  and that  $\frac{df(z)}{dz} = \frac{2}{(1+z)^3} > 0$ . As  $z_k = (1 - \beta_k)\theta(\bar{\pi})V$  is increasing in  $\bar{\pi}$  through  $\theta(\bar{\pi})$ , and the bias composition shifts toward higher- $z$  match types as  $\bar{\pi}$  rises,  $\bar{S}$  is strictly increasing in  $\bar{\pi}$  on any path where all three effects are aligned. It follows immediately that for any two interior steady states with  $\bar{\pi}_1^* < \bar{\pi}_3^*$ , the low-bias steady state strictly welfare-dominates the high-bias one:  $W(\bar{\pi}_1^*) > W(\bar{\pi}_3^*)$ . The ranking reflects two effects: a direct feedback effect and a composition effect. High-bias environments require more effort to sustain positive interactions ( $\theta(\bar{\pi}_3^*) > \theta(\bar{\pi}_1^*)$ ) and, in the high-bias steady state, more individuals are biased, so a larger share of out-group matches are biased pairs (with  $\beta = 0$ ) and a smaller share are unbiased pairs (with  $\beta = \beta_0$ ). The distribution of matches shifts toward higher-effort types, lowering welfare. Define the welfare gap as

$$\Delta W = \frac{1}{2} [\bar{S}(\bar{\pi}_3^*) - \bar{S}(\bar{\pi}_1^*)]. \quad (17)$$

Now, suppose that all individuals discount future payoffs by the common factor  $\delta \in (0, 1)$ . The present discounted value of steady-state welfare is  $\frac{W}{1-\delta}$ , so the long-run value of transitioning permanently from  $\bar{\pi}_3^*$  to  $\bar{\pi}_1^*$  (the *value of regime change R*) is

$$R = \frac{\Delta W}{1-\delta} = \frac{\bar{S}(\bar{\pi}_3^*) - \bar{S}(\bar{\pi}_1^*)}{2(1-\delta)}. \quad (18)$$

$R$  captures the present value of the perpetual per-period reduction in effort costs from operating in a low-bias environment.

## 5.2 Transition-Path Welfare

To incorporate the transition dynamics, I define the  $V$ -bonus  $B_V(t)$ , which captures the per-period welfare gains under a  $V$ -intervention. Let

$$B_V(t) = W(\bar{\pi}^t; V'', \mu^0) - W(\bar{\pi}^t; V^0, \mu^0) = (V'' - V^0) - \frac{1}{2} [\bar{S}(\bar{\pi}^t; V'', \mu^0) - \bar{S}(\bar{\pi}^t; V^0, \mu^0)]. \quad (19)$$

Observe that differentiating  $u^*$  with respect to  $V$  yields

$$\frac{\partial u^*}{\partial V} = 1 - \frac{(1-\beta)\theta}{(1+z)^3}$$

As  $z \in (0, 1)$ ,  $(1-\beta)\theta \leq 1$  and  $\frac{(1-z)}{(1+z)^3} < 1$ , so  $\frac{\partial u^*}{\partial V} > 0$ . Hence,  $u^*$  is weakly increasing in  $V$  for every match type. Aggregating implies that  $B_V(t) \geq 0$ . Increasing  $V$  has two effects: it increases the prize from each successful interaction and it induces additional effort. The prize effect always dominates because the effort response is a second-order correction. Individuals optimally choose effort that equates the marginal cost to the marginal benefit, so the net payoff from additional effort is always non-negative by the envelope theorem.  $V$ -interventions are therefore welfare-improving in every period, even before bias adjusts.

Define  $J_\mu(t)$  as the  $\mu$ -penalty; i.e., the per-period welfare loss from operating under  $\tilde{\mu}$  rather than  $\mu^0$  for a given  $\pi^t$ :

$$J_\mu(t) = W(\bar{\pi}^t; V^0, \mu^0) - W(\bar{\pi}^t; V^0, \tilde{\mu}) = \frac{1}{2} [\bar{S}(\bar{\pi}^t; V^0, \tilde{\mu}) - \bar{S}(\bar{\pi}^t; V^0, \mu^0)],$$

which can be rewritten as

$$J_\mu(t) = \frac{1}{2}(\tilde{\mu} - \mu^0)q(1-q) \sum_{\tau} [\bar{s}_{out,\tau}^*(\pi^t) - s_0^*] \geq 0, \quad (20)$$

where  $\bar{s}_{out,\tau}^*(\pi^t)$  is the average aggregate effort from a type- $\tau$  individual in an out-group match and  $s_0^*$  is the aggregate effort from an in-group match. The penalty is proportional to the mandate intensity  $\tilde{\mu} - \mu^0$  and the effort-cost gap between out-group and in-group matches. The gap is largest when bias is high (many biased pairs with  $\beta_2 = 0$ ) and shrinks as bias falls (more unbiased pairs with  $\beta_0 > 0$ , which require less effort). It's nonnegativity follows from the fact that effort is always at its lowest with either in-group matches or unbiased matches. Hence  $\bar{s}_{out,\tau}^*(\pi^t) - s_0^* \geq 0$ , with equality only if  $\bar{\pi} = 0$ , so  $\pi^t = (1, 1)$ .  $J_\mu(t)$  is monotonically declining along any trajectory where bias is falling. As  $\bar{\pi}$  falls, the fraction of high-effort biased out-group pairs shrinks, so  $s_{out}$  converges toward  $s_0^*$ , and the gap narrows.

Unlike  $V$ -interventions, which uniformly increase welfare (as  $B_V(t)$  is nonnegative),  $\mu$ -interventions exhibit a  $J$ -shape. Let  $\bar{\pi}_{int}^t$  denote the aggregate bias at time  $t$  in the scenario in which there is a  $\mu$ -intervention and let  $\bar{\pi}_{cf}$  denote the counterfactual in which there is no intervention, with  $\pi_{int}^t$  and  $\pi_{cf}^t$  defined analogously. The net change in per-period welfare from the intervention is

$$W(\bar{\pi}_{int}^t; \tilde{\mu}) - W(\bar{\pi}_{cf}^t; \mu^0) = \frac{1}{2} [\bar{S}(\pi_{cf}^t; \mu^0) - \bar{S}(\pi_{int}^t; \tilde{\mu})].$$

Let  $\bar{S}(\pi_{int}^t; \mu^0)$  denote the average aggregate effort if bias had followed the intervention trajectory but matching had stayed at baseline. Adding and subtracting this value from the change in welfare yields

$$\begin{aligned} W(\bar{\pi}_{int}^t; \tilde{\mu}) - W(\bar{\pi}_{cf}^t; \mu^0) &= \frac{1}{2} [\bar{S}(\pi_{cf}^t; \mu^0) - \bar{S}(\pi_{int}^t; \mu^0)] + \frac{1}{2} [\bar{S}(\pi_{int}^t; \mu^0) - \bar{S}(\pi_{int}^t; \tilde{\mu})] \\ &= \frac{1}{2} \underbrace{[\bar{S}(\pi_{cf}^t; \mu^0) - \bar{S}(\pi_{int}^t; \mu^0)]}_{\text{bias-reduction gain } \geq 0} + \underbrace{-J_\mu(t)}_{\mu\text{-penalty } \leq 0} . \end{aligned}$$

At the start of the intervention  $t_1$ ,  $\pi_{cf}^{t_s} = \pi_{int}^{t_1}$ , so  $W(\bar{\pi}_{int}^{t_1}; \tilde{\mu}) - W(\bar{\pi}_{cf}^{t_1}; \mu^0) < 0$  and the initial change in welfare is strictly negative. At some  $\hat{t} > t_1$ ,  $W(\bar{\pi}_{int}^{\hat{t}}; \tilde{\mu}) = W(\bar{\pi}_{cf}^{\hat{t}}; \mu^0)$  and remains positive for all  $t$  thereafter. As  $t \rightarrow \infty$ , the intervention converges to the low-bias steady state. The bias-reduction gain stabilizes at  $\Delta W$  (the welfare gap). The penalty stabilizes at  $J_\mu(\infty)$  evaluated at  $\bar{\pi}_1$ , which is small (bias is low and the effort gap is narrow). Total welfare gains are permanently positive.

**Proposition 6.** *Total welfare under a  $V$ -intervention relative to the counterfactual is strictly positive in every period. Total welfare under a  $\mu$ -intervention relative to the counterfactual is initially negative, eventually positive during the intervention, and permanently positive thereafter.*

The  $\mu$ -penalty  $J_\mu(t)$  provides an economic foundation for political resistance to integration mandates.  $\mu$ -interventions impose visible, immediate welfare costs on participants. They are being asked to engage in costlier out-group interactions while the bias-reduction gains materialize only later.  $V$ -interventions face no such asymmetry:  $B_V(t) \geq 0$  from the start. This welfare-based rationale for  $V$ -first sequencing is distinct from the dynamic rationale in Proposition 5.  $V$ -first is preferred not only because it expands the basin of attraction, but also because it avoids front-loading welfare costs on participants.

### 5.3 Welfare-Efficient Intervention Design

Proposition 6 establishes that a transition from the high-bias to the low-bias steady state increases welfare, that  $V$ -interventions generate per-period welfare gains throughout the transition, and that  $\mu$ -interventions impose front-loaded welfare costs that are eventually offset by bias-reduction gains. This subsection uses these properties to determine which instrument, or combination of instruments, achieves a successful reduction of in-group bias at the lowest welfare-adjusted cost.

Because all successful policies converge to the same low-bias steady state  $\bar{\pi}_1^*(\mu^0, V^0)$ , the long-run value of regime change  $R$  is identical across them. The welfare ranking therefore reduces to

a comparison of costs incurred during the intervention phase, where costs include both the fiscal expenditure on the policy and the transition-path welfare effects characterized above. To formalize this, define  $c_V > 0$  and  $c_\mu > 0$  denote the per-period (monetary) costs of the  $V$ - and  $\mu$ -interventions, respectively. Then, define the *welfare-adjusted per-period costs*

$$\tilde{c}_V(t) \equiv c_V - B_V(t), \quad \tilde{c}_\mu(t) \equiv c_\mu + J_\mu(t). \quad (21)$$

The effective cost of a  $V$ -intervention in period  $t$  is its fiscal cost  $c_V$  minus the welfare bonus  $B_V(t) \geq 0$  it generates. The effective cost of a  $\mu$ -intervention in period  $t$  is its fiscal cost  $c_\mu$  plus the welfare penalty  $J_\mu(t) \geq 0$  it imposes. Accounting for transition-path welfare therefore makes  $V$  cheaper and  $\mu$  more expensive than their fiscal costs alone suggest.

I compare three candidate policies, each of which successfully crosses the tipping point.

1. *Policy M ( $\mu$ -only)*. Raise  $\mu$  from  $\mu^0$  to  $\mu'$  for  $T_M$  periods, where

$$T_M \equiv \min \{t \geq 0 : \bar{\pi}^t < \bar{\pi}_2^*(\mu^0, V^0)\}$$

along the trajectory induced by  $(\mu', V^0)$ . This policy can only be successfully implemented when  $\bar{\pi}^0 \in (\bar{\pi}_2^*(\mu^0, V^0), \bar{\pi}_2^*(\frac{1-q}{q}, V^0))$ .

2. *Policy V ( $V$ -only)*. Raise  $V$  from  $V^0$  to  $V''$  for  $T_V$  periods. Feasible from any  $\bar{\pi}^0 \in (\bar{\pi}_2^*(\mu^0, V^0), 1)$ .

3. *Policy S (sequencing)*. Raise  $V$  from  $V^0$  to  $V''$  for  $\tilde{T}_V$  periods, then raise  $\mu$  from  $\mu^0$  to  $\tilde{\mu}$  for  $T_\mu$  periods. Feasible from any  $\bar{\pi}^0 \in (\bar{\pi}_2^*(\mu^0, V^0), 1)$ .

For each policy  $P \in \{M, V, S\}$ , the total welfare-adjusted cost is the discounted sum of welfare-adjusted per-period costs along its trajectory:

$$C_M = \sum_{t=0}^{T_M-1} \delta^t \tilde{c}_\mu^M(t), \quad (22)$$

$$C_V = \sum_{t=0}^{T_V-1} \delta^t \tilde{c}_V^V(t), \quad (23)$$

$$C_S = \sum_{t=0}^{\tilde{T}_V-1} \delta^t \tilde{c}_V^S(t) + \sum_{t=\tilde{T}_V}^{\tilde{T}_V+T_\mu-1} \delta^t \tilde{c}_\mu^S(t), \quad (24)$$

where the superscript on  $\tilde{c}$  indicates that  $B_V$  and  $J_\mu$  are evaluated along the respective policy's trajectory.

**Proposition 7.** *For  $\delta$  sufficiently close to 1, the welfare-optimal policy minimizes the total welfare-adjusted cost among  $\{C_M, C_V, C_S\}$ .*

- (i) *V-only versus sequencing.  $V$  is welfare-preferred to  $S$  iff  $C_V < C_S$ .*
- (ii) *V-only versus  $\mu$ -only.  $V$  is welfare-preferred to  $M$  iff  $C_V < C_M$ .*
- (iii) *Sequencing versus  $\mu$ -only.  $S$  is welfare-preferred to  $M$  iff  $C_S < C_M$ .*

The proposition establishes that the welfare-optimal instrument depends on the relative magnitudes of  $B_V$ ,  $J_\mu$ , and the fiscal costs, and can differ from the policy that minimizes fiscal cost. Two forces drive this divergence. First,  $B_V \geq 0$  reduces the effective cost of the  $V$ -instrument below its fiscal cost, making  $V$ -only relatively more attractive than a pure fiscal comparison suggests. Second,  $J_\mu \geq 0$  raises the effective cost of the  $\mu$ -instrument above its fiscal cost, making  $\mu$ -only and the  $\mu$ -phase of sequencing relatively less attractive. These forces work in the same direction: both favor  $V$  over  $\mu$ , so the welfare ranking is unambiguously tilted toward  $V$ -intensive policies relative to the fiscal ranking. When comparing  $V$ -only and  $\mu$ -only interventions (statement (ii)), it merits mentioning that because  $B_V(t) \geq 0$  reduces the left side and  $J_\mu(t) \geq 0$  increases the right side, this condition can be satisfied even when  $c_V > c_\mu$ .

An immediate implication of Proposition 7 is if  $B_V(t) \geq c_V$  for all  $t \in [0, T_V)$ , then  $\tilde{c}_V^V(t) \leq 0$  for all  $t$  and Policy  $V$  is strictly welfare-preferred to both  $M$  and  $S$ . When the per-period welfare bonus from elevated  $V$  exceeds the fiscal cost of providing it, every period of the  $V$ -phase generates net welfare gains. Cutting the  $V$ -phase short to introduce a  $\mu$ -phase, which carries both a fiscal cost and a welfare penalty, is strictly worse. This condition is satisfied when  $V$ -interventions are inexpensive relative to the cooperative surplus they create, as in team-based pay structures where the additional compensation is modest relative to the gains from successful cooperation.

Furthermore, policy  $M$  is welfare-improving relative to no intervention only if

$$\frac{R}{\sum_{t=0}^{T_M-1} \delta^t \tilde{c}_\mu^M(t)} > 1,$$

i.e., the discounted value of regime change exceeds the total welfare-adjusted cost of the  $\mu$ -intervention. No such condition is required for Policy  $V$  because  $B_V(t) \geq 0$ ,  $V$ -only is welfare-improving for sufficiently small  $c_V$  at any  $\delta > 0$ . Because  $\mu$ -interventions impose front-loaded welfare costs while their gains are back-loaded, they require the planner to be sufficiently patient. The threshold is increasing in  $c_\mu$  and in the severity of the  $\mu$ -penalty.  $V$ -interventions face no such asymmetry:  $B_V(t) \geq 0$

ensures that welfare is weakly higher in every period of the intervention, so the fiscal cost is the only barrier to welfare improvement.

Lastly, there exist parameter configurations under which  $c_S < c_V$  (sequencing has lower fiscal cost than  $V$ -only) but  $C_V < C_S$ . Hence, minimizing fiscal cost can be a misleading criterion. Sequencing has lower fiscal cost than  $V$ -only interventions while  $V$ -only interventions have lower welfare-adjusted cost than sequencing (and vice versa). When the  $V$ -phase periods displaced by sequencing generate welfare bonuses that exceed  $c_V$ , the fiscal savings from shortening the  $V$ -phase come at a welfare cost. Those periods were net positive for the population. This divergence is sharpest in high-bias environments, where  $B_V$  is large (high  $\theta$  induces substantial effort, so the effort-cost reduction from higher  $V$  is valuable) and  $J_\mu$  is simultaneously large (the effort gap between out-group and in-group matches is wide).

#### 5.4 Distributional Effects

Does the majority or minority benefit more from a successful intervention? At any interior state  $\pi$  with  $\mu > -1$ , a type-2 (minority) individual's expected per-period welfare is strictly lower than a type-1 (majority) individual's. This is because the average effort of a minority individual is greater than that of a majority individual (Corollary 1). Consequently, the minority gains strictly more from any reduction in aggregate bias, and strictly more from a transition from the high-bias to the low-bias steady state.

The minority interacts out-group more frequently ( $\eta_2 < \eta_1$ ), so a larger fraction of their matches are costly out-group encounters. When aggregate bias falls, the effort-cost reduction is concentrated on out-group matches, and the minority has more of them. This distributional prediction complements Corollary 2. The majority is more biased in steady state, but the minority bears more of the welfare cost. Successful interventions therefore reduces both the level asymmetry (bias gap narrows) and the welfare asymmetry (effort-burden gap narrows).

## 6 Policy Implications

The model's three main results each carry a distinct policy implication, developed in turn below.

### 6.1 Effort and the Value of Successful Interaction

Policies that increase  $V$  correspond to institutional interventions that increase the returns to successful interaction. This includes team-based pay that ties compensation to cooperative output,

school environments that create shared goals across group lines, or professional settings structured around joint production. A sufficiently large temporary increase in  $V$  eliminates the high-bias steady state for the duration of the intervention, driving the economy toward a unique low-bias steady state from any initial condition (Proposition 4). The welfare analysis of Section 5 reinforces this conclusion.  $V$ -interventions raise per-period welfare on impact (Proposition 6), and when the welfare bonus exceeds the fiscal cost,  $V$ -only is strictly welfare-preferred to any policy involving a  $\mu$ -phase.

This prediction finds support in studies of integration into high-value settings (Rao, 2019; Boisjoly et al., 2006; Beaman et al., 2009; Hjort, 2014). The model’s mechanism implies a sharper reading of these findings than a directional prediction alone. Proposition 3 implies that the parameter region supporting  $\mu$ -based tipping shrinks monotonically in  $V$ . Environments with higher interaction stakes should therefore exhibit lower steady-state bias but a narrower region in which temporary contact interventions generate lasting change. Consistent with this, Rao (2019) finds that the largest reductions in discriminatory behavior arise specifically in economic exchange tasks, which are precisely the interactions where  $V$  is highest and, by the model’s logic, where effort is most strongly induced. A further practical constraint is that if institutional integration is perceived as culturally threatening, individuals may strategically withdraw, endogenously recreating assortative sorting and offsetting the gains from increasing  $V$  (Carvalho et al., 2017; Carvalho and Sacks, 2024; Carvalho et al., 2024b).

## 6.2 Assortative and Disassortative Matching

Policies that raise  $\mu$  toward disassortative matching correspond to integration mandates, diversity requirements, school assignment policies, and workplace composition rules that increase the frequency of out-group interaction. Conditional on the economy beginning in the basin of attraction of the high-bias steady state, a temporary increase in  $\mu$  generates a persistent reduction in bias if and only if prevailing average bias lies between the tipping point under the current contact rate and the tipping point under the maximum feasible contact rate.<sup>16</sup>

This window-dependence has a direct empirical counterpart. The tipping literature in urban economics documents that neighborhoods and schools exhibit sharp threshold dynamics in racial composition. Once the share of one group crosses a critical level, the social environment transitions

---

<sup>16</sup>Formally, the condition is  $\bar{\pi}^0 \in (\bar{\pi}_2^*(\mu^0), \bar{\pi}_2^*(\frac{1-q}{q}))$ , where  $\frac{1-q}{q}$  is the upper bound on  $\mu$  imposed by market clearing,  $q(1 - \eta_1) = (1 - q)(1 - \eta_2)$ .

rapidly and persistently to a segregated state (Schelling, 1971; Card et al., 2008). In the model, these are the kinds of environments where in-group bias feeds back strongly enough on interaction outcomes to support multiple steady states, and where  $\mu$ -based tipping is feasible.<sup>17</sup>

The window condition offers a specific, falsifiable account of cross-setting heterogeneity in contact interventions. Lowe (2021) finds persistent reductions in discrimination following a cooperative cricket program in India, while Mousa (2020) finds that a structurally similar intervention in post-ISIS Iraq produces effects confined to the contact setting. Consistent with the same logic, Ghosh et al. (2026) find that a structured youth camp intervention in West Bengal generates persistent reductions in in-group bias one year after the camps ended.

The model identifies three distinct channels through which such divergence can arise. First, the two populations may differ in their position relative to the tipping point, with prevailing bias in one setting lying inside the success window and in the other lying outside it. Second, the settings may differ in whether societal feedback is sufficiently sigmoidal to generate tipping dynamics at all, so that the same initial bias level produces a unique steady state in one environment and multiple steady states in the other. Third, even when both settings exhibit tipping dynamics, differences in the shape of  $\theta$  may shift the location and width of the tipping window itself, so that identical initial bias levels fall inside the window in one context and outside it in the other. All three channels generate predictions absent from purely heterogeneity-based explanations: whether a contact program succeeds depends on measurable features of the societal environment, not merely on implementation details. Mousa et al. (2025) provide complementary evidence from a randomized field experiment in Lebanon, where Syrian refugees comprise a quarter of the population and baseline prejudice is severe. Structured contact had null effects on prejudicial attitudes and actually reduced participants' willingness to engage in future out-group activities, consistent with the model's prediction that when prevailing bias lies well above the tipping threshold, increases in  $\mu$  fail to cross it and leave steady-state bias unchanged.

Steinmayr (2021) finds that communities exposed to refugee resettlement exhibit reduced support for far-right parties when meaningful direct contact occurs, but increased support when exposure occurs without genuine interaction. The model offers a specific account of this asymmetry. In-

---

<sup>17</sup>Interventions that successfully pushed neighborhoods or schools across such thresholds, as in sustained court-ordered desegregation, generated lasting integration (Guryan, 2004); interventions that failed to sustain the mandate were reversed once legal pressure lapsed, with corresponding deterioration in outcomes (Lutz, 2011; Billings et al., 2014).

creasing  $\mu$  raises the rate of out-group contact but does not directly improve interaction quality; without a corresponding increase in  $V$ , exposure alone may fail to cross the tipping threshold. Lebow et al. (2024) find that anti-migrant sentiment worsened nationally during the Venezuelan exodus, but not in the within-country regions receiving the most migrants, consistent with the idea that local exposure need not generate local backlash. The model does not account for backlash of the kind documented by Enos (2014), where out-group contact actively increases exclusionary attitudes; that pattern likely reflects mechanisms outside the model’s scope, such as identity threat or status competition. Within interventions that do succeed, however, the model predicts that effects should be asymmetric across majority and minority participants. Ghosh et al. (2026) find, in a randomized youth camp experiment in West Bengal, India, that higher out-group contact intensity backfires among Hindu majority participants but not Muslim minority participants, consistent with the model’s prediction that majority members, facing higher baseline in-group exposure, are more adversely affected by contact interventions that fall short of the tipping threshold.

The model does, however, offer a limited account of a weaker form of post-intervention rebound. A temporary increase in  $\mu$  or  $V$  shifts the low-bias steady state downward for the duration of the intervention; if the intervention runs long enough, the economy begins converging toward that lower steady state. When the intervention ends and parameters return to baseline, the low-bias steady state snaps back to its original level, and the economy must correct upward before settling there. The resulting rebound is bounded within the low-bias regime and is proportional to the duration of the intervention. Longer interventions push the economy further below the terminal steady state and therefore generate a larger, though still temporary, withdrawal effect. This mechanism is distinct from Enos-type backlash, in which out-group exposure actively pushes bias above its pre-intervention level; here, reversion never crosses into the high-bias basin.

These transition dynamics have a further implication for empirical evaluation. When an intervention ends shortly after the economy crosses the tipping point, it has entered the basin of attraction of the low-bias steady state but has not yet converged to it. Bias will continue falling for some time after withdrawal, so evaluations with short follow-up windows will understate the long-run treatment effect. Taken together, the two biases cut in opposite directions depending on intervention duration: short-lived interventions near the tipping point generate effects that grow after the study ends, while long-running interventions generate a transient post-withdrawal spike that can be misread as deterioration.

The welfare analysis provides a further constraint on  $\mu$ -based interventions. Even when prevailing bias lies within the success window, the welfare J-curve (Proposition 6) implies that  $\mu$ -interventions impose front-loaded welfare costs on participants before bias-reduction gains materialize. When these costs are severe, as in deeply biased environments where the effort gap between out-group and in-group matches is wide, a planner may prefer  $V$ -first sequencing even when  $\mu$ -only is dynamically feasible (Proposition 7).

The model also implies that contact policy changes the cross-group distribution of bias, not just its average level. Corollary 2 establishes that, under empirically plausible conditions, the majority is strictly more biased than the minority in any interior steady state, reflecting the majority's higher in-group interaction rate and greater accumulated in-group reinforcement.<sup>18</sup> A disassortative matching policy reduces in-group exposure for both groups, but the reduction is proportionally larger for the majority. A uniform increase in  $\mu$  therefore narrows the bias gap between groups, generating convergence in the distribution of bias alongside a reduction in average bias. This is consistent with evidence that contact effects are often stronger among majority than minority group members (Tropp and Pettigrew, 2005; Mousa et al., 2025).

### 6.3 Sequencing

Proposition 5 establishes that a  $V$ -first,  $\mu$ -second policy can shorten the  $V$ -phase relative to  $V$ -only. The welfare analysis adds a second dimension to this comparison. Proposition 7 shows that the welfare-optimal choice among  $V$ -only,  $\mu$ -only, and sequencing depends on welfare-adjusted costs that incorporate both fiscal expenditure and transition-path welfare effects. The welfare-adjusted ranking can differ from the fiscal ranking. Sequencing may reduce fiscal cost while increasing welfare-adjusted cost, because the  $V$ -phase periods eliminated by sequencing were generating net welfare gains that exceeded their fiscal cost. This divergence is sharpest in high-bias environments, where  $B_V$  is large and  $J_\mu$  is simultaneously large. The welfare analysis therefore implies that sequencing is most attractive when  $V$ -interventions are expensive ( $c_V$  high), the welfare bonus from elevated  $V$  is modest ( $B_V$  small), and the J-curve penalty is mild ( $J_\mu$  small); a configuration more likely in moderately biased environments near the tipping point than in deeply entrenched ones.

---

<sup>18</sup>The conditions are  $\mu > -1$ , so that out-group contact occurs with positive probability, and  $\varepsilon > 0$ , so that in-group interactions generate positive reinforcement ( $\varepsilon$  can be arbitrarily small).

## 7 Concluding Remarks

This paper develops a dynamic model of in-group bias to explain why some temporary contact interventions produce lasting reductions in bias while others do not. The central insight is that persistence requires self-reinforcing societal feedback strong enough to generate a tipping point. When that condition holds, a temporary increase in out-group contact has lasting effects only if it moves the population across the unstable threshold separating a high-bias regime from a low-bias regime; otherwise, bias declines during the intervention and reverts once contact returns to baseline.

The model also shows that policies that increase contact are not equivalent to policies that raise the value of successful interaction. Because individuals choose effort in out-group interactions, higher interaction stakes induce greater effort and reduce steady-state bias, but also dampen the feedback that sustains multiplicity. The same force that improves interaction quality therefore narrows the set of environments in which temporary contact alone can generate permanent change. This tension yields the paper's main empirical prediction: environments with higher stakes for successful interaction should exhibit lower steady-state bias, but a smaller region in which temporary increases in contact have lasting effects.

These results imply that when bias is too entrenched for contact alone to succeed, policies that raise the value of successful interaction can create the conditions under which temporary contact becomes permanently effective. The framework therefore provides a threshold-based way to think about the heterogeneous effects of contact interventions, the distinct roles of exposure and incentives, and the conditions under which sequencing the two can be effective. It also implies a cross-group asymmetry in steady-state bias and shows that welfare comparisons across instruments need not coincide with simple fiscal rankings. More broadly, the paper offers a tractable framework for interpreting heterogeneity in contact-based interventions and for designing policies that can produce durable reductions in in-group bias.

## References

- Akerlof, George A. and Rachel E. Kranton**, "Economics and identity," *Quarterly Journal of Economics*, 2000, 115 (3), 715–753.
- Allcott, Hunt and Matthew Gentzkow**, "Social media and fake news in the 2016 election," *Journal of Economic Perspectives*, 2017, 31 (2), 211–236.
- Allport, Gordon W.**, *The Nature of Prejudice*, Cambridge, MA: Addison-Wesley, 1954. Abridged edition.
- Arrow, Kenneth J.**, *The Theory of Discrimination*, Princeton University Press, 1973.

- Bail, Christopher A., Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky**, “Exposure to opposing views on social media can increase political polarization,” *Proceedings of the National Academy of Sciences*, 2018, *115* (37), 9216–9221.
- Bazzi, Samuel, Arya Gaduh, Alexander D. Rothenberg, and Maisy Wong**, “Unity in diversity? How intergroup contact can foster nation building,” *American Economic Review*, 2019, *109* (11), 3978–4025.
- Beaman, Lori, Raghendra Chattopadhyay, Esther Duflo, Rohini Pande, and Petia Topalova**, “Powerful women: Does exposure reduce bias?,” *Quarterly Journal of Economics*, 2009, *124* (4), 1497–1540.
- Becker, Gary S.**, *The Economics of Discrimination*, University of Chicago Press, 1957.
- Bénabou, Roland and Jean Tirole**, “Identity, morals, and taboos: Beliefs as assets,” *Quarterly Journal of Economics*, 2011, *126* (2), 805–855.
- Bernhard, Helen, Ernst Fehr, and Urs Fischbacher**, “Group affiliation and altruistic norm enforcement,” *American Economic Review: Papers and Proceedings*, 2006, *96* (2), 217–221.
- Billings, Stephen B., David J. Deming, and Jonah Rockoff**, “School segregation, educational attainment, and crime: Evidence from the end of busing in Charlotte-Mecklenburg,” *Quarterly Journal of Economics*, 2014, *129* (1), 435–476.
- Bisin, Alberto and Thierry Verdier**, “Beyond the melting pot: Cultural transmission, marriage, and the evolution of ethnic and religious traits,” *Quarterly Journal of Economics*, 2000, *115* (3), 955–988.
- and —, “The economics of cultural transmission and the dynamics of preferences,” *Journal of Economic Theory*, 2001, *97* (2), 298–319.
- Bohren, J. Aislinn, Peter Hull, and Alex Imas**, “Systemic discrimination: Theory and measurement,” *Quarterly Journal of Economics*, 2025, *140* (3), 1743–1799.
- Boisjoly, Johanne, Greg J. Duncan, Michael Kremer, Dan M. Levy, and Jacque Eccles**, “Empathy or antipathy? The impact of diversity,” *American Economic Review*, 2006, *96* (5), 1890–1905.
- Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer**, “Stereotypes,” *Quarterly Journal of Economics*, 2016, *131* (4), 1753–1794.
- Bowles, Samuel and Herbert Gintis**, “Social capital and community governance,” *Economic Journal*, 2002, *112* (483), F419–F436.
- Boxell, Levi, Matthew Gentzkow, and Jesse M. Shapiro**, “Greater internet use is not associated with faster growth in political polarization among US demographic groups,” *Proceedings of the National Academy of Sciences*, 2017, *114* (40), 10612–10617.
- Card, David, Alexandre Mas, and Jesse Rothstein**, “Tipping and the dynamics of segregation,” *Quarterly Journal of Economics*, 2008, *123* (1), 177–218.
- Carvalho, Jean-Paul and Michael Sacks**, “Radicalisation,” *The Economic Journal*, 2024, *134* (659), 1019–1068.
- , **Jared Rubin, and Michael Sacks**, “Failed secular revolutions: Religious belief, competition, and extremism,” *Public Choice*, 2024, *200*, 561–586.
- , **Mark Koyama, and Cole Williams**, “Resisting education,” *Journal of the European Economic Association*, 2024, *22* (6), 2549–2597.
- , —, and **Michael Sacks**, “Education, identity, and community: Lessons from Jewish emancipation,” *Public Choice*, 2017, *171*, 119–143.

- Chen, Yan and Sherry Xin Li**, “Group identity and social preferences,” *American Economic Review*, 2009, *99* (1), 431–57.
- Cinyabuguma, Mathieu, Talbot Page, and Louis Putterman**, “Cooperation under the threat of expulsion in a public goods experiment,” *Journal of Public Economics*, 2005, *89* (8), 1421–1435.
- Clochard, Gwen-Jirō, Guillaume Hollard, and Omar Sene**, “Bringing contact interventions to the lab: Effects of brief bilateral discussions on interethnic trust in Senegal,” *World Development*, 2026, *199*, Article No. 107247.
- Currarini, Sergio, Matthew O. Jackson, and Paolo Pin**, “An economic model of friendship: Homophily, minorities, and segregation,” *Econometrica*, 2009, *77* (4), 1003–1045.
- Enos, Ryan D.**, “Causal effect of intergroup contact on exclusionary attitudes,” *Proceedings of the National Academy of Sciences*, 2014, *111* (10), 3699–3704.
- Fehr, Ernst and Simon Gächter**, “Cooperation and punishment in public goods experiments,” *American Economic Review*, 2000, *90* (4), 980–994.
- Gentzkow, Matthew and Jesse M. Shapiro**, “Ideological segregation online and offline,” *Quarterly Journal of Economics*, 2011, *126* (4), 1799–1839.
- Ghosh, Arkadev, Prerna Kundu, Matt Lowe, and Gareth Nellis**, “Creating cohesive communities: A youth camp experiment in India,” *Review of Economic Studies*, 2026, *93* (1), 438–475.
- Goette, Lorenz, David Huffman, and Stephan Meier**, “The impact of group membership on cooperation and norm enforcement: Evidence using random assignment to real social groups,” *American Economic Review: Papers and Proceedings*, 2006, *96* (2), 212–216.
- Gürerk, Özgür, Bernd Irlenbusch, and Bettina Rockenbach**, “The competitive advantage of sanctioning institutions,” *Science*, 2006, *312* (5770), 108–111.
- Guryan, Jonathan**, “Desegregation and black dropout rates,” *American Economic Review*, 2004, *94* (4), 919–943.
- Hjort, Jonas**, “Ethnic divisions and production in firms,” *Quarterly Journal of Economics*, 2014, *129* (4), 1899–1946.
- Holmström, Bengt**, “Moral hazard in teams,” *Bell Journal of Economics*, 1982, *13* (2), 324–340.
- Kandori, Michihiro, George J. Mailath, and Rafael Rob**, “Learning, mutation, and long run equilibria in games,” *Econometrica*, 1993, *61* (1), 29–56.
- Kuran, Timur**, *Private Truths, Public Lies: The Social Consequences of Preference Falsification*, Harvard University Press, 1995.
- Lebow, Jeremy, Jonathan Moreno-Medina, Salma Mousa, and Horacio Coral**, “Migrant exposure and anti-migrant sentiment: The case of Venezuelan exodus,” *Journal of Public Economics*, 2024, *236*, Article No. 105169.
- Lowe, Matt**, “Types of contact: A field experiment on collaborative and adversarial caste integration,” *American Economic Review*, 2021, *111* (6), 1807–1844.
- , “Has intergroup contact delivered?,” *Annual Review of Economics*, 2025, *17*, 321–344.
- Lutz, Byron F.**, “The end of court-ordered desegregation,” *American Economic Journal: Economic Policy*, 2011, *3* (2), 130–168.
- Mousa, Salma**, “Building social cohesion between Christians and Muslims through soccer in post-ISIS Iraq,” *Science*, 2020, *369* (6505), 866–870.

- , **Lennard Naumann**, and **Alexandra Scacco**, “Intergroup contact, empathy education, and refugee-native integration: Evidence from a field experiment in Lebanon,” *Working Paper*, 2025.
- Paluck, Elizabeth Levy, Roni Porat, Chelsey S. Clark, and Donald P. Green**, “Prejudice reduction: Progress and challenges,” *Annual Review of Psychology*, 2021, 72, 533–560.
- , **Seth A. Green**, and **Donald P. Green**, “The contact hypothesis re-evaluated,” *Behavioural Public Policy*, 2019, 3 (2), 129–158.
- Pettigrew, Thomas F. and Linda R. Tropp**, “A meta-analytic test of intergroup contact theory,” *Journal of Personality and Social Psychology*, 2006, 90 (5), 751–783.
- Rao, Gautam**, “Familiarity does not breed contempt: Generosity, discrimination, and diversity in Delhi schools,” *American Economic Review*, 2019, 109 (3), 774–809.
- Scacco, Alexandra and Shana S. Warren**, “Can social contact reduce prejudice and discrimination? Evidence from a field experiment in Nigeria,” *American Political Science Review*, 2018, 112 (3), 654–677.
- Schelling, Thomas C.**, “Dynamic models of segregation,” *Journal of Mathematical Sociology*, 1971, 1 (2), 143–186.
- Shayo, Moses**, “Social identity and economic policy,” *Annual Review of Economics*, 2020, 12, 355–389.
- Steinmayr, Andreas**, “Contact versus exposure: Refugee presence and voting for the far right,” *Review of Economics and Statistics*, 2021, 103 (2), 310–327.
- Tajfel, Henri and John C. Turner**, “An Integrative Theory of Intergroup Conflict,” in W. G. Austin and S. Worchel, eds., *The Social Psychology of Intergroup Relations*, Brooks/Cole, 1979, pp. 33–47.
- , **Michael G. Billig, R. P. Bundy, and Claude Flament**, “Social categorization and intergroup behaviour,” *European Journal of Social Psychology*, 1971, 1 (2), 149–178.
- Tropp, Linda R. and Thomas F. Pettigrew**, “Relationships between intergroup contact and prejudice among minority and majority status groups,” *Psychological Science*, 2005, 16 (12), 951–957.
- Young, H. Peyton**, “The evolution of conventions,” *Econometrica*, 1993, 61 (1), 57–84.

## Appendix

### Proof of Lemma 1.

*Proof.* Conditional on  $i$  and  $j$  being matched with type-bias pairs  $(\tau_i, b_i)$  and  $(\tau_j, b_j)$ ,  $i$  chooses effort

$$e_i^*(b_i, b_j) = \arg \max_{e_i} \left( s(e_i, e_j) + (1 - s(e_i, e_j))[\beta(b_i, b_j) + (1 - \beta(b_i, b_j))(1 - \theta(\bar{\pi}))]V - \frac{1}{2}e_i^2 \right),$$

where  $s(e_i, e_j) = e_i + e_j - e_i e_j$ . Each  $j$  performs an analogous optimization. The corresponding first-order condition, evaluated at  $e_i = e_i^*$  and  $e_j = e_j^*$ , is

$$e_i^* = (1 - e_j^*)(1 - \beta(b_i, b_j))\theta(\bar{\pi})V \iff \frac{e_i^*}{1 - e_j^*} = (1 - \beta(b_i, b_j))\theta(\bar{\pi})V.$$

As  $\beta(b_i, b_j) = \beta(b_j, b_i)$ ,  $e_i^* = e_j^*$ . Setting  $e_i^* = e_j^* = e^*$ , substituting, and solving for  $e^*$  yields the expression in the lemma. The comparative statics follow immediately.  $\square$

The following proposition formalizes the individual-feedback benchmark results.

**Proposition A1.** *Fix  $s(e_i, e_j) = 0$  and  $\theta(\bar{\pi}) = 1$  for all  $e_i, e_j$ , and  $\bar{\pi}$ . The biased steady state always exists. If  $\varepsilon = 0$  and  $\mu > -1$ , then the following statements hold.*

- (i) *Almost all initial states converge to the biased steady state iff either  $\beta_1 < \frac{1}{2}$  or  $\beta_1 = \frac{1}{2}$  and  $\beta_0 < 1$ .*
- (ii) *If  $\beta_0 = 1$ , then an unbiased steady state exists, to which almost all initial states converge iff  $\beta_1 > \frac{1}{2}$ .*
- (iii) *If  $\beta_1 > \frac{1}{2}$  and  $\beta_0 < 1$ , then there exists a unique interior steady state to which almost all states converge. If  $\beta_1 = \frac{1}{2}$  and  $\beta_0 = 1$ , then each initial state converges to  $\pi_\tau^\infty = q\pi_1^0 + (1 - q)\pi_2^0$ .*

If  $\varepsilon > 0$  and  $\mu > -1$ , then the following statements hold.

- (iv) *If  $\frac{\partial F_1(1; \varepsilon)}{\partial \pi_2} \times \frac{\partial F_2(1; \varepsilon)}{\partial \pi_1} \leq 1$ , then the biased steady state is the unique steady state.*
- (v) *If  $\frac{\partial F_1(1; \varepsilon)}{\partial \pi_2} \times \frac{\partial F_2(1; \varepsilon)}{\partial \pi_1} > 1$ , then there exists an interior (asymmetric) steady state, which is componentwise strictly increasing in  $\varepsilon$ , to which almost all initial conditions converge.*

If  $\mu = -1$ , then  $\pi_\tau^\infty = 1$  when  $\varepsilon > 0$  while  $\pi_\tau^\infty = \pi_\tau^0$  when  $\varepsilon = 0$ .

The following two Lemmas are useful for proving Proposition A1

**Lemma 5.** *Suppose that  $s(e_i, e_j) = 0$  for all  $e_i, e_j$ ,  $\theta(\bar{\pi}) = 1$  for all  $\bar{\pi}$ ,  $\varepsilon = 0$ , and  $\mu > -1$ . If  $(\pi_1^\infty, \pi_2^\infty)$  solves  $\Delta\pi_\tau = 0$  for  $\tau = 1, 2$ , then  $\pi_1^\infty = \pi_2^\infty$ .*

*Proof.* Set  $s(e_i, e_j) = 0$  for all  $e_i, e_j$  and  $\theta(\bar{\pi}) = 1$  for all  $\bar{\pi}$ . Then, by (9),

$$\Delta\pi_\tau = (1 - \eta_\tau) \left[ (1 - \pi_\tau) \left( (1 - \pi_{-\tau})(1 - \beta_0) + \pi_{-\tau}(1 - \beta_1) \right) - \pi_\tau(1 - \pi_{-\tau})\beta_1 \right], \quad (25)$$

If  $\mu > -1$ , then  $\eta_1, \eta_2 \in (0, 1)$ , so  $\Delta\pi_\tau = 0$  if and only if

$$\underbrace{(1 - \pi_\tau) \left( (1 - \pi_{-\tau})(1 - \beta_0) + \pi_{-\tau}(1 - \beta_1) \right) - \pi_\tau(1 - \pi_{-\tau})\beta_1}_{\equiv F_\tau(\pi_\tau, \pi_{-\tau})} = 0$$

for  $\tau = 1, 2$ , which is equivalent to  $F_1(\pi_1, \pi_2) = 0$  and  $F_2(\pi_2, \pi_1) = 0$ . Subtracting  $F_2(\pi_2, \pi_1)$  from  $F_1(\pi_1, \pi_2)$  yields  $F_1(\pi_1, \pi_2) - F_2(\pi_2, \pi_1) = \pi_2 - \pi_1$ . Thus,  $F_1(\pi_1^\infty, \pi_2^\infty) = F_2(\pi_2^\infty, \pi_1^\infty) = 0$  implies  $\pi_1^\infty = \pi_2^\infty$ .  $\square$

**Lemma 6.** Suppose that  $s(e_i, e_j) = 0$  for all  $e_i, e_j$ ,  $\theta(\bar{\pi}) = 1$  for all  $\bar{\pi}$ ,  $\varepsilon = 0$ ,  $\mu > -1$ ,  $\beta_0 = 1$ , and  $\beta_1 = \frac{1}{2}$ . Given any initial shares  $\pi_1^0$  and  $\pi_2^0$ , there is convergence to  $\pi_1^\infty = \pi_2^\infty = q\pi_1^0 + (1-q)\pi_2^0$ .

*Proof.* Given the conditions set forth in the lemma, the law of motion reduces to

$$\begin{pmatrix} \pi_1^{t+1} \\ \pi_2^{t+1} \end{pmatrix} = \underbrace{\begin{pmatrix} 1 - \frac{(1-q)(1+\mu)}{2} & \frac{(1-q)(1+\mu)}{2} \\ \frac{q(1+\mu)}{2} & 1 - \frac{q(1+\mu)}{2} \end{pmatrix}}_{\equiv \mathbf{Q}} \begin{pmatrix} \pi_1^t \\ \pi_2^t \end{pmatrix},$$

with eigenvalues 1 and  $\frac{1-\mu}{2} \in [0, 1)$  for  $\mu > -1$ . A direct calculation confirms that  $q\pi_1^{t+1} + (1-q)\pi_2^{t+1} = q\pi_1^t + (1-q)\pi_2^t$ , so the weighted average  $q\pi_1^t + (1-q)\pi_2^t$  is constant at  $q\pi_1^0 + (1-q)\pi_2^0$ . As the second eigenvalue lies in  $[0, 1)$ , the solution converges to a multiple of the unit eigenvector  $(1, 1)^T$ , so  $\pi_1^t - \pi_2^t \rightarrow 0$ . Taking  $t \rightarrow \infty$  with  $\pi_1^\infty = \pi_2^\infty \equiv \pi^\infty$  and using the preserved quantity gives  $(q+(1-q))\pi^\infty = q\pi_1^0 + (1-q)\pi_2^0$ , hence  $\pi_1^\infty = \pi_2^\infty = q\pi_1^0 + (1-q)\pi_2^0$ .  $\square$

### Proof of Proposition A1.

*Proof.* Set  $s(e_i, e_j) = 0$  for all  $e_i, e_j$ ,  $\theta(\bar{\pi}) = 1$  for all  $\bar{\pi}$ , and  $\varepsilon = 0$ . First, observe that evaluating (9) at  $(\pi_1, \pi_2) = (1, 1)$  yields  $\Delta\pi_\tau = 0$  for each  $\tau$ , so the biased steady state always exists.

To prove statements (i)–(iii), suppose  $\varepsilon = 0$  and  $\mu > -1$ . By Lemma 5, any steady state is symmetric, so  $\pi_1^\infty = \pi_2^\infty = \pi^*$  and the steady states are characterized by

$$F(\pi^*) = 0 \iff (1 - \pi^*)[1 - \beta_0 + \pi^*(\beta_0 - 2\beta_1)] = 0. \quad (26)$$

From (26), the bracketed factor of  $F(\pi)$  is linear in  $\pi$ , equaling  $1 - \beta_0 \geq 0$  at  $\pi = 0$  and  $1 - 2\beta_1$  at  $\pi = 1$ . If  $\beta_1 < \frac{1}{2}$ , the bracket is positive on all of  $[0, 1)$ , so  $F > 0$  and the biased steady state is globally attracting, proving statement (i) (the knife-edge  $\beta_0 = 1$ ,  $\beta_1 = \frac{1}{2}$  is Lemma 6). If  $\beta_0 = 1$ , then  $F(\pi) = \pi(1-\pi)(1-2\beta_1)$ , so the unbiased steady state exists, and  $F < 0$  on  $(0, 1)$  whenever  $\beta_1 > \frac{1}{2}$ , proving statement (ii). If  $\beta_0 < 1$  and  $\beta_1 > \frac{1}{2}$ , the unique interior root  $\pi^* = \frac{1-\beta_0}{2\beta_1-\beta_0} \in (0, 1)$  is the unique interior steady state, and the sign change of  $F$  at  $\pi^*$  together with monotonicity of the joint map implies global convergence from almost all initial conditions, proving statement (iii).

To prove statements (iv) and (v), suppose that  $\varepsilon > 0$  and  $\mu > -1$ . Steady states solve the fixed-point system  $\pi_1 = F_1(\pi_2; \varepsilon)$  and  $\pi_2 = F_2(\pi_1; \varepsilon)$ , where  $F_\tau(\cdot; \varepsilon)$  is given by (10). Note that  $F_\tau(\cdot; \varepsilon)$  is continuous and strictly increasing on  $[0, 1]$  since  $\frac{\partial F_\tau(\pi_{-\tau}; \varepsilon)}{\partial \pi_{-\tau}} > 0$  for all  $\pi_{-\tau} \in [0, 1]$  by (11). Moreover, the biased steady state  $(\pi_1^\infty, \pi_2^\infty) = (1, 1)$  is always a steady state because  $F_\tau(1; \varepsilon) = 1$  for each  $\tau$ .

Define the composite map  $H(x; \varepsilon) \equiv F_1(F_2(x; \varepsilon); \varepsilon)$  and  $g(x; \varepsilon) \equiv H(x; \varepsilon) - x$ . Then  $(\pi_1^\infty, \pi_2^\infty)$  is a steady state iff  $\pi_1^\infty$  solves  $g(\pi_1^\infty; \varepsilon) = 0$  and  $\pi_2^\infty = F_2(\pi_1^\infty; \varepsilon)$ . Differentiating  $H$  and evaluating at  $x = 1$  yields

$$\frac{dH(1; \varepsilon)}{dx} = \frac{\partial F_1(1; \varepsilon)}{\partial \pi_2} \times \frac{\partial F_2(1; \varepsilon)}{\partial \pi_1}.$$

Suppose that  $\frac{dH(1; \varepsilon)}{dx} \leq 1$ . As  $F_\tau(\cdot; \varepsilon)$  is strictly increasing,  $H(\cdot; \varepsilon)$  is strictly convex on  $[0, 1]$ , so  $\frac{dH(x; \varepsilon)}{dx} \leq \frac{dH(1; \varepsilon)}{dx} \leq 1$  for all  $x \in [0, 1)$ . As  $\varepsilon > 0$  implies  $H(0; \varepsilon) > 0$ , we have  $g(0; \varepsilon) > 0$ , and combined with  $g'(x; \varepsilon) < 0$  on  $[0, 1)$ , it follows that  $g(x; \varepsilon) > 0$  on  $[0, 1)$ , so no interior steady state exists, proving statement (iv).

Now, suppose that  $\frac{dH(1;\varepsilon)}{dx} > 1$ . Then,  $\frac{dg(1;\varepsilon)}{dx} > 0$ , so for some small  $\delta > 0$ ,  $g(1 - \delta; \varepsilon) = H(1 - \delta; \varepsilon) - (1 - \delta) < 0$ . However,  $\varepsilon > 0$  implies that  $F_\tau(x; \varepsilon) > 0$  for all  $x \in [0, 1]$ , so  $H(0; \varepsilon) > 0$ . Hence,  $g(0; \varepsilon) > 0$ . By continuity, there exists a value  $x^* \in (0, 1)$  with  $g(x^*; \varepsilon) = 0$ , yielding an interior steady state  $(\pi_1^\infty, \pi_2^\infty) = (x^*, F_2(x^*; \varepsilon))$ . Finally, for any  $p \in [0, 1]$ ,  $\frac{\partial F_\tau(p; \varepsilon)}{\partial \varepsilon} > 0$ , so both  $F_\tau(\cdot; \varepsilon)$  shift up pointwise in  $\varepsilon$ , and thus so does  $H(\cdot; \varepsilon)$ . As  $H(\cdot; \varepsilon)$  is linear-fractional,  $H(x; \varepsilon) - x$  is quadratic with root  $x = 1$ , hence there is at most one additional (interior) fixed point. Because the interior fixed point is unique, it follows that  $x^*(\varepsilon)$  is strictly increasing in  $\varepsilon$ , and therefore  $\pi_2^\infty(\varepsilon) = F_2(x^*(\varepsilon); \varepsilon)$  is strictly increasing as well, proving statement (v).

As  $\eta_\tau = 1$  for  $\tau = 1, 2$  when  $\mu = -1$ , the final statement follows immediately from evaluating (9) at  $\eta_\tau = 1$ .  $\square$

The following supporting results for Section 3.3 is useful for proving Proposition 1.

**Proposition A2.** *Fix  $s(e_i, e_j) = 0$  and  $\beta(b_i, b_j) = 0$  for all  $e_i, e_j, b_i$ , and  $b_j$ . The biased steady state exists. Moreover, if  $\varepsilon = 0$  and  $\mu > -1$ , then the following statements hold.*

- (i) *The unbiased steady state exists.*
- (ii) *If  $\theta(\bar{\pi})$  is convex, then almost all initial conditions converge to the unbiased steady state.*
- (iii) *If  $\theta(\bar{\pi})$  is linear, then initial conditions  $(\pi_1^0, \pi_2^0)$  converge to the steady state*

$$\pi_1^\infty = \pi_2^\infty = \frac{q^2}{q^2 + (1-q)^2} \pi_1^0 + \frac{(1-q)^2}{q^2 + (1-q)^2} \pi_2^0.$$

If  $\varepsilon > 0$  and  $\mu > -1$ , then the following statements hold.

- (iv) *If  $\theta(\bar{\pi})$  is linear, then all initial conditions converge to the biased steady state.*
- (v) *If  $\theta(\bar{\pi})$  is convex and  $\bar{\gamma} \frac{d\theta(1)}{d\bar{\pi}} > 1$ , then in addition to the biased steady state, there is a unique interior steady state  $\bar{\pi}^* \in (0, 1)$  to which almost all initial conditions converge. Otherwise, the biased steady state is the unique globally stable steady state.*

If  $\mu = -1$ , then  $\pi_\tau^\infty = 1$  when  $\varepsilon > 0$  and  $\pi_\tau^\infty = \pi_\tau^0$  when  $\varepsilon = 0$ .

*Proof.* Fix  $s(e_i, e_j) = 0$  for all  $e_i, e_j$  and  $\beta(b_i, b_j) = 0$  for all  $b_i, b_j$ . Evaluating (15) at  $\bar{\pi} = 1$  immediately gives  $1 = 1$ , so the biased steady state always exists.

To prove statements (i)–(iii), suppose that  $\varepsilon = 0$  and  $\mu > -1$ . At  $\varepsilon = 0$ ,  $\gamma_1 = \gamma_2 = 1$ , and evaluating at  $\bar{\pi} = 0$  gives  $0 = 0$ , so the unbiased steady state exists, proving statement (i). As  $\theta(0) = 0$  and both the unbiased and biased steady states exist, there are no additional steady states. Additionally,

$$\Delta \pi_\tau = (1 - \eta_\tau)(\theta(\bar{\pi}) - \pi_\tau). \tag{27}$$

Both types are pulled each period toward the same target  $\theta(\bar{\pi})$ . Hence, if  $\pi_1 > \pi_2$ , the higher type decreases relative to the lower whenever  $\theta(\bar{\pi}) \leq \pi_1$ , and the lower type increases relative to the higher whenever  $\theta(\bar{\pi}) \geq \pi_2$ . Thus, the dynamics pull  $\pi_1$  and  $\pi_2$  toward the diagonal  $\pi_1 = \pi_2$ ; in particular, any steady state must satisfy  $\pi_1 = \pi_2$  and thus  $\bar{\pi} = \theta(\bar{\pi})$ . For all  $\bar{\pi} > 0$ , the convexity of  $\theta(\bar{\pi})$  coupled with  $\theta(0) = 0$  implies that  $\theta(\bar{\pi}) - \bar{\pi} < 0$ , so there is convergence from almost all initial conditions to the unbiased steady state, , proving statement (ii).<sup>19</sup>

<sup>19</sup>If  $\theta(\bar{\pi})$  is concave, then by an analogous argument,  $\theta(\bar{\pi}) - \bar{\pi} > 0$ , so there is convergence from almost all initial conditions to the biased steady state.

Now suppose  $\theta(\bar{\pi}) = \bar{\pi}$ . The law of motion (13) reduces to

$$\begin{pmatrix} \pi_1^{t+1} \\ \pi_2^{t+1} \end{pmatrix} = \underbrace{\begin{pmatrix} 1 - (1 - \eta_1)(1 - q) & (1 - \eta_1)(1 - q) \\ (1 - \eta_2)q & 1 - (1 - \eta_2)q \end{pmatrix}}_{\equiv \mathbf{Q}} \begin{pmatrix} \pi_1^t \\ \pi_2^t \end{pmatrix},$$

with eigenvalues 1 and  $(1 - q)\eta_1 + q\eta_2 \in (0, 1)$  for  $\mu > -1$ . A direct calculation confirms that  $q(1 - \eta_2)\pi_1^{t+1} + (1 - q)(1 - \eta_1)\pi_2^{t+1} = q(1 - \eta_2)\pi_1^t + (1 - q)(1 - \eta_1)\pi_2^t$ , so the linear combination on the left is constant at its initial value, and since the second eigenvalue lies in  $(0, 1)$ , the solution converges to a multiple of the unit eigenvector  $(1, 1)^T$  so  $\pi_1^t - \pi_2^t \rightarrow 0$ , proving statement (iii).

To prove statements (iv) and (v), suppose that  $\varepsilon > 0$  and  $\mu > -1$ . In this case  $(\gamma_1, \gamma_2) \in (0, 1)^2$ . Now suppose that  $\theta(\bar{\pi})$  is linear. At  $\pi_1^0 = \pi_2^0 = 0$  (so  $\bar{\pi} = 0$ ),

$$[q\gamma_1 + (1 - q)\gamma_2](0 - 1) + 1 > 0.$$

As the biased steady state exists, it immediately follows (by an analogous argument to statement (ii)) that all initial conditions converge to the biased steady state, proving statement (iv).<sup>20</sup>

Now suppose that  $\theta(\bar{\pi})$  is convex and  $[q\gamma_1 + (1 - q)\gamma_2] \frac{d\theta(1)}{d\bar{\pi}} > 1$ . Define  $H(\bar{\pi}) \equiv [q\gamma_1 + (1 - q)\gamma_2](\theta(\bar{\pi}) - 1) + 1$  and  $g(\bar{\pi}) \equiv H(\bar{\pi}) - \bar{\pi}$ . As  $\varepsilon > 0$  and  $(\gamma_1, \gamma_2) \in (0, 1)^2$ ,  $g(0) > 0$  and  $g(1) = 0$ . The assumed condition gives  $\frac{dg(1)}{d\bar{\pi}} > 0$ , so  $g < 0$  in a left-neighborhood of 1; by the intermediate value theorem there exists  $\bar{\pi}^* \in (0, 1)$  with  $g(\bar{\pi}^*) = 0$ . The convexity of  $\theta$  implies convexity of  $g$ , so  $\bar{\pi}^*$  is unique and  $\frac{dg(\bar{\pi}^*)}{d\bar{\pi}} < 0$ , establishing global stability. Uniqueness of the constituent  $\pi_1^*, \pi_2^*$  follows because the right-hand side of (14) depends on  $\pi_\tau$  only through  $\bar{\pi}^*$ , proving statement (v). If the slope condition fails,  $g(\bar{\pi}) > 0$  for all  $\bar{\pi} < 1$  and the biased steady state is the unique globally stable steady state.

If  $\mu = -1$ , then as is the case in Proposition A1, evaluating (13) at  $\mu = -1$  (so  $\eta_\tau = 1$  for  $\tau = 1, 2$ ) yields the last statement.  $\square$

### Proof of Proposition 1.

*Proof.* Fix  $s(e_i, e_j) = 0$  for all  $e_i, e_j$  and  $\beta(b_i, b_j) = 0$  for all  $b_i, b_j$ . By the exact same argument as Proposition A2, the biased steady state exists. Suppose that  $\varepsilon > 0$  and  $\mu > -1$ . Define

$$H(\bar{\pi}) = [q\gamma_1 + (1 - q)\gamma_2](\theta(\bar{\pi}) - 1) + 1$$

as the right-hand side of (15) and  $g(\bar{\pi}) = H(\bar{\pi}) - \bar{\pi}$ . Then,  $H(0) = 1 - [q\gamma_1 + (1 - q)\gamma_2] > 0$  (as  $(\gamma_1, \gamma_2) \in (0, 1)^2$ ), so the unbiased steady state no longer exists. As  $\frac{dH(\bar{\pi})}{d\bar{\pi}} = [q\gamma_1 + (1 - q)\gamma_2] \frac{d\theta(\bar{\pi})}{d\bar{\pi}}$ , additional interior steady states exist iff  $g(\bar{\pi}) < 0$  for some  $\bar{\pi} \in (0, 1)$ . Because  $g(0) = 1 - [q\gamma_1 + (1 - q)\gamma_2] > 0$  and  $g(1) = 0$ , this condition requires that  $\frac{dg(\bar{\pi})}{d\bar{\pi}}$  becomes positive somewhere, i.e.,  $\frac{d\theta(\bar{\pi})}{d\bar{\pi}} > [q\gamma_1 + (1 - q)\gamma_2]^{-1}$  for some  $\bar{\pi}$ . Interior steady states therefore exist iff  $\frac{d\theta(\bar{x})}{d\bar{\pi}} > [q\gamma_1 + (1 - q)\gamma_2]^{-1}$ , where  $\bar{x}$  maximizes  $\frac{d\theta(\bar{\pi})}{d\bar{\pi}}$  on  $(0, 1)$ . As  $\theta$  is sigmoidal,  $g$  has exactly one local minimum followed by one local maximum on  $(0, 1)$ ; together with  $g(0) > 0$ ,  $g(1) = 0$ , and  $\frac{dg(1)}{d\bar{\pi}} < 0$ , this permits exactly two interior zeros when the condition holds. If this condition holds, there are three steady states: a stable interior steady state  $\bar{\pi}^*$ , an unstable interior tipping point  $\bar{\pi}_U^* > \bar{\pi}^*$ , and the biased steady state, with  $g(\bar{\pi}) > 0$  on  $(0, \bar{\pi}^*)$ ,  $g(\bar{\pi}) < 0$  on  $(\bar{\pi}^*, \bar{\pi}_U^*)$ , and  $g(\bar{\pi}) > 0$  on  $(\bar{\pi}_U^*, 1)$ , giving the respective basins of attraction. If this condition fails,  $g(\bar{\pi}) > 0$  for all  $\bar{\pi} \in [0, 1)$  and the biased steady state is the unique steady state, proving statement (i).

<sup>20</sup>Allowing weak concavity does not affect this argument.

Suppose that the conditions of statement (i) hold. Consider  $\mu \in (-1, \frac{1-q}{q}]$

Note that

$$\gamma_\tau = \frac{1 - \eta_\tau}{1 - \eta_\tau(1 - \varepsilon)}$$

is decreasing in  $\eta_\tau$  and thus increasing in  $\mu$ . Therefore, if  $H(\bar{\pi}_U^*(\mu)) = \bar{\pi}_U^*(\mu)$ , then for all  $\mu' > \mu$ , evaluating  $\gamma_1$  and  $\gamma_2$  at  $\mu'$  and  $\bar{\pi} = \bar{\pi}_U^*(\mu)$  yields

$$[q\gamma(\eta_1(\mu'), \varepsilon) + (1 - q)\gamma(\eta_2(\mu'), \varepsilon)](\theta(\bar{\pi}_U^*(\mu)) - 1) + 1 < \bar{\pi}_U^*(\mu),$$

so  $\bar{\pi}_U^*(\mu') > \bar{\pi}_U^*(\mu)$ .

I now verify that  $\bar{\pi}^*(\mu)$  is strictly decreasing in  $\mu$ . At a stable interior steady state,  $g(\bar{\pi}^*(\mu)) = 0$  and  $\frac{\partial g(\bar{\pi})}{\partial \bar{\pi}} < 0$ . Differentiating implicitly yields

$$\frac{d\bar{\pi}^*}{d\mu} = -\frac{\partial g / \partial \mu}{\partial g / \partial \bar{\pi}^*}.$$

Because  $g(\bar{\pi}) = [q\gamma_1 + (1 - q)\gamma_2](\theta(\bar{\pi}) - 1) + 1 - \bar{\pi}$  and  $\theta(\bar{\pi}) - 1 < 0$  for  $\bar{\pi} \in (0, 1)$ , an increase in  $\mu$  lowers  $g$  pointwise, so  $\frac{\partial g(\bar{\pi})}{\partial \mu} < 0$ . Combined with  $\frac{\partial g(\bar{\pi}^*)}{\partial \bar{\pi}} < 0$  at a stable fixed point,  $\frac{d\bar{\pi}^*}{d\mu} < 0$ . Hence,  $\bar{\pi}^*(\mu)$  is strictly decreasing in  $\mu$ .

Now, consider an initial  $\mu = \mu^0$  and  $\bar{\pi}^0 > \bar{\pi}_U^*(\mu^0)$ . In this case, under  $\mu = \mu^0$ , the trajectory remains in the basin of attraction of the biased steady state and converges toward it. As  $\bar{\pi}_U^*(\mu)$  is continuous in  $\mu$  (by the implicit function theorem, since  $\frac{dH(\bar{\pi}_U^*(\mu))}{d\bar{\pi}} > 1$  implies  $\frac{\partial g(\bar{\pi}_U^*)}{\partial \bar{\pi}} \neq 0$ ) and strictly increasing in  $\mu$ , and  $\bar{\pi}^0 < \bar{\pi}_U^*(\frac{1-q}{q})$  by hypothesis, there exists  $\mu' \in (\mu^0, \frac{1-q}{q})$  such that  $\bar{\pi}^0 < \bar{\pi}_U^*(\mu')$ , so at  $\mu = \mu'$ ,  $\bar{\pi}^0$  is in the basin of attraction for the lower steady state.

Under  $\mu = \mu'$ , the trajectory converges to the lower interior steady state, so there exists  $t' > 0$  such that  $\bar{\pi}^{t'} < \bar{\pi}_U^*(\mu^0)$  (as  $\bar{\pi}^*(\mu') < \bar{\pi}^*(\mu^0) < \bar{\pi}_U^*(\mu^0)$ , where the first inequality follows from  $\frac{d\bar{\pi}^*}{d\mu} < 0$  and  $\mu' > \mu^0$ , and the second from above. Therefore, returning  $\mu$  to  $\mu^0$  implies that there will be persistent convergence of  $\bar{\pi}^t$  to  $\bar{\pi}^*(\mu^0)$ , proving statement (ii).  $\square$

## Proof of Lemma 2.

*Proof.* Recall from Section 2.2 that

$$\rho_{b_i, b_j} = s^*(b_i, b_j) + (1 - s^*(b_i, b_j))[\beta(b_i, b_j) + (1 - \beta(b_i, b_j))(1 - \theta(\bar{\pi}))].$$

Because efforts are symmetric in the equilibrium of the social interaction stage game, letting  $e_i^*(b_i, b_j) = e_j^*(b_i, b_j) = e^*$  (suppressing the bias values),

$$s^*(b_i, b_j) = e^* + e^* - e^* \times e^* = 1 - (1 - e^*)^2.$$

Now, consider the probability of a negative interaction between  $i$  and  $j$ :

$$\begin{aligned} 1 - \rho_{b_i, b_j} &= 1 - s^*(b_i, b_j) - (1 - s^*(b_i, b_j))[\beta(b_i, b_j) + (1 - \beta(b_i, b_j))(1 - \theta(\bar{\pi}))] \\ &= (1 - e^*)^2(1 - \beta(b_i, b_j))\theta(\bar{\pi}), \end{aligned}$$

so

$$\rho_{b_i, b_j} = 1 - (1 - e^*)^2 (1 - \beta(b_i, b_j)) \theta(\bar{\pi}),$$

Substituting (5) for  $e^*$  yields

$$\rho_{b_i, b_j} = 1 - \frac{(1 - \beta(b_i, b_j)) \theta(\bar{\pi})}{[1 + (1 - \beta(b_i, b_j)) \theta(\bar{\pi}) V]^2}.$$

Differentiating this expression with respect to  $\bar{\pi}$  yields

$$\frac{\partial \rho_{b_i, b_j}}{\partial \bar{\pi}} = -\frac{d\theta(\bar{\pi})}{d\bar{\pi}} \times \frac{(1 - \beta(b_i, b_j)) [1 - (1 - \beta(b_i, b_j)) \theta(\bar{\pi}) V]}{[1 + (1 - \beta(b_i, b_j)) \theta(\bar{\pi}) V]^3}.$$

As  $\frac{d\theta(\bar{\pi})}{d\bar{\pi}} > 0$  and  $(1 - \beta(b_i, b_j)) \in [0, 1]$ ,  $\frac{\partial \rho_{b_i, b_j}}{\partial \bar{\pi}}$  is monotonic in  $\bar{\pi}$  if and only if

$$1 - (1 - \beta(b_i, b_j)) \theta(\bar{\pi}) V$$

is monotonic; i.e., if at  $\bar{\pi} = 1$ ,

$$\begin{aligned} 1 - (1 - \beta(b_i, b_j)) V &\geq 0 \\ \iff V &\leq \frac{1}{1 - \beta(b_i, b_j)} \equiv \hat{V}, \end{aligned}$$

proving the Lemma. □

### Proof of Lemma 3.

*Proof.* Suppose that  $\mu > -1$ . Recall by (6) that

$$\Delta \pi_\tau = (1 - \eta_\tau) [1 - \bar{\rho}_\tau(\pi) - \pi_\tau] + \eta_\tau (1 - \pi_\tau) \varepsilon,$$

which when evaluated at the biased steady state  $(\pi_1, \pi_2) = (1, 1)$  yields

$$\Delta \pi_\tau = -(1 - \eta_\tau) \bar{\rho}_\tau(1).$$

Hence, for  $\mu > -1$  (so  $\eta_\tau < 1$ ), the biased steady state exists if and only if  $\bar{\rho}_\tau(1) = 0$ , where by the proof of Lemma 2,

$$\bar{\rho}_\tau(1) = \rho_{1,1} = 1 - \frac{\theta(1)}{[1 + \theta(1)V]^2} = 1 - \frac{1}{(1 + V)^2} = \frac{V(2 + V)}{(1 + V)^2} > 0.$$

Therefore, the biased steady state does not exist.

Now, evaluating (6) at the unbiased steady state  $(\pi_1, \pi_2) = (0, 0)$  yields

$$\Delta \pi_\tau = -(1 - \eta_\tau)(1 - \bar{\rho}_\tau(0)) - \eta_\tau \varepsilon.$$

Hence, for  $\mu > -1$ , the unbiased steady state exists if and only if  $\bar{\rho}_\tau(0) = 1$  and  $\varepsilon = 0$ , where by the proof of Lemma 2,

$$\bar{\rho}_\tau(0) = \rho_{0,0} = 1 - \frac{(1 - \beta_0) \theta(0)}{[1 + (1 - \beta_0) \theta(0)V]^2} = 1$$

Thus, the unbiased steady state exists whenever  $\varepsilon = 0$  and  $\mu > -1$ . □

**Proof of Proposition 2.**

*Proof.* Fix  $\mu > -1$ . Recall by (6) that

$$\Delta\pi_\tau = (1 - \eta_\tau)[1 - \bar{\rho}_\tau(\pi) - \pi_\tau] + \eta_\tau(1 - \pi_\tau)\varepsilon$$

where  $\bar{\rho}_\tau(\pi)$  is the (type- $\tau$ ) expected probability of a positive interaction under equilibrium effort.

Fix a realized pair of biases  $(b_i, b_j)$  and write  $\beta \equiv \beta(b_i, b_j) \in [0, 1]$  and  $\theta \equiv \theta(\bar{\pi})$ . From Lemma 1, equilibrium effort satisfies

$$1 - s^*(b_i, b_j) = (1 - e^*(b_i, b_j))^2 = \frac{1}{(1 + (1 - \beta)\theta V)^2},$$

so the match-level negative-interaction probability is

$$m_\beta(\bar{\pi}) \equiv \Pr(\text{negative} \mid b_i, b_j) = \frac{(1 - \beta)\theta(\bar{\pi})}{(1 + (1 - \beta)\theta(\bar{\pi})V)^2}. \quad (28)$$

Equivalently,  $\rho_{b_i, b_j}(\bar{\pi}) = 1 - m_\beta(\bar{\pi})$ .

Define

$$\tilde{p}'(\bar{\pi}) := \frac{d\theta(\bar{\pi})}{d\bar{\pi}} \times \max_{\beta \in [0, 1]} \frac{(1 - \beta)|1 - (1 - \beta)\theta(\bar{\pi})V|}{(1 + (1 - \beta)\theta(\bar{\pi})V)^3}.$$

Define  $z = (1 - \beta)\theta(\bar{\pi})V$ , so the above can be represented by the equivalent optimization problem

$$\frac{1}{\theta(\bar{\pi})V} \times \max_{z \in [0, \theta(\bar{\pi})V]} \frac{z|1 - z|}{(1 + z)^3}.$$

Solving this maximization problem yields

$$(1 - \beta)\theta V = z^* = \begin{cases} 2 - \sqrt{3} & \text{if } \theta(\bar{\pi})V \leq 2 + \sqrt{3} \\ 2 + \sqrt{3} & \text{if } \theta(\bar{\pi})V > 2 + \sqrt{3}. \end{cases}$$

As  $\frac{z^*|1 - z^*|}{(1 + z^*)^3} = \frac{\sqrt{3}}{18}$  for both  $z \leq 1$  and  $z > 1$ ,  $|\frac{dm_\beta}{d\bar{\pi}}| \leq \tilde{p}'(\bar{\pi})$ . Define  $\tilde{P}' := \sup_{\bar{\pi} \in [0, 1]} \tilde{p}'(\bar{\pi})$ .

Now, denote by

$$\bar{\Delta} = \sup_{\bar{\pi} \in [0, 1]} \max_{\beta, \beta' \in \{0, \beta_1, \beta_0\}} |m_\beta(\bar{\pi}) - m_{\beta'}(\bar{\pi})| \geq 0$$

the largest gap between the  $m_\beta(\bar{\pi})$  terms across all bias configurations. Expanding  $M_\tau(\pi) = 1 - \bar{\rho}_\tau(\pi)$  yields

$$M_\tau(\pi) = (1 - \pi_\tau)(1 - \pi_{-\tau})m_{\beta_0} + [(1 - \pi_\tau)\pi_{-\tau} + \pi_\tau(1 - \pi_{-\tau})]m_{\beta_1} + \pi_\tau\pi_{-\tau}m_{\beta_2}.$$

Differentiating this expression and letting  $q_\tau = \lambda(I_\tau)$  yields

$$\frac{\partial M_\tau}{\partial \pi_\tau} = \underbrace{(1 - \pi_{-\tau})(m_{\beta_1} - m_{\beta_0}) + \pi_{-\tau}(m_{\beta_2} - m_{\beta_1})}_{W_\tau^{\text{own}}} + q_\tau \frac{\partial M_\tau}{\partial \bar{\pi}}$$

$$\frac{\partial M_\tau}{\partial \pi_{-\tau}} = \underbrace{(1 - \pi_\tau)(m_{\beta_1} - m_{\beta_0}) + \pi_\tau(m_{\beta_2} - m_{\beta_1})}_{W_\tau^{\text{cross}}} + q_{-\tau} \frac{\partial M_\tau}{\partial \bar{\pi}}.$$

Each  $W_\tau^{\text{own}}$  is a convex combination of  $(m_{\beta_1} - m_{\beta_0})$  and  $(m_{\beta_2} - m_{\beta_1})$ , so  $|W_\tau^{\text{own}}| \leq \bar{\Delta}$ . Likewise,  $|W_\tau^{\text{cross}}| \leq \bar{\Delta}$ . By the triangle inequality and the fact that  $\left|\frac{dm_\beta}{d\bar{\pi}}\right| \leq \tilde{p}'(\bar{\pi})$ ,

$$\sum_{k=1}^2 \left| \frac{\partial M_\tau}{\partial \pi_k} \right| \leq |W_\tau^{\text{own}}| + |W_\tau^{\text{cross}}| + \sum_{k=1}^2 q_k \left| \frac{\partial M_\tau}{\partial \bar{\pi}} \right| \leq 2\bar{\Delta} + \tilde{P}'. \quad (29)$$

At a steady state, (6) implies

$$0 = (1 - \eta_\tau)[M_\tau(\pi) - \pi_\tau] + \eta_\tau(1 - \pi_\tau)\varepsilon,$$

which can be rearranged to

$$\pi_\tau = \frac{(1 - \eta_\tau)M_\tau(\pi) + \eta_\tau\varepsilon}{1 - \eta_\tau(1 - \varepsilon)} \equiv T_\tau(\pi), \quad (30)$$

$\tau \in \{1, 2\}$ . Thus, steady states are fixed points of the mapping  $T = (T_1, T_2) : [0, 1]^2 \rightarrow [0, 1]^2$ . Because  $T_\tau(\pi) = \gamma_\tau M_\tau(\pi) + (1 - \gamma_\tau)$ ,

$$\sum_{k=1}^2 \left| \frac{\partial T_\tau}{\partial \pi_k} \right| = \gamma(\eta_\tau, \varepsilon) \sum_{k=1}^2 \left| \frac{\partial M_\tau}{\partial \pi_k} \right| \leq \gamma(\eta_\tau, \varepsilon)(2\bar{\Delta} + \tilde{P}'). \quad (31)$$

As  $\gamma(\eta_\tau, \varepsilon)$  is strictly decreasing in  $\eta$  and  $\eta_1 \geq \eta_2$ , the maximum row sum is attained at  $\tau = 2$ . Now, assume that  $2\bar{\Delta} + \tilde{P}' < \gamma(\eta_2, \varepsilon)^{-1}$ . Then (31) implies that

$$\max_{\tau \in \{1, 2\}} \sum_{k=1}^2 \left| \frac{\partial T_\tau}{\partial \pi_k} \right| < 1$$

uniformly on  $[0, 1]^2$ , so  $T$  is a global contraction. By Banach's fixed-point theorem,  $T$  has a unique fixed point in  $[0, 1]^2$ . Lemma 3 rules out  $(1, 1)$  for  $\mu > -1$ , so the unique steady state lies in  $[0, 1)^2$ , proving statement (i).

Lastly, assume  $\varepsilon > 0$  and that  $\frac{dm_{\beta_0}(\bar{\pi})}{d\bar{\pi}}$  is single-peaked on  $(0, 1)$ , attaining its unique maximum at  $\bar{\pi}_M \in (0, 1)$ . Let  $M_\tau(\pi) = m_{\beta_0} + R_\tau(\pi)$ , where  $|R_\tau(\pi)| \leq \bar{\Delta}$  for all  $\pi \in [0, 1]^2$  and  $\bar{\gamma} = q\gamma_1 + (1 - q)\gamma_2$ . Define the aggregate map

$$\Psi(\bar{\pi}) = \bar{\gamma}m_{\beta_0}(\bar{\pi}) + \sum_{\tau} q_\tau \frac{\eta_\tau\varepsilon}{1 - \eta_\tau(1 - \varepsilon)} > 0,$$

so that at any steady state,  $\bar{\pi} = \Psi(\bar{\pi}) + \mathcal{E}(\pi)$  where  $\mathcal{E}(\pi)$  is an error term such that  $|\mathcal{E}(\pi)| \leq \bar{\Delta}$  uniformly. As  $\frac{d\Psi}{d\bar{\pi}} = \bar{\gamma} \times \frac{dm_{\beta_0}}{d\bar{\pi}}$  and  $\frac{dm_{\beta_0}}{d\bar{\pi}}$  is single-peaked at  $\bar{\pi}_M$  by hypothesis,  $\frac{d\Psi}{d\bar{\pi}}$  is also single-peaked at  $\bar{\pi}_M$ . The conditions

$$\frac{dm_{\beta_0}(\bar{\pi}_M)}{d\bar{\pi}} > \bar{\gamma}^{-1}, \quad \tilde{p}'(0), \tilde{p}'(1) < \bar{\gamma}^{-1}$$

imply that  $\frac{d\Psi}{d\bar{\pi}} < 1$  near  $\bar{\pi} = 0$  and  $\bar{\pi} = 1$  (using  $\frac{dm_{\beta_0}}{d\bar{\pi}} \leq \tilde{p}'$  at the endpoints), while  $\frac{d\Psi}{d\bar{\pi}} > 1$  at  $\bar{\pi}_M$ . As

$$\Psi(0) = \sum_{\tau} q_\tau \frac{\eta_\tau\varepsilon}{1 - \eta_\tau(1 - \varepsilon)}$$

$\Psi$  lies above the diagonal near 0 and  $\Psi(1) < 1$  (as  $m_{\beta_0}(1) < 1$  and  $\bar{\gamma} < 1$ ), the graph of  $\Psi$  crosses the 45-degree line exactly three times at  $\bar{\pi}_1^* < \bar{\pi}_2^* < \bar{\pi}_3^*$ .

For each  $\bar{\pi}_k^*$ , the fixed-point equations  $\pi_\tau = T_\tau(\pi)$  subject to  $q\pi_1 + (1-q)\pi_2 = \bar{\pi}_k^*$  reduce to a single equation in one free variable. By the monotonicity and continuity of  $T_\tau$ , this equation has at least one solution, yielding a steady state  $\pi^{(k)} \in (0, 1)^2$ .

More precisely, at each crossing  $\bar{\pi}_k^*$ , define the two-dimensional map  $\Phi(\pi) = T(\pi) - \pi$ . At  $\bar{\Delta} = 0$  (i.e.,  $\beta_1 = \beta_0$  and  $\beta_2 = \beta_0$ ), the system reduces exactly to the one-dimensional map.  $M_\tau(\pi)$  depends on  $\pi$  only through  $\bar{\pi}$ , so the fixed-point equations collapse to  $\bar{\pi} = \Psi(\bar{\pi})$  together with  $\pi_\tau = \gamma_\tau m_{\beta_0}(\bar{\pi}) + (1 - \gamma_\tau)$ . Each of the three aggregate crossings then yields a unique  $(\pi_1^{(k)}, \pi_2^{(k)})$  at which  $\Phi = 0$ . At the two stable crossings ( $k = 1, 3$ ),  $\frac{d\Psi}{d\bar{\pi}} < 1$ , and at the unstable crossing ( $k = 2$ ),  $\frac{d\Psi}{d\bar{\pi}} > 1$ . In each case  $D\Phi$  is nonsingular at the corresponding fixed point. By the implicit function theorem, each fixed point persists under small perturbations of the bias parameters. As  $\bar{\Delta}$  is continuous in  $(\beta_0, \beta_1, \beta_2)$  and equals zero when  $\beta_0 = \beta_1 = \beta_2$ , the three fixed points persist for all  $\bar{\Delta} < \hat{\Delta}$  for some  $\hat{\Delta} > 0$ .<sup>21</sup>

Let  $g(\bar{\pi}) = \Psi(\bar{\pi}) - \bar{\pi}$ . By the above analysis,  $g$  has exactly three zeros at  $\bar{\pi}_1^* < \bar{\pi}_2^* < \bar{\pi}_3^*$ . On the interval  $[\bar{\pi}_1^*, \bar{\pi}_2^*]$ ,  $g$  is strictly negative (as  $\Psi$  crosses the diagonal from above at  $\bar{\pi}_1^*$  and from below at  $\bar{\pi}_2^*$ ). The error  $\mathcal{E}(\pi)$  perturbs each crossing level by at most  $\bar{\Delta}$ . Define

$$d_1 = - \min_{\bar{\pi} \in [\bar{\pi}_1^*, \bar{\pi}_2^*]} g(\bar{\pi}) > 0.$$

Similarly, on  $[\bar{\pi}_2^*, \bar{\pi}_3^*]$ ,  $g$  is strictly positive. Define

$$d_2 = \max_{\bar{\pi} \in [\bar{\pi}_2^*, \bar{\pi}_3^*]} g(\bar{\pi}) > 0.$$

The full steady-state system requires  $g(\bar{\pi}) = \mathcal{E}(\pi)$ , where  $|\mathcal{E}(\pi)| \leq \bar{\gamma}\bar{\Delta}$  uniformly.

Assume  $\bar{\Delta} < \min(d_1/\bar{\gamma}, d_2/\bar{\gamma}, (1 - \bar{\gamma})/\bar{\gamma}, 1 - m_{\beta_0}(1)) \equiv \hat{\Delta}$ . At  $\bar{\pi} = 0$ ,  $g(0) = 1 - \bar{\gamma} > 0$ . At the minimizer  $\bar{\pi}_{\min} \in (\bar{\pi}_1^*, \bar{\pi}_2^*)$ ,  $g(\bar{\pi}_{\min}) = -d_1 < -\bar{\gamma}\bar{\Delta}$ . Hence, for any  $\mathcal{E}$  with  $|\mathcal{E}| \leq \bar{\gamma}\bar{\Delta}$ , the  $g(\bar{\pi}) - \mathcal{E}$  satisfies  $g(0) - \mathcal{E} \geq (1 - \bar{\gamma}) - \bar{\gamma}\bar{\Delta} > 0$  and  $g(\bar{\pi}_{\min}) - \mathcal{E} \leq -d_1 + \bar{\gamma}\bar{\Delta} < 0$ . By the intermediate value theorem, there exists a root  $\hat{\pi}_1 \in (0, \bar{\pi}_{\min})$ . At the maximizer  $\bar{\pi}_{\max} \in (\bar{\pi}_2^*, \bar{\pi}_3^*)$ ,  $g(\bar{\pi}_{\max}) = d_2 > \bar{\gamma}\bar{\Delta}$ . As  $g(\bar{\pi}_{\min}) - \mathcal{E} < 0$  and  $g(\bar{\pi}_{\max}) - \mathcal{E} > d_2 - \bar{\gamma}\bar{\Delta} > 0$ ,  $g(\bar{\pi}) - \mathcal{E}$  changes sign on  $[\bar{\pi}_{\min}, \bar{\pi}_{\max}]$ , yielding a root  $\hat{\pi}_2$ . Finally,  $g(1) = \Psi(1) - 1 < 0$  (as  $m_{\beta_0}(1) < 1$  and  $\bar{\gamma} < 1$ ), so  $g(1) - \mathcal{E} \leq g(1) + \bar{\gamma}\bar{\Delta}$ . Because  $g(1) = -\bar{\gamma}(1 - m_{\beta_0}(1)) < 0$ , this remains strictly negative for  $\bar{\Delta} < 1 - m_{\beta_0}(1)$ . As  $g(\bar{\pi}_{\max}) - \mathcal{E} > 0$ ,  $g(\bar{\pi}) - \mathcal{E}$  changes sign on  $[\bar{\pi}_{\max}, 1]$ , yielding a third root  $\hat{\pi}_3$ .

For each aggregate crossing  $\hat{\pi}_k$ , the fixed-point equations  $\pi_\tau = T_\tau(\pi)$  subject to  $q\pi_1 + (1-q)\pi_2 = \hat{\pi}_k$  reduce to a single equation in one free variable. By the monotonicity and continuity of  $T_\tau$ , and because  $T_\tau$  maps  $[0, 1]^2$  into  $(0, 1)^2$  (since  $\varepsilon > 0$  and  $V > 0$ ), this equation has at least one solution, yielding a steady state  $\pi^{(k)} \in (0, 1)^2$ . The three crossings  $\hat{\pi}_1 < \hat{\pi}_2 < \hat{\pi}_3$  are distinct, so the corresponding steady states are also distinct, proving statement (ii).  $\square$

**Proposition A3.** *Suppose  $\varepsilon > 0$  and parameters are contained in  $\Omega$  for all  $\mu \in [\mu^0, \frac{1-q}{q}]$ . If  $\bar{\pi}^0 \in (\bar{\pi}_2^*(\mu^0), \bar{\pi}_2^*(\frac{1-q}{q}))$ , then there exists a  $\mu' > \mu^0$  such that temporarily increasing  $\mu^0$  to  $\mu'$  yields a persistent reduction in long-run bias with convergence to  $\bar{\pi}_1^*(\mu^0)$ .*

<sup>21</sup>The binding constraint is not proximity of  $\beta_1$  to  $\beta_0$  per se, but rather that the amplitude of the sigmoidal nonlinearity in  $\Psi$  exceeds the perturbation introduced by heterogeneous  $\beta$  values. Concretely,  $\hat{\Delta}$  is bounded below by  $\min_k |1 - \frac{d\Psi(\bar{\pi}_k^*)}{d\bar{\pi}}|$  scaled by a Lipschitz constant. When the three crossings of  $\Psi$  with the diagonal are well separated and steep, the fixed points survive large differences between  $\beta_0$  and  $\beta_1$ . The condition  $\bar{\Delta} < \hat{\Delta}$  is therefore least restrictive precisely when tipping dynamics are most pronounced.

*Proof.* From the proof of Proposition 2, the aggregate map is

$$\Psi(\bar{\pi}; \mu) = \bar{\gamma}(\mu)m_{\beta_0}(\bar{\pi}) + \sum_{\tau} q_{\tau} \frac{\eta_{\tau}(\mu)\varepsilon}{1 - \eta_{\tau}(\mu)(1 - \varepsilon)}.$$

As  $1 - \gamma_{\tau} = \frac{\eta_{\tau}\varepsilon}{(1 - \eta_{\tau}(1 - \varepsilon))}$ , summing gives  $\sum_{\tau} q_{\tau}(1 - \gamma_{\tau}) = 1 - \bar{\gamma}$ , so

$$\Psi(\bar{\pi}; \mu) = 1 - \bar{\gamma}(\mu)(1 - m_{\beta_0}(\bar{\pi})).$$

Define  $g(\bar{\pi}; \mu) \equiv \Psi(\bar{\pi}; \mu) - \bar{\pi}$ . Because  $\bar{\gamma}(\mu)$  is strictly increasing in  $\mu$  (since  $\eta_{\tau}$  is strictly decreasing in  $\mu$  and  $\gamma_{\tau} = \frac{(1 - \eta_{\tau})}{(1 - \eta_{\tau}(1 - \varepsilon))}$  is strictly decreasing in  $\eta_{\tau}$ ), and  $m_{\beta_0}(\bar{\pi}) < 1$  for all  $\bar{\pi} \in [0, 1]$  (as  $V > 0$ ), it follows that

$$\frac{\partial g(\bar{\pi}; \mu)}{\partial \mu} = -\frac{d\bar{\gamma}}{d\mu}(1 - m_{\beta_0}(\bar{\pi})) < 0$$

for all  $\bar{\pi} \in [0, 1)$ .

At the unstable crossing  $\bar{\pi}_2^*$ , the map  $\Psi$  crosses the diagonal from below, so  $\frac{\partial g(\bar{\pi}; \mu)}{\partial \bar{\pi}} > 0$ . The implicit function theorem applied to  $g(\bar{\pi}_2^*(\mu); \mu) = 0$  yields

$$\frac{d\bar{\pi}_2^*}{d\mu} = -\frac{\partial g/\partial \mu}{\partial g/\partial \bar{\pi}} > 0,$$

so  $\bar{\pi}_2^*(\mu)$  is strictly increasing in  $\mu$ . At  $\bar{\pi}_1^*$ , the map  $\Psi$  crosses from above, so  $\partial g/\partial \bar{\pi} < 0$ . By the same argument,  $\frac{d\bar{\pi}_1^*}{d\mu} < 0$ , so  $\bar{\pi}_1^*(\mu)$  is strictly decreasing in  $\mu$ .

The remainder of the argument follows the proof of Proposition 1. As  $\bar{\pi}_2^*(\mu)$  is continuous and strictly increasing in  $\mu$  with  $\bar{\pi}^0 < \bar{\pi}_2^*(\frac{1-q}{q})$  by hypothesis, there exists  $\mu' \in (\mu^0, \frac{1-q}{q})$  such that  $\bar{\pi}^0 < \bar{\pi}_2^*(\mu')$ . Thus,  $\bar{\pi}_1^*(\mu') < \bar{\pi}_1^*(\mu^0) < \bar{\pi}_2^*(\mu^0) < \bar{\pi}^0 < \bar{\pi}_2^*(\mu')$ , and the trajectory from  $\bar{\pi}^0$  under  $\mu'$  lies in the basin of attraction of  $\bar{\pi}_1^*(\mu')$  and converges toward it. Because  $\bar{\pi}_1^*(\mu') < \bar{\pi}_1^*(\mu^0) < \bar{\pi}_2^*(\mu^0)$  (using strict decrease of  $\bar{\pi}_1^*$  and the three-steady-state ordering), there exists  $t' > 0$  such that  $\bar{\pi}^{t'} < \bar{\pi}_2^*(\mu^0)$ . Returning  $\mu$  to  $\mu^0$  places the trajectory in the basin of attraction of  $\bar{\pi}_1^*(\mu^0)$ , yielding persistent convergence to  $\bar{\pi}_1^*(\mu^0) < \bar{\pi}^0$ .  $\square$

The following Lemma is useful for proving Proposition 3.

**Lemma 7.** *For all  $\bar{\pi} \in [0, 1]$  and  $V \geq 0$ ,  $\frac{dm_{\beta_0}(\bar{\pi})}{d\bar{\pi}} \leq \frac{d\theta(\bar{\pi})}{d\bar{\pi}}$ , with strict inequality for all  $\bar{\pi} > 0$  whenever  $V > 0$  and  $\beta_0 \in [0, 1)$ .*

*Proof.* Differentiating  $m_{\beta_0}$  yields

$$\frac{dm_{\beta_0}}{d\bar{\pi}} = \frac{d\theta(\bar{\pi})}{d\bar{\pi}} \times \frac{(1 - \beta_0)[1 - (1 - \beta_0)\theta(\bar{\pi})V]}{[1 + (1 - \beta_0)\theta(\bar{\pi})V]^3}.$$

Set  $z = (1 - \beta_0)\theta(\bar{\pi})V \geq 0$  and define

$$A(z) = \frac{(1 - \beta_0)(1 - z)}{(1 + z)^3},$$

so  $\frac{dm_{\beta_0}}{d\bar{\pi}} = \frac{d\theta(\bar{\pi})}{d\bar{\pi}} \times A(z)$ . I now show that  $A(z) < 1$  for all  $z > 0$ . If  $z \geq 1$ , then  $1 - z \leq 0$ , so  $A(z) \leq 0 < 1$ . If  $0 < z < 1$ , then  $(1 - \beta_0) < 1$  and  $(1 - z) < 1 \leq (1 + z)^3$ , so  $A(z) < 1$ . At  $z = 0$ ,  $A(0) = 1 - \beta_0 \leq 1$ , with equality only when  $\beta_0 = 0$ . Hence,  $A(z) < 1$  for all  $z > 0$ , i.e., whenever  $\theta(\bar{\pi}) > 0$ ,  $V > 0$ , and  $\beta_0 \in [0, 1)$ .  $\square$

### Proof of Proposition 3.

*Proof.* Recall from the definition of the slope envelope that

$$\tilde{p}'(\bar{\pi}; V) = \frac{d\theta(\bar{\pi})}{d\bar{\pi}} \times \max_{\beta \in [0,1]} h(\beta, \bar{\pi}, V),$$

where  $h(\beta, \bar{\pi}, V) = \frac{(1-\beta)|1-(1-\beta)\theta V|}{(1+(1-\beta)\theta V)^3}$ . Setting  $z = (1-\beta)\theta(\bar{\pi})V$ , the maximum over  $\beta \in [0, 1]$  is equivalent to

$$\max_{\beta \in [0,1]} h = \frac{1}{\theta(\bar{\pi})V} \max_{z \in [0, \theta(\bar{\pi})V]} g(z), \quad g(z) \equiv \frac{z|1-z|}{(1+z)^3}.$$

Let  $w = \theta(\bar{\pi})V > 0$ . If  $w \geq 2 - \sqrt{3}$ , the interior maximum of  $g$  at  $z^* = 2 - \sqrt{3}$  is feasible and  $g(z^*) = \frac{\sqrt{3}}{18}$ , giving  $\max_{\beta} h = \frac{\sqrt{3}}{18w}$ . If  $w < 2 - \sqrt{3}$ , the maximum of  $g$  on  $[0, w]$  is attained at  $z = w$  (since  $g$  is increasing on  $[0, 2 - \sqrt{3}]$ ), giving  $\max_{\beta} h = \frac{g(w)}{w} = \frac{1-w}{(1+w)^3}$ .

I show that  $\max_{\beta} h$  is strictly decreasing in  $V$  for each  $\bar{\pi} > 0$  in both regimes.

*Case 1:*  $w \geq 2 - \sqrt{3}$ .  $\max_{\beta} h = \frac{\sqrt{3}}{18\theta(\bar{\pi})V}$ , so  $\frac{\partial}{\partial V} \max_{\beta} h = -\frac{\sqrt{3}}{18\theta(\bar{\pi})V^2} < 0$ .

*Case 2:*  $w < 2 - \sqrt{3}$ .  $\max_{\beta} h = \frac{1-\theta V}{(1+\theta V)^3}$ , so

$$\frac{\partial}{\partial V} \max_{\beta} h = \frac{2\theta(\theta V - 2)}{(1 + \theta V)^4} < 0,$$

where the inequality holds because  $\theta V < 2 - \sqrt{3} < 2$ .

Continuity at the boundary  $w = 2 - \sqrt{3}$  is verified by observing that both expressions equal  $\frac{\sqrt{3}-1}{(3-\sqrt{3})^3}$  at  $w = 2 - \sqrt{3}$ .

As  $\frac{d\theta}{d\bar{\pi}} > 0$  for  $\bar{\pi} > 0$  and  $\max_{\beta} h$  is strictly decreasing in  $V$ , the pointwise envelope  $\tilde{p}'(\bar{\pi}; V) = \frac{d\theta}{d\bar{\pi}} \times \max_{\beta} h$  is strictly decreasing in  $V$  for each  $\bar{\pi} > 0$ . As the supremum of pointwise-decreasing functions is decreasing,  $\tilde{P}'(V) = \sup_{\bar{\pi}} \tilde{p}'(\bar{\pi}; V)$  is weakly decreasing in  $V$ . The inequality is strict because, at any maximizer  $\bar{\pi}^*(V)$  of  $\tilde{p}'(\cdot; V)$ ,  $\theta(\bar{\pi}^*) > 0$  (since the maximizer lies in the interior where  $\theta' > 0$ ), so the pointwise strict decrease holds at the maximizer.

To verify that  $\bar{\Delta}(V)$  is strictly decreasing, recall that  $\bar{\Delta}(V) = \sup_{\bar{\pi}} \max_{\beta, \beta'} |m_{\beta}(\bar{\pi}; V) - m_{\beta'}(\bar{\pi}; V)|$ . For fixed  $\bar{\pi} > 0$  and any  $\beta \neq \beta'$ , the gap  $|m_{\beta} - m_{\beta'}|$  is strictly decreasing in  $V$  because higher  $V$  compresses all  $m_{\beta}$  terms toward zero (as  $m_{\beta}(\bar{\pi}; V) \rightarrow 0$  for each  $\beta$  as  $V \rightarrow \infty$ , with the rate depending on  $\beta$ , and the highest- $m$  term decreases fastest). Taking the supremum preserves the strict decrease by the same maximizer argument.

Continuity of  $\tilde{P}'(V)$  and  $\bar{\Delta}(V)$  follows from continuity of the constituent functions and compactness of  $[0, 1]$ , proving statement (i).

For statement (ii), I show that three steady states persist for all  $V \in (0, V')$  whenever they exist at  $V = V'$ . The argument proceeds in two steps: first establishing persistence of the three crossings in the one-dimensional aggregate map, then extending to the full two-dimensional system. Recall  $g(\bar{\pi}; V) = \Psi(\bar{\pi}; V) - \bar{\pi}$ , where  $\Psi(\bar{\pi}; V) = 1 - \bar{\gamma}(1 - m_{\beta_0}(\bar{\pi}; V))$ . I verify the four conditions that jointly ensure exactly three zeros of  $g(\cdot; V)$  for all  $V \in (0, V']$ : (a)  $g(0; V) = 1 - \bar{\gamma} > 0$ , independent of  $V$ ; (b)  $g(1; V) = -\bar{\gamma}(1 - m_{\beta_0}(1; V)) < 0$  for all  $V > 0$ , since  $m_{\beta_0}(1; V) = \frac{(1-\beta_0)}{(1+(1-\beta_0)V)^2} < 1$ ; (c) the endpoint slopes

satisfy  $\frac{d\Psi(0;V)}{d\bar{\pi}} = \bar{\gamma}(1 - \beta_0)\theta'(0) < 1$  (independent of  $V$ , using  $\theta'(0) < 1$  by the sigmoidal assumption and  $\bar{\gamma}(1 - \beta_0) < 1$ ) and  $\frac{d\Psi(1;V)}{d\bar{\pi}} \leq \bar{\gamma}\theta'(1) < 1$  (where the first inequality follows from Lemma 7 and the second from the sigmoidal assumption); (d) by Lemma 7,  $\frac{dm_{\beta_0}(\bar{\pi};V)}{d\bar{\pi}}$  is pointwise non-decreasing as  $V$  decreases, so  $\sup_{\bar{\pi}} \frac{d\Psi(\bar{\pi};V)}{d\bar{\pi}} \geq \sup_{\bar{\pi}} \frac{d\Psi(\bar{\pi};V')}{d\bar{\pi}} > 1$  for all  $V \leq V'$ . Conditions (a)–(d) together imply that  $g(\cdot; V)$  starts positive, decreases below zero (by (a), (c) at  $\bar{\pi} = 0$ , and (d)), rises above zero (by (d)), and ends negative (by (b), (c) at  $\bar{\pi} = 1$ ). By the intermediate value theorem and the sigmoidal single-peaked structure of  $\frac{d\Psi}{d\bar{\pi}}$ ,  $g(\cdot; V)$  has exactly three zeros  $\bar{\pi}_1^*(V) < \bar{\pi}_2^*(V) < \bar{\pi}_3^*(V)$  for every  $V \in (0, V']$ .

At  $V = V'$ , three steady states  $\pi_k^*(V')$  of the full map  $T(\cdot; V')$  exist by hypothesis. Each is a non-degenerate fixed point:  $\det(DT(\pi_k^*; V') - I) \neq 0$ , as established in the proof of Proposition 2(ii) via the implicit function theorem applied to the bias parameters  $(\beta_0, \beta_1, \beta_2)$ . By the implicit function theorem now applied to the parameter  $V$ , each  $\pi_k^*(V)$  extends uniquely as a  $C^1$  function of  $V$  in a neighborhood of  $V'$ . Loss of a steady state as  $V$  decreases from  $V'$  requires a bifurcation at which  $\det(DT - I) = 0$ , i.e., an eigenvalue of  $DT$  equals 1. In the unperturbed system ( $\bar{\Delta} = 0$ ), the eigenvalues of  $DT$  at a crossing  $\bar{\pi}_k^*$  are  $\bar{\gamma} \frac{dm_{\beta_0}}{d\bar{\pi}} \Big|_{\bar{\pi}_k^*}$  and 0, so the non-degeneracy gap at crossing  $k$  is  $|1 - \bar{\gamma} \frac{dm_{\beta_0}(\bar{\pi}_k^*)}{d\bar{\pi}}|$ . For  $\bar{\Delta} > 0$ , the eigenvalues are perturbed by  $O(\bar{\Delta})$ , so  $\det(DT - I)$  remains nonzero as long as  $\bar{\Delta}$  is small relative to the non-degeneracy gap. By the above argument, the three one-dimensional crossings are non-degenerate for all  $V \in (0, V']$  (the slopes  $\bar{\gamma} \frac{dm_{\beta_0}}{d\bar{\pi}}$  at the crossings are strictly different from 1 and vary continuously in  $V$ ), so the minimum non-degeneracy gap

$$\sigma \equiv \inf_{V \in (0, V')} \min_{k \in \{1, 2, 3\}} \left| 1 - \bar{\gamma} \frac{dm_{\beta_0}(\bar{\pi}_k^*(V); V)}{d\bar{\pi}} \right|$$

is strictly positive by continuity and compactness of  $[0, V']$  (the infimum extends by continuity to  $V = 0$  where the three crossings also exist). As  $\bar{\Delta}(V) \leq \bar{\Delta}(0) = \beta_0 < \infty$  is bounded, the non-degeneracy condition holds for all  $V \in (0, V']$  whenever  $\bar{\Delta}$  is small relative to  $\sigma$ , precisely the content of the maintained condition  $\bar{\Delta} < \hat{\Delta}$  from Proposition 2(ii), whose threshold  $\hat{\Delta}$  is bounded below by  $\sigma$  scaled by a Lipschitz constant (see footnote 21). Hence, no bifurcation occurs on  $(0, V']$ , the three two-dimensional steady states persist throughout, and  $\Omega(V') \subset \Omega(V)$ . □

#### Proof of Proposition 4.

*Proof.* From the proof of Proposition A3,

$$\Psi(\bar{\pi}; V) = 1 - \bar{\gamma}(1 - m_{\beta_0}(\bar{\pi}; V)).$$

Steady states of the full system correspond to fixed points of  $\Psi(\cdot; V)$  up to the perturbation  $\mathcal{E}(\pi)$  bounded by  $\bar{\Delta}$ , as established in the proof of Proposition 2.

Differentiating  $m_{\beta_0}(\bar{\pi}; V)$  with respect to  $V$  yields

$$\frac{\partial m_{\beta_0}}{\partial V} = - \frac{2(1 - \beta_0)^2 \theta(\bar{\pi})^2}{[1 + (1 - \beta_0)\theta(\bar{\pi})V]^3} < 0$$

for all  $\bar{\pi} > 0$ . Therefore

$$\frac{\partial \Psi(\bar{\pi}; V)}{\partial V} = \bar{\gamma} \frac{\partial m_{\beta_0}}{\partial V} < 0$$

for all  $\bar{\pi} \in (0, 1]$ . Define  $g(\bar{\pi}; V) \equiv \Psi(\bar{\pi}; V) - \bar{\pi}$ . At any stable steady state  $\bar{\pi}_k^*(V)$ ,  $\Psi$  crosses the diagonal from above, so  $\frac{\partial g(\bar{\pi})}{\partial \bar{\pi}} < 0$ . As  $\frac{\partial g(\bar{\pi})}{\partial V} < 0$  and  $\frac{\partial g(\bar{\pi})}{\partial \bar{\pi}} < 0$  at  $\bar{\pi}_1^*$ , the implicit function theorem gives

$$\frac{d\bar{\pi}_1^*}{dV} = -\frac{\partial g/\partial V}{\partial g/\partial \bar{\pi}} < 0,$$

so  $\bar{\pi}_1^*(V)$  is strictly decreasing in  $V$ .

From Lemma 7,

$$\frac{dm_{\beta_0}(\bar{\pi}; V)}{d\bar{\pi}} = \frac{d\theta(\bar{\pi})}{d\bar{\pi}} \times A(z), \quad z = (1 - \beta_0)\theta(\bar{\pi})V,$$

where  $A(z) \rightarrow 0$  as  $z \rightarrow \infty$ . Fix  $\delta > 0$ . For all  $\bar{\pi} \in [\delta, 1]$ ,  $\theta(\bar{\pi}) \geq \theta(\delta) > 0$ , so  $z \geq (1 - \beta_0)\theta(\delta)V \rightarrow \infty$ , and  $A(z) \rightarrow 0$  uniformly on  $[\delta, 1]$  as  $V \rightarrow \infty$ . On  $[0, \delta]$ , by the hypothesis  $\tilde{p}'(0) < \bar{\gamma}^{-1}$  in Proposition 2(ii), we have  $\bar{\gamma} \frac{dm_{\beta_0}}{d\bar{\pi}} \leq \bar{\gamma} \tilde{p}'(\bar{\pi}) < 1$  near  $\bar{\pi} = 0$  for any  $V \geq V^0$ . As  $\delta \rightarrow 0$  this region vanishes. Hence,

$$\lim_{V \rightarrow \infty} \sup_{\bar{\pi} \in [0, 1]} \frac{d\Psi(\bar{\pi}; V)}{d\bar{\pi}} = \lim_{V \rightarrow \infty} \bar{\gamma} \sup_{\bar{\pi} \in [0, 1]} \frac{dm_{\beta_0}(\bar{\pi}; V)}{d\bar{\pi}} = 0.$$

By Proposition 3, both  $\tilde{P}'(V) = \sup_{\bar{\pi}} \tilde{p}'(\bar{\pi}; V)$  and  $\bar{\Delta}(V) = \sup_{\bar{\pi}} [p(\bar{\pi}; V) - m_{\beta_0}(\bar{\pi}; V)]$  are strictly decreasing and continuous in  $V$ . By the above,  $\tilde{P}'(V) \rightarrow 0$  as  $V \rightarrow \infty$ . As  $p(\bar{\pi}; V) \rightarrow 0$  and  $m_{\beta_0}(\bar{\pi}; V) \rightarrow 0$  as  $V \rightarrow \infty$  for each  $\bar{\pi}$ , with  $p(\bar{\pi}; V) \geq m_{\beta_0}(\bar{\pi}; V)$ , the gap  $\bar{\Delta}(V) \rightarrow 0$  as well. Therefore, there exists  $V'$  such that for all  $V \geq V'$ ,

$$2\bar{\Delta}(V) + \tilde{P}'(V) < \gamma(\eta_2, \varepsilon)^{-1},$$

which is the uniqueness condition of Proposition 2(i). By that proposition, the full model admits a unique steady state  $\bar{\pi}_1^*(V) \in [0, 1]^2$  for all  $V \geq V'$ . By Lemma 3, this steady state is not  $(1, 1)$ , so  $\bar{\pi}_1^*(V) \in (0, 1)$ . Moreover,  $\bar{\pi}_1^*(V) \leq \bar{\pi}_1^*(V^0) < \bar{\pi}_2^*(V^0)$ , proving statement (i).

Now fix  $V'' \geq V'$  satisfying statement (i) and let  $\bar{\pi}^0 \in (\bar{\pi}_2^*(V^0), 1)$ . Under  $V''$ , the unique steady state  $\bar{\pi}_1^*(V'') < \bar{\pi}_2^*(V^0)$  is globally attracting by the contraction established in the proof of Proposition 2(i). Hence, the trajectory  $\{\bar{\pi}^t\}_{t \geq 0}$  converges to  $\bar{\pi}_1^*(V'')$  from  $\bar{\pi}^0$ . As  $\bar{\pi}_1^*(V'') < \bar{\pi}_2^*(V^0)$  and convergence is guaranteed in finite time since the unique steady lies strictly below  $\bar{\pi}_2^*(V_0)$ , there exists a  $t' < \infty$  such that  $\bar{\pi}^{t'} < \bar{\pi}_2^*(V^0)$ . Returning  $V$  to  $V^0$  at time  $t'$  restores the original three-steady-state configuration. Because  $\bar{\pi}^{t'} < \bar{\pi}_2^*(V^0)$ , the trajectory now lies in the basin of attraction of  $\bar{\pi}_1^*(V^0)$  under  $V^0$ , and converges to  $\bar{\pi}_1^*(V^0) < \bar{\pi}^0$ , proving statement (ii).  $\square$

#### Proof of Lemma 4.

*Proof.* Recall that  $\Delta\pi_\tau = (1 - \eta_\tau)[1 - \bar{\rho}_\tau(\pi) - \pi_\tau] + \eta_\tau(1 - \pi_\tau)\varepsilon$ . As  $\bar{\rho}_\tau(\pi) \in [0, 1]$  and  $\varepsilon \geq 0$ , the expression is minimized at  $\bar{\rho}_\tau = 1$  and  $\varepsilon = 0$ , giving  $\Delta\pi_\tau \geq -(1 - \eta_\tau)\pi_\tau$ . Therefore,  $-\Delta\pi_\tau \leq (1 - \eta_\tau)\pi_\tau \leq 1 - \eta_\tau$ , and

$$\bar{\pi}^t - \bar{\pi}^{t+1} = -(q\Delta\pi_1 + (1 - q)\Delta\pi_2) \leq q(1 - \eta_1)\pi_1 + (1 - q)(1 - \eta_2)\pi_2 \leq q(1 - \eta_1) + (1 - q)(1 - \eta_2).$$

Substituting  $1 - \eta_1 = (1 - q)(1 + \mu)$  and  $1 - \eta_2 = q(1 + \mu)$  yields  $\bar{\pi}^t - \bar{\pi}^{t+1} \leq q(1 - q)(1 + \mu) + (1 - q)q(1 + \mu) = 2q(1 - q)(1 + \mu) = C^*(\mu, q)$ .  $\square$

The following Lemma corresponds to footnote 15.

**Lemma 8.** Under  $\Omega$  with  $\mu > -1$  and  $\varepsilon > 0$ , the unstable tipping point  $\bar{\pi}_2^*(\mu, V)$  is strictly increasing in both  $\mu$  and  $V$ .

*Proof.* From the proof of Proposition A3,  $g(\bar{\pi}; \mu, V) \equiv \Psi(\bar{\pi}; \mu, V) - \bar{\pi}$  satisfies  $\frac{\partial g(\bar{\pi}_2^*)}{\partial \bar{\pi}} > 0$  at the unstable tipping point (where  $\Psi$  crosses the diagonal from below). From the proof of Proposition A3,  $\frac{\partial g}{\partial \mu} < 0$ . From the proof of Proposition 4,  $\frac{\partial g}{\partial V} < 0$ . Applying the implicit function theorem to  $g(\bar{\pi}_2^*; \mu, V) = 0$  in each argument separately yields

$$\frac{d\bar{\pi}_2^*}{d\mu} = -\frac{\partial g/\partial \mu}{\partial g/\partial \bar{\pi}} > 0, \quad \frac{d\bar{\pi}_2^*}{dV} = -\frac{\partial g/\partial V}{\partial g/\partial \bar{\pi}} > 0. \quad \square$$

### Proof of Proposition 5.

*Proof.* By Proposition 4(i), under  $(V'', \mu^0)$  the model admits a unique globally attracting steady state  $\bar{\pi}_1^*(\mu^0, V'') < \bar{\pi}_2^*(\mu^0, V^0)$ . Since the trajectory  $\{\bar{\pi}^t\}$  converges to  $\bar{\pi}_1^*(\mu^0, V'')$ , it eventually falls strictly below both  $\bar{\pi}_2^*(\mu^0, V^0)$  and  $\bar{\pi}_2^*(\tilde{\mu}, V^0)$ , so  $T_V, \tilde{T}_V < \infty$ .

By hypothesis,  $\bar{\pi}^0 \geq \bar{\pi}_2^*(\tilde{\mu}, V^0) > \bar{\pi}_2^*(\mu^0, V^0)$ , so  $\tilde{T}_V \geq 1$ . By definition of  $\tilde{T}_V$  as a minimum,  $\bar{\pi}^{\tilde{T}_V} < \bar{\pi}_2^*(\tilde{\mu}, V^0)$  and  $\bar{\pi}^{\tilde{T}_V-1} \geq \bar{\pi}_2^*(\tilde{\mu}, V^0)$ .

Suppose for contradiction that  $\bar{\pi}^{\tilde{T}_V} \leq \bar{\pi}_2^*(\mu^0, V^0)$ . Then

$$\bar{\pi}^{\tilde{T}_V-1} - \bar{\pi}^{\tilde{T}_V} \geq \bar{\pi}_2^*(\tilde{\mu}, V^0) - \bar{\pi}_2^*(\mu^0, V^0) > C^*(\mu^0, q),$$

where the second inequality is the assumption on  $C^*(\mu^0, q)$ . Lemma 4 applied under matching parameter  $\mu^0$  gives  $\bar{\pi}^{\tilde{T}_V-1} - \bar{\pi}^{\tilde{T}_V} \leq C^*(\mu^0, q)$ , a contradiction. Hence,  $\bar{\pi}^{\tilde{T}_V} \in (\bar{\pi}_2^*(\mu^0, V^0), \bar{\pi}_2^*(\tilde{\mu}, V^0))$ . In particular  $\bar{\pi}^{\tilde{T}_V} > \bar{\pi}_2^*(\mu^0, V^0)$ , so the trajectory has not yet crossed the lower threshold at time  $\tilde{T}_V$ , establishing  $\tilde{T}_V < T_V$ .

Under  $(\tilde{\mu}, V^0)$ , the maintained conditions ensure the three-steady-state structure holds. From the proof of Proposition A3,  $\bar{\pi}_1^*(\mu, V)$  is strictly decreasing in  $\mu$ , so  $\bar{\pi}_1^*(\tilde{\mu}, V^0) < \bar{\pi}_1^*(\mu^0, V^0) < \bar{\pi}_2^*(\mu^0, V^0)$ . As  $\bar{\pi}^{\tilde{T}_V} < \bar{\pi}_2^*(\tilde{\mu}, V^0)$ , the trajectory under  $(\tilde{\mu}, V^0)$  initiated at  $\bar{\pi}^{\tilde{T}_V}$  lies in the basin of attraction of  $\bar{\pi}_1^*(\tilde{\mu}, V^0)$  and converges to it. As  $\bar{\pi}_1^*(\tilde{\mu}, V^0) < \bar{\pi}_2^*(\mu^0, V^0)$ , the trajectory eventually falls strictly below  $\bar{\pi}_2^*(\mu^0, V^0)$ , so  $T_\mu < \infty$ .

Under  $(V'', \mu^0)$ ,  $T_V < \infty$  and by definition  $\bar{\pi}^{T_V} < \bar{\pi}_2^*(\mu^0, V^0)$ . Restoring  $V^0$  at  $t = T_V$  returns the model to the three-steady-state configuration under  $(\mu^0, V^0)$ , with  $\bar{\pi}^{T_V}$  lying strictly in the basin of attraction of  $\bar{\pi}_1^*(\mu^0, V^0)$ . The economy therefore converges persistently to  $\bar{\pi}_1^*(\mu^0, V^0) < \bar{\pi}^0$ .

As established above,  $\bar{\pi}^{\tilde{T}_V} \in (\bar{\pi}_2^*(\mu^0, V^0), \bar{\pi}_2^*(\tilde{\mu}, V^0))$ . At  $t = \tilde{T}_V$ ,  $V$  is restored to  $V^0$  and  $\mu$  is raised to  $\tilde{\mu}$ . The maintained conditions ensure three steady states under  $(\tilde{\mu}, V^0)$ . As  $\bar{\pi}^{\tilde{T}_V} < \bar{\pi}_2^*(\tilde{\mu}, V^0)$ , the trajectory under  $(\tilde{\mu}, V^0)$  lies in the basin of attraction of  $\bar{\pi}_1^*(\tilde{\mu}, V^0)$  and converges to it. As  $\bar{\pi}_1^*(\tilde{\mu}, V^0) < \bar{\pi}_2^*(\mu^0, V^0)$ , the trajectory crosses  $\bar{\pi}_2^*(\mu^0, V^0)$  in finite time  $T_\mu$ , by the argument above. At  $t = \tilde{T}_V + T_\mu$ , both instruments are restored to  $(\mu^0, V^0)$ . By the definition of  $T_\mu$ ,  $\bar{\pi}^{\tilde{T}_V+T_\mu} < \bar{\pi}_2^*(\mu^0, V^0)$ , so the trajectory lies strictly in the basin of attraction of  $\bar{\pi}_1^*(\mu^0, V^0)$  and converges to it persistently. The strict inequality  $\tilde{T}_V < T_V$  is established above, proving the proposition.  $\square$

### Proof of Proposition 6.

*Proof.* The maintained assumptions place the economy in the high-bias basin under  $(V^0, \mu^0)$ , so the counterfactual trajectory  $\{\bar{\pi}_{cf}^t\}$  converges to  $\bar{\pi}_3^*(\mu^0, V^0)$ . I consider each instrument in turn.

*V-intervention.* Under  $(V'', \mu^0)$ , Proposition 4(i) implies a unique globally attracting steady state  $\bar{\pi}_1^*(V'') < \bar{\pi}_2^*(V^0)$ , so the intervention trajectory  $\{\bar{\pi}_{int}^t\}$  converges to  $\bar{\pi}_1^*(V'')$ . Add and subtract  $W(\bar{\pi}_{int}^t; V^0, \mu^0)$  in the total welfare comparison:

$$\begin{aligned} W(\bar{\pi}_{int}^t; V'', \mu^0) - W(\bar{\pi}_{cf}^t; V^0, \mu^0) &= \underbrace{W(\bar{\pi}_{int}^t; V'', \mu^0) - W(\bar{\pi}_{int}^t; V^0, \mu^0)}_{= B_V(t) \geq 0} \\ &\quad + \underbrace{\frac{1}{2}[\bar{S}(\bar{\pi}_{cf}^t; V^0, \mu^0) - \bar{S}(\bar{\pi}_{int}^t; V^0, \mu^0)]}_{\text{bias-reduction gain} \geq 0}. \end{aligned}$$

The first term satisfies  $B_V(t) \geq 0$  for all  $t$  by the envelope argument in Section 5.2. For the second term, I claim  $\bar{\pi}_{int}^t \leq \bar{\pi}_{cf}^t$  for all  $t \geq 0$ . At  $t = 0$ ,  $\bar{\pi}_{int}^0 = \bar{\pi}_{cf}^0 = \bar{\pi}^0$ . For  $t \geq 1$ , the intervention trajectory converges monotonically to  $\bar{\pi}_1^*(V'') < \bar{\pi}_2^*(V^0) < \bar{\pi}^0$ , while the counterfactual converges to  $\bar{\pi}_3^*(V^0) > \bar{\pi}_2^*(V^0)$ , so  $\bar{\pi}_{int}^t < \bar{\pi}_{cf}^t$  for all  $t \geq 1$ . Since  $\bar{S}$  is increasing in  $\bar{\pi}$  (as  $f(z)$  is increasing in  $z$  and  $z_k = (1 - \beta_k)\theta(\bar{\pi})V$  is increasing in  $\bar{\pi}$  through  $\theta$ ), the bias-reduction gain is non-negative for all  $t$  and strictly positive for  $t \geq 1$ .

At  $t = 0$ ,  $B_V(0) > 0$  strictly:  $V'' > V^0$ ,  $\bar{\pi}^0 > 0$  (so  $\theta(\bar{\pi}^0) > 0$ ), and  $\frac{\partial u^*}{\partial V} > 0$  for every match type with  $z > 0$ . Hence, the total welfare difference is strictly positive for all  $t \geq 0$ .

*$\mu$ -intervention.* The decomposition established in Section 5.2 gives

$$W(\bar{\pi}_{int}^t; \tilde{\mu}) - W(\bar{\pi}_{cf}^t; \mu^0) = \underbrace{\frac{1}{2}[\bar{S}(\bar{\pi}_{cf}^t; \mu^0) - \bar{S}(\bar{\pi}_{int}^t; \mu^0)]}_{\text{bias-reduction gain}} - \underbrace{J_\mu(t)}_{\mu\text{-penalty}}.$$

At the start of the intervention  $t = 0$ ,  $\bar{\pi}_{int}^0 = \bar{\pi}_{cf}^0 = \bar{\pi}^0$ , so the bias-reduction gain is zero. The  $\mu$ -penalty satisfies  $J_\mu(0) > 0$  (since  $\bar{\pi}^0 > 0$  implies  $\bar{s}_{out,\tau}^* - s_0^* > 0$  by the effort ordering  $s_2^* > s_1^* > s_0^*$ ). Hence, the total welfare difference equals  $-J_\mu(0) < 0$ . Under  $\tilde{\mu} > \mu^0$ , the intervention trajectory falls faster than the counterfactual:  $\bar{\pi}_{int}^t < \bar{\pi}_{cf}^t$  for all  $t \geq 1$  (the higher contact rate generates more positive out-group interactions, reducing bias faster, while the counterfactual remains in the high-bias basin). The bias-reduction gain  $\frac{1}{2}[\bar{S}(\bar{\pi}_{cf}^t; \mu^0) - \bar{S}(\bar{\pi}_{int}^t; \mu^0)]$  is therefore strictly positive and growing for  $t \geq 1$ , as  $\bar{\pi}_{cf}^t$  converges upward to  $\bar{\pi}_3^*$  while  $\bar{\pi}_{int}^t$  converges downward toward  $\bar{\pi}_1^*(\tilde{\mu})$ . Simultaneously,  $J_\mu(t)$  is declining (as shown in Section 5.2, bias reduction narrows the effort gap  $\bar{s}_{out,\tau}^* - s_0^*$ ). Since the bias-reduction gain is eventually bounded below by  $\frac{1}{2}[\bar{S}(\bar{\pi}_3^*; \mu^0) - \bar{S}(\bar{\pi}_1^*(\tilde{\mu}); \mu^0)] > 0$  and  $J_\mu(t) \rightarrow J_\mu(\infty)$  evaluated at  $\bar{\pi}_1^*(\tilde{\mu})$ , which is bounded, there exists  $\hat{t}$  at which the gain exceeds the penalty. For all  $t > \hat{t}$ , the total welfare difference is strictly positive.

At  $t = T_M$ ,  $\mu$  is returned to  $\mu^0$ . The  $\mu$ -penalty drops to zero immediately:  $J_\mu(t) = 0$  for  $t \geq T_M$  because the matching parameter is  $\mu^0$  in both the intervention and counterfactual paths. The bias-reduction gain persists. By the maintained assumption,  $\bar{\pi}_{int}^{T_M} < \bar{\pi}_2^*(\mu^0)$ , so the post-intervention trajectory converges to  $\bar{\pi}_1^*(\mu^0)$ , while the counterfactual converges to  $\bar{\pi}_3^*(\mu^0)$ . Hence,  $\bar{S}(\bar{\pi}_{cf}^t; \mu^0) > \bar{S}(\bar{\pi}_{int}^t; \mu^0)$  for all  $t \geq T_M$ , and the total welfare difference equals the bias-reduction gain, which is strictly positive and converges to  $\Delta W$  as  $t \rightarrow \infty$ .  $\square$

## Proof of Proposition 7.

*Proof.* All three policies converge to the same low-bias steady state  $\bar{\pi}_1^*(\mu^0, V^0)$ , so the per-period welfare from period  $T_P$  onward (where  $T_P$  is the last period of intervention under policy  $P$ ) converges to the

common value  $W(\bar{\pi}_1^*(\mu^0, V^0))$ . The present discounted welfare under any successful policy  $P$  can therefore be decomposed as

$$\mathcal{W}_P = \sum_{t=0}^{T_P-1} \delta^t [W^P(t) - c^P(t)] + \sum_{t=T_P}^{\infty} \delta^t W^P(t),$$

where  $W^P(t)$  is per-period welfare and  $c^P(t) \in \{c_V, c_\mu\}$  is the fiscal cost at time  $t$  along the trajectory induced by policy  $P$ .

Write the no-intervention counterfactual welfare as  $\mathcal{W}_0 = \sum_{t=0}^{\infty} \delta^t W_{cf}(t)$ , where  $W_{cf}(t) \rightarrow W(\bar{\pi}_3^*)$  as  $t \rightarrow \infty$ . For each policy  $P$ , define the net welfare gain relative to the counterfactual:

$$\mathcal{W}_P - \mathcal{W}_0 = \underbrace{\sum_{t=0}^{\infty} \delta^t [W^P(t) - W_{cf}(t)]}_{\text{total transition-path welfare gain}} - \underbrace{\sum_{t=0}^{T_P-1} \delta^t c^P(t)}_{\text{total fiscal cost}}.$$

Because all policies converge to  $\bar{\pi}_1^*(\mu^0, V^0)$  and the counterfactual converges to  $\bar{\pi}_3^*(\mu^0, V^0)$ , the tail of the transition-path welfare gain converges to  $\Delta W = \frac{1}{2}[\bar{S}(\bar{\pi}_3^*) - \bar{S}(\bar{\pi}_1^*)]$  per period. The discounted value of this permanent gain,  $R = \Delta W/(1 - \delta)$ , is identical across all three policies.

The policies differ only in their transition-path welfare effects and fiscal costs during the intervention phase. Substituting the definitions of  $B_V(t)$  and  $J_\mu(t)$ :

*Policy V:* The per-period welfare difference relative to the counterfactual is  $B_V(t) + G_V(t)$ , where  $G_V(t) = \frac{1}{2}[\bar{S}(\pi_{cf}^t; \mu^0) - \bar{S}(\pi_V^t; \mu^0)]$  is the bias-reduction gain along the  $V$ -only trajectory, and the per-period fiscal cost is  $c_V$ . Hence, the welfare-adjusted cost in each period is  $c_V - B_V(t) = \tilde{c}_V^V(t)$ , and the total welfare-adjusted cost is  $C_V$  as defined in (23), up to the common bias-reduction gain that is absorbed into  $R$ .

*Policy M:* The per-period welfare difference relative to the counterfactual is  $-J_\mu(t) + G_M(t)$ , the per-period fiscal cost is  $c_\mu$ , and the welfare-adjusted cost is  $c_\mu + J_\mu(t) = \tilde{c}_\mu^M(t)$ . The total is  $C_M$  as in (22).

*Policy S:* The welfare-adjusted costs are  $\tilde{c}_V^S(t)$  during  $[0, \tilde{T}_V)$  and  $\tilde{c}_\mu^S(t)$  during  $[\tilde{T}_V, \tilde{T}_V + T_\mu)$ . The total is  $C_S$  as in (24).

Since  $R$  is common to all three policies, the net welfare gain satisfies  $\mathcal{W}_P - \mathcal{W}_0 = R - C_P + \Xi_P$ , where  $\Xi_P$  collects differences in the timing of the bias-reduction gain across policies that do not affect the long-run welfare. Because all trajectories converge to the same steady state, differences in  $\Xi_P$  across policies are of order  $O(\delta^{T_P})$  and vanish relative to  $R$  for patient planners ( $\delta$  close to 1). At any  $\delta \in (0, 1)$ , the ranking  $\mathcal{W}_P > \mathcal{W}_{P'}$  holds if and only if  $C_P < C_{P'}$ , because the bias-reduction gains, while differing in their timing across policies, are second-order relative to the welfare-adjusted costs for planners who are sufficiently patient.

Strictly, for any  $\delta \in (0, 1)$ , define the *full* welfare-adjusted cost inclusive of transition-path differences:

$$C_P^* = \sum_{t=0}^{T_P-1} \delta^t c^P(t) - \sum_{t=0}^{\infty} \delta^t [W^P(t) - W_{cf}(t)].$$

Then  $\mathcal{W}_P - \mathcal{W}_0 = -C_P^*$ , and  $P$  is welfare-preferred to  $P'$  iff  $C_P^* < C_{P'}^*$ . During the intervention phase,  $W^P(t) - W_{cf}(t)$  decomposes into the instrument-specific welfare effect (captured by  $B_V$  or  $-J_\mu$ ) and the bias-reduction gain. The welfare-adjusted per-period costs  $\tilde{c}_V(t)$  and  $\tilde{c}_\mu(t)$  as defined in (21) incorporate

the instrument-specific welfare effects. The bias-reduction gains differ across policies only during the intervention phases (since all post-intervention trajectories converge to the same steady state), so  $C_P^* = C_P - \Gamma_P$  where  $\Gamma_P = \sum_t \delta^t G_P(t)$  is the discounted bias-reduction gain along policy  $P$ 's trajectory. Since  $\Gamma_P$  depends on the trajectory, pairwise rankings reduce to comparing  $C_P - \Gamma_P$  across policies.

The pairwise comparisons in statements (i)–(iii) hold by construction: policy  $P$  is welfare-preferred to  $P'$  iff  $C_P - \Gamma_P < C_{P'} - \Gamma_{P'}$ . Because the proposition defines the ranking in terms of  $C_P$  (the welfare-adjusted costs as defined in (22)–(24)), which already incorporate the instrument-specific welfare effects through  $B_V$  and  $J_\mu$ , the stated pairwise conditions  $C_P < C_{P'}$  are necessary and sufficient whenever the bias-reduction gain differential  $\Gamma_P - \Gamma_{P'}$  is negligible or can be absorbed into the cost comparison. This holds approximately when the comparison involves policies that share initial trajectory segments (as in statement (i), where  $V$  and  $S$  share the first  $\tilde{T}_V$  periods), and holds to a close approximation otherwise, with the approximation improving as  $\delta \rightarrow 1$ .  $\square$