The Polarisation Threshold

Gabriel Bontemps

February 27, 2026

Common information is often expected to discipline beliefs toward consensus. Yet in many real-world episodes of mass communication, shared signals appear to *sharpen* disagreement: identical public announcements can be followed by increasingly divergent risk assessments, compliance choices, or political interpretations. This paper provides a parsimonious continuous-time mechanism for this "common-information polarisation" that does *not* rely on network segregation, selective exposure, heterogeneous priors, or bounded rationality. The key behavioural ingredient is instead a standard implication of cognitive-dissonance adjustment with fixed costs: when belief revision is costly and discrete, the cross-sectional belief distribution acquires *endogenous reinjection flows* that can interact with even the simplest form of social feedback to destabilise consensus.

At the individual level, we model a belief deviation u_t from an anchor (reference point) as an Ornstein–Uhlenbeck diffusion,

$$\mathrm{d}u_t = -\lambda u_t \,\mathrm{d}t + \sigma \,\mathrm{d}W_t,$$

and assume that holding beliefs far from the anchor generates convex dissonance costs (quadratic in u). Revising one's belief is not frictionless: agents may pay a fixed cognitive adjustment cost K_0 to perform a discrete *reset* that partially shrinks the deviation, $u \mapsto c_0 u$ with $c_0 \in (0,1)$. This yields an impulse-control problem with a quasi-variational inequality. Under the usual smooth-fit regularity conditions, the optimal policy takes a *band-and-reset* form: agents tolerate deviations within an endogenous interval $(-\Theta, \Theta)$, and when the process hits $\pm\Theta$ they optimally intervene, paying the fixed cost and reinjecting the belief to $\pm c_0\Theta$. Thus, microfounded cognitive adjustment implies a belief process with boundary exits and reinjection at interior points.

We then embed these microfounded resets in a population model with the simplest possible social interaction: *mean-only social feedback*. In an N-agent system, each belief evolves within the band according to

$$\mathrm{d}u_t^i = \big(-\lambda u_t^i + \kappa \bar{u}_t^N\big)\,\mathrm{d}t + \sigma \,\mathrm{d}W_t^i, \qquad \bar{u}_t^N = \frac{1}{N}\sum_{j=1}^{N} u_t^j,$$

with the same boundary-reset rule at $|u_t^i| = \Theta$. Under standard propagation-of-chaos conditions, the empirical measure converges to a deterministic density ρ_t

solving a McKean–Vlasov/Fokker–Planck equation on $[-\Theta, \Theta]$ with (i) Dirichlet "exit" boundary conditions at $\pm\Theta$ and (ii) *transmission (flux-jump) conditions* that encode reinjection at $\pm c_0\Theta$. The resulting dynamics are not a smooth diffusion on R: they are a diffusion with endogenous, state-dependent probability currents created by the reset policy.

A first result is that the system admits a *consensus stationary equilibrium* for every feedback strength $\kappa \geq 0$: the unique even stationary density of the underlying OU-with-boundary-resets process, denoted $\bar{\rho}^{\mathrm{BR}}$, together with mean zero, is always a stationary solution of the mean-field equation. The substantive question is therefore not existence but *stability*: when is this symmetric equilibrium locally attractive, and when does it become fragile to polarising perturbations?

The main contribution is an explicit and tractable stability theory. Linearising the mean-field dynamics around $\bar{\rho}^{\mathrm{BR}}$ yields a linear operator that decomposes into a baseline generator (the OU–reset operator at $\kappa = 0$) plus a *rank-one perturbation* induced by mean-only coupling. Restricting attention to the *odd* subspace (polarising deviations), the infinite-dimensional spectral problem collapses to a scalar characteristic equation

$$1 = \kappa\, \Phi(z),$$

where $\Phi(z)$ is a resolvent-based *gain map* (a susceptibility of the distribution to a mean shift). In particular, at $z = 0$ the model delivers a closed-form tipping condition

$$\kappa^\star = \frac{1}{\Phi(0)}.$$

We prove that there exists a unique analytic leading eigen-branch crossing the origin at $\kappa^\star$ with strictly positive slope (transversality). Hence, the consensus equilibrium is locally asymptotically stable to polarising (odd) perturbations for $\kappa < \kappa^\star$ and loses odd-subspace stability at $\kappa = \kappa^\star$. Intuitively, $\Phi(0)$ is an aggregate polarisation elasticity: it accumulates over time the mechanical response of the stationary density to small mean shifts, taking into account both OU stabilisation and the endogenous reinjection flow created by resets. The product $\kappa\Phi(0)$ therefore acts as a *social multiplier*: consensus is stable when the feedback loop is weaker than unity and becomes fragile as $\kappa\Phi(0) \uparrow 1$.

Because $\Phi(0)$ is pinned down by primitives of the OU–reset process, the framework offers transparent comparative statics and a computation workflow that does not require simulation-heavy black boxes. In the paper, $\Phi(0)$ is obtained from a boundary/transmission value problem tied to the stationary density $\bar{\rho}^{\mathrm{BR}}$, enabling direct numerical evaluation and sensitivity analysis with respect to $(\lambda, \sigma, \Theta, c_0)$. Numerically, once κ passes $\kappa^\star$, the model exhibits the emergence of a bimodal cross-sectional density, consistent with a symmetry-breaking "polarisation" pattern.

Finally, the theory is designed to be empirically operational. The parameters (λ, σ) can be estimated from within-band belief increments (conditioning on no reset), the reset primitives (Θ, c_0) from detected revision episodes (pre-jump magnitudes and jump ratios), and the feedback strength κ from how

within-band drift responds to the contemporaneous mean (with experimental or quasi-experimental variation in the salience/visibility of the mean providing a natural instrument). Given these estimates, one computes $\Phi(0)$ and the implied threshold $\kappa^\star$, and can track an early-warning distance-to-instability statistic $\Delta(\kappa) = 1 - \kappa\Phi(0)$. The policy implication is correspondingly sharp: interventions can target either (i) the strength of norm pressure κ (e.g., amplification of "what others think" cues) or (ii) susceptibility $\Phi(0)$ through levers that affect cognitive adjustment costs and reset effectiveness. Overall, the paper provides a minimal microfounded route from cognitive dissonance with fixed costs to a computable macroscopic tipping point at which common information ceases to unify and instead fuels persistent disagreement.