PNAS RESEARCH ARTICLE | SOCIAL SCIENCES

# The cultural evolution of pluralistic ignorance

Sergey Gavrilets[a,b,c,d,1,2], Johannes Karl[e,f,g,1], and Michele J. Gelfand[g,1]

Edited by Susan Fiske, Princeton University, Jamaica, VT; received August 23, 2025; accepted December 27, 2025

**Pluralistic ignorance—the systematic misperception of others' attitudes—can entrench suboptimal norms, yet its dynamics remain poorly understood. We develop a mathematical model of the coevolution of actions, private attitudes, and beliefs about others, with societal tightness as a central parameter. Our framework integrates theories of spirals of silence, preference falsification, and cultural mismatch into a single dynamic system capturing the effects of material payoffs, cognitive forces, and social influence. The model shows that pluralistic ignorance can arise from lags between attitude change and belief updating, even without silence or deception. Dynamics unfold faster in loose cultures and slower in tight ones: loose societies display sharp but transient peaks of pluralistic ignorance, while tight societies sustain slower, persistent mismatches. Both can experience cultural evolutionary mismatch but through distinct pathways—internalized norm adherence in loose cultures vs. conformity pressure in tight ones. These mechanisms may help explain global patterns where private support exceeds perceived support, such as climate action, women's rights, and abortion attitudes. Interventions must therefore be culturally tailored: accelerating attitude change through highlighting benefits is effective in loose cultures, whereas lowering expression costs (via anonymity or legal protections) empowers norm entrepreneurs in tight cultures. Our framework identifies policy levers and clarifies when apparent opinion stability conceals underlying shifts, offering insights for democratic societies navigating rapid social change.**

evolution of beliefs about others | cultural evolutionary mismatch | spirals of silence | tight and loose cultures | preference falsification

## Significance

People often get public opinion wrong, assuming their own views are unpopular when in fact many others share them. This widespread misperception, called pluralistic ignorance, can trap societies in harmful or outdated norms. We build a mathematical model showing how these misperceptions form and change over time, depending on whether cultures are "tight" (with strict norms) or "loose" (with flexible ones). Our results explain why support for issues like climate action or women's rights is often underestimated, and why change happens faster in some societies than others. The model also points to practical solutions: in loose cultures, sharing accurate information works best, while in tight ones, lowering the costs of speaking up can spark social change.

Human societies systematically misperceive their own collective preferences. In every aspect of public life—from climate change and women's rights to political polarization—people vastly misestimate how many others share their views. For example, a study across 125 nations showed that while 69% would donate income to fight climate change, people believed that support was much lower (1). The same pattern appears for women's rights across 60 countries, where solid majorities support basic rights but assume others do not (2). For affirmative action, the bias reverses: in nations where a majority approves it, approval is underestimated, while in nations where it is a minority view, it is overestimated (2). Similarly, Americans significantly overestimate how likely others—especially those from the opposing political party—are to engage in canceling behavior (3). Even on divisive issues like abortion, Americans drastically underestimate how much support exists for access, with both sides viewing their own positions as more extreme than they actually are (4).

This phenomenon, known as pluralistic ignorance (PI) (5–10), has profound implications for social and political processes. At the individual level, PI generates psychological distress and isolation, as citizens conceal their authentic views. Collectively, it suppresses dissent, drains informational diversity, and can lock groups into outdated or harmful norms, producing suboptimal, or even disastrous, decisions. Legislators may misread the electorate, passing laws that most citizens secretly oppose, while rival factions can exaggerate each other's extremism, fueling polarization. And when hidden majorities finally recognize their numbers, opinion can flip abruptly and spark mass mobilization.

Pluralistic ignorance emerges through several distinct but interacting mechanisms including spirals of silence, preference falsification, misinterpretation, biased sampling, and structural distortion, though these mechanisms tend to be siloed in different disciplines and poorly integrated. Spirals of silence (11) occur when individuals fear social sanctions and choose to remain silent. For example, Republican supporters of childhood vaccines—despite being the majority—anticipate social conflict when exposed to an online environment dominated by an antivaccine minority within their party, leading them to reduce their participation in online discussions (12). Preference falsification (13) arises when people voice views they do not hold, as when college students profess enthusiasm for heavy drinking to appear socially aligned. Even honest signals are easily

Author affiliations: [a]Department of Ecology and Evolutionary Biology, University of Tennessee, Knoxville, TN 37996; [b]Department of Mathematics, University of Tennessee, Knoxville, TN 37996; [c]Complexity Science Hub, Vienna 1030, Austria; [d]Institute for Advanced Study Toulouse, Toulouse School of Economics, Toulouse 31000, France; [e]School of Psychology, Department of Psychology, Te Herenga Waka, Victoria University of Wellington, Wellington 6140, New Zealand; [f]Department of Psychology, University of Zurich, Zurich CH-8006, Switzerland; and [g]Department of Psychology, Stanford Graduate School of Business, Stanford University, Stanford, CA 94305

Author contributions: S.G. and M.J.G. designed research; S.G. performed research; S.G., J.K., and M.J.G. analyzed data; and S.G., J.K., and M.J.G. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.



[1]S.G., J.K., and M.J.G. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: gavrila@utk.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2522998123/-/DCSupplemental.

Published February 13, 2026.

misinterpreted. Observers may undertrust them, assuming that visible behavior hides opposite motives, or overtrust them, e.g., when treating a casual social-media "like" as wholehearted endorsement. Misinterpretation can also spring from projection, when people simply assume others think as they do. Biased sampling misleads observers who rely on unrepresentative networks such as when a voter encircled by partisan friends may infer the entire electorate shares that stance. Structural distortion comes from gatekeepers or algorithms that magnify some opinions and muffle others, such as news outlets that saturate coverage with antivaccine protests that attract only a vocal minority. All these mechanisms illustrate how humans' pronounced sensitivity to social influence can cause them to align not only with what others do, but also with what we think they approve or believe (14–17).

Understanding PI dynamics is an urgent societal concern. In an era of rapid social change, political polarization, and global challenges requiring collective action, the ability to accurately gauge public sentiment is critical for the success of everything from policy implementation to social movements to public health campaigns. Yet despite its prevalence and importance, our understanding of the evolution of PI remains very limited. Research has remained fragmented, with little attention to how the above mechanisms interact to predict PI over time. Mathematical modeling holds great promise to better understand and quantify these interactions, making it possible to predict how misperceptions influence collective behavior and societal outcomes. Nevertheless, formal models of pluralistic ignorance remain sparse and are narrow in scope.

For example, Taylor's pioneering study (18) treated the problem as a one-person game, while Granovetter and Soong (19) recast the threshold framework of Granovetter (20). Similar approaches were used by Kuran (13) and Centola et al. (21) to model preference falsification. Bénabou and Tirole (22) showed how pluralistic ignorance can arise from individuals' inferences about societal values based on observed laws and others' behaviors, emphasizing the role of expressive law in shaping perceived norms. In all these studies, expressed opinions evolve over time but private attitudes remain fixed. Fernández-Duque (23) retained that assumption, yet added sequential moves and explicit group-size effects, again in a game-theoretic setting. More recent work couples private attitudes and public expression in DeGroot-style opinion dynamics (24), though it still forces agents to speak in every round (25, 26). Agent-based simulations of online behavior do allow attitude change, optional silence, and evolving expression, but focus almost exclusively on how network topology shapes spirals of silence, leaving other drivers largely unexplored (19, 27–31).

Valuable as they are, these early models leave out several key ingredients. Second-order beliefs are never modeled directly; they are simply equated with the visible average behavior of one's partners. Conformity is the sole cognitive motive considered, while forces such as cognitive dissonance, social projection, and theory-of-mind reasoning are ignored. Most dynamic treatments also drop material pay-offs, blocking any analysis of cultural–evolutionary mismatch, i.e. cases in which once-adaptive norms (e.g., child marriage) persist after conditions change (32–34). Finally, these models assume cultural homogeneity. No models have examined how cultural factors—such as the strength of social norms and tolerance for deviant behavior(35) can dramatically affect PI dynamics. In reality, the costs of dissent and pressures to conform vary dramatically across cultures, institutional settings, and historical contexts. These factors must be explicitly modeled to accurately capture the emergence and persistence of pluralistic ignorance globally.

To fill these gaps, we develop a formal model of PI that builds upon a recently developed framework that merges norm-utility theory with belief dynamics (17, 36–40) that has been validated in behavioral experiments (41, 42).

In our model, we track the coevolution of each individual's action, private attitude, and second-order beliefs about others' actions and attitudes. Individuals choose actions that maximize a utility combining material payoffs, personal normative inclinations, and perceived social approval. Personal norms and perceived approval update endogenously through social learning, cognitive-dissonance reduction, and social projection. We also incorporate norm strength as an explicit parameter, allowing us to capture how pluralistic ignorance differs between tight cultures (where norms are rigid and dissent costly) and loose cultures (where norms are flexible and dissent tolerated). Our model endogenously produces spirals of silence, preference falsification, and misinterpretations of others' behavior, while deliberately leaving biased sampling and structural information distortion outside its scope.

We compare short- and long-run trajectories of actions, beliefs, and PI, show how PI can emerge from temporal lags between attitude change and belief updating, and identify conditions that favor cultural evolutionary mismatch. The analysis explains several observed empirical patterns and yields testable predictions. By illuminating the hidden multilevel and dynamic architecture of collective misbelief, this model brings together decades of fragmented theory under one roof. It offers a generalizable framework for understanding social change, norm entrenchment, and cultural evolutionary mismatch-and opens paths for policy, intervention, and cultural foresight.

The next section lays out the model: we introduce its core elements, define the utility function, specify the dynamics for private attitudes and second-order beliefs, and show how cultural tightness enters as an explicit parameter. We then present and interpret the analytical results and numerical simulations. The final section distills the main insights and connects them to key empirical patterns.

## 1. Results

**1.1. Model.** Consider a population of individuals engaged in social interactions involving two competing options, which may represent formal policies or informal social norms. For simplicity, we refer to these options as the old norm and the new norm. Each individual has a personal attitude $y$ toward these norms, that is, their internal belief about what behavior ought to be performed. Below, attitude and personal norm are used interchangeably, consistent with earlier formal models of norm internalization. Attitude $y$ ranges between $[-1, 1]$. A positive $y$ indicates a preference for the new norm, while a negative $y$ reflects a preference for the old norm. The absolute value of $y$, denoted $|y|$, represents the strength of the individual's support for the corresponding norm. The distribution of attitudes in the population, $f(y)$, is unknown to individuals. Instead, each individual forms a second-order belief about the average attitude in the population, denoted $\tilde{y}$. This belief $\tilde{y}$ captures injunctive norms or normative expectations, reflecting what individuals believe others think ought to be done (43).

Time is discrete. At each time step, every individual sequentially revises their action, attitude, and second-order belief. Individual actions are specified by variable $x$. Individuals may i)

act in line with their preference—choose $x = 1$ when $y > 0$ or $x = -1$ when $y < 0$; ii) abstain, $x = 0$; or iii) support the norm they privately oppose—choose $x = -1$ when $y > 0$ or $x = 1$ when $y < 0$. Acting in support of either norm incurs a cost $c$. Choosing action $x = 1$ yields a net benefit of $b \geq 0$, while choosing action $x = -1$ results in a net loss of $-b$. The actions $x = -1$ and $x = 1$ can be interpreted either as publicly voicing support for the corresponding norm or as adopting behaviors aligned with one of the norms.

For example, political endorsements often provide financial or social rewards, whereas detractors risk social exclusion or professional setbacks. Similarly, aligning with prevailing trends on social media can enhance one's influence, whereas opposing such trends might result in follower loss or reduced visibility. Taking controversial positions may lead to backlash, yet these actions can also attract dedicated niche followings. Public actions demonstrating support or opposition to a government can similarly result in tangible gains or losses. Under conditions of religious suppression, individuals might openly practice a traditional religion, conform publicly to a state-imposed religion, or privately maintain their beliefs without outward expression. In rapidly evolving industries, professionals can choose to uphold traditional practices, embrace emerging technologies, or withdraw from participation altogether. Similarly, dietary behaviors illustrate comparable choices: people may continue to consume animal products, adopt a strictly plant-based diet, or opt for a flexible, intermediate (flexitarian) approach.

Let $p$, $q$, and $1 - p - q$ be the frequencies of people choosing $x = 1, -1$, and 0, respectively. The average value of the expressed behaviors is $\bar{x} = \frac{p-q}{p+q}$ which is assumed to be known from observations.

**1.2. Utility Function and Best Response.** Assume that when deciding on an action, each individual aims to maximize their utility, expressed as follows:

$$u = \underbrace{bx}_{\text{action payoff}} - \underbrace{c|x|}_{\text{action cost}} + \underbrace{k_1 yx}_{\text{cognitive dissonance}} + \underbrace{k_2 \tilde{y} x}_{\text{social influence}}. \quad [1]$$

This equation captures four key considerations influencing individual choice. The first two terms represent material or social incentives and costs associated with publicly supporting a norm. The next two terms introduce psychological and social dimensions: the cognitive dissonance term captures the discomfort or satisfaction resulting from the misalignment or alignment of an individual's expressed action ($x$) with their private attitude ($y$); the social influence term reflects the social pressure or encouragement stemming from the perceived average attitude of others. Notice that the signs of these two psychological terms depend explicitly on whether the chosen action $x$ aligns or conflicts with an individual's private attitude $y$ and their perception of the group's prevailing sentiment, $\tilde{y}$. Choosing to abstain from expressing any preference ($x = 0$) results in a baseline utility, defined as $u = 0$.

The nonnegative parameters $b$, $c$, $k_1$, and $k_2$ quantify the relative strengths of these competing influences and may vary according to individual cognitive, psychological, or cultural characteristics.

In *SI Appendix*, we provide equations that describe the best response actions. These equations show that as the benefit $b$ increases, more individuals are inclined to choose action $x = 1$. Conversely, as the cost $c$ increases, more individuals will prefer to remain silent. An increase in cognitive dissonance $k_1$ encourages more individuals to express their opinions, while a stronger conformity effect $k_2$ raises the likelihood that individuals choose actions aligning with the perceived majority opinion. These results are intuitive.

**1.3. Dynamics of Attitudes and Second-Order Beliefs.** After taking an action and observing the behavior of others, personal attitudes change according to the recurrence equation:

$$y' = y + \underbrace{\alpha_1(x - y)}_{\text{cognitive dissonance}} + \underbrace{\beta_1|\bar{x}|(\bar{x} - y)}_{\text{social influence}} + \underbrace{\gamma_1 b.}_{\text{benefit-driven attitude adjustment}}. \quad [2]$$

Here, cognitive dissonance shifts attitude $y$ toward the action taken ($x$), while social influence pulls $y$ toward the observed average behavior $\bar{x}$. If both norms are equally frequent (so that $p = q$ and $\bar{x} = 0$), the social influence term vanishes. The last term captures how the benefit $b$ directly modifies attitude $y$. Parameters $\alpha_1, \beta_1, \gamma_1$ measure sensitivity to each corresponding force.

We further postulate that, after observing peer behavior, second-order beliefs $\tilde{y}$ shift according to the recurrence equation:

$$\tilde{y}' = \tilde{y} + \underbrace{\alpha_2(y - \tilde{y})}_{\text{social projection}} + \underbrace{\beta_2(\bar{x} - \tilde{y})}_{\text{learning}} + \underbrace{\gamma_2 b.}_{\text{benefit-driven belief adjustment}} \quad [3]$$

Social projection (44) shifts $\tilde{y}$ toward the individual's own attitude, while learning acts to align $\tilde{y}$ with observed average behavior $\bar{x}$. The last term reflects how the benefit $b$ modifies second-order beliefs. Parameters $\alpha_2, \beta_2, \gamma_2$ control the sensitivity of beliefs to these forces.

Notice that the effects of observations on $y$ and $\tilde{y}$ are modeled via the mean observed behavior $\bar{x}$. In binary behavior settings, the mean also defines variance: $\text{var}\, x = \bar{x}(1 - \bar{x})$.

**1.4. Psychological Characteristics of Tight and Loose Cultures.** In the model, cognitive effects on actions and beliefs are quantified by parameters $k_1, \alpha_1$, and $\alpha_2$; social influence is measured by $k_2, \beta_1$, and $\beta_2$; and motivational factors—driven by the desire to align attitudes and beliefs with the beneficial action—are represented by $\gamma_1$ and $\gamma_2$.

In tight cultures, behavior is primarily guided by beliefs about socially appropriate actions while in loose cultures behavior is guided by internal values (45, 46). Accordingly, individuals in tight cultures are theorized to have less adjustment to internal cognitive dissonance, reflected in smaller $k_1, \alpha_1, \alpha_2$, and have larger $k_2$. Observed behaviors are theorized to be less influential when updating personal attitudes and second-order beliefs, corresponding to smaller $\beta_1, \beta_2$, as such behaviors are less reliable indicators of true preferences in tight cultures.

In contrast, individuals in loose cultures are theorized to be more sensitive to internal cognitive factors like cognitive dissonance (47–49) and social projection (50, 51), corresponding to larger $k_1, \alpha_1, \alpha_2$. However, conformity and second-order beliefs play a smaller role, leading to smaller $k_2$. Individuals are strongly influenced by observed behaviors when updating beliefs, resulting in larger $\beta_1, \beta_2$. This occurs because such behaviors are viewed as reliable indicators of attitudes (52, 53).

Tight cultures resist small psychological or material benefits that challenge norms, due to strong adherence to tradition and high costs of deviation. Loose cultures, by contrast, are more open to integrating such benefits, driven by tolerance for diversity and change (54, 55). However, in loose cultures, second-order

beliefs shift less strongly or uniformly than personal attitudes due to weaker conformity pressures. These dynamics imply smaller $\gamma_i$ values in tight cultures, reflecting reduced benefit-driven adjustments. In loose cultures, $\gamma_1 > \gamma_2$, indicating that personal attitudes $y$ adjust more readily than second-order beliefs $\tilde{y}$ to benefits.

In numerical simulations, we capture these patterns by introducing a parameter $\tau$ representing cultural tightness ($0 \leq \tau \leq 1$) and assuming that the mean values of the parameters governing decision-making and belief update depend on $\tau$ in specific ways (explicitly defined in *SI Appendix*) that reflect the effects described above. These functional forms are theoretically motivated but necessarily stylized. They rely on mathematically simple linear and quadratic dependencies on $\tau$, and a more direct empirical calibration of the $\tau$–parameter mappings using cross-cultural data would be an important direction for future work.

Table. 1 summarizes main model variables and parameters

## 2. Modeling Predictions

**2.1. Simple Special Case.** To gain analytical insight, *SI Appendix* considers a simplified setting in which individuals differ only in their attitudes $y$, infer beliefs exclusively from the observed average behavior, so that $\tilde{y}_i = \bar{x}$ for every individual, and are unable to choose actions that contradict their preferences. We show that if the distribution of attitudes $y$ remains fixed, the population evolves toward an equilibrium, which can take multiple forms.

Stable equilibria include a fully silent state, where no one expresses their opinion due to a high cost $c$, and states where individuals supporting a single norm express their opinions while others remain silent. These latter states are more likely in tight societies, characterized by large $k_2$. Notably, two such states can be stable simultaneously, meaning that the eventual outcome depends on initial conditions. However, the new norm is sustained over a broader range of parameters than the old norm.

Equilibria where both opinions are expressed require moderate values of the cost $c$ and benefit $b$. Increasing cultural tightness $\tau$ expands the range of $b$ values that support such equilibria.

The proportion of silent individuals, $1 - p - q$, as well as pluralistic ignorance $I$ (measured by the difference between $\bar{y}$ and $\bar{\tilde{y}}$), is greater in tight societies compared to loose societies.

When attitudes $y$ evolve, their distribution develops up to three sharp peaks, corresponding to $x = -1$, 0, and 1.

**2.2. Numerical Simulations.** To model more realistic situations when individuals vary in their psychological characteristics, can endorse the norm they privately oppose, and when their attitudes and second-order beliefs are updated incrementally, agent-based simulations are required. Choosing the right parameters and initial conditions is crucial for such simulations, as discussed next.

We generate individual values of the parameters $k_i$, $\alpha_i$, $\beta_i$ and $\gamma_i$ by random sampling from independent Beta distributions with mean values dependent on cultural tightness $\tau$ (*SI Appendix*, expressions S6) and a common SD $\sigma$. We consider a range of values of $\tau$ from small (in loose cultures) to large (in tight cultures) and vary parameters $b$ and $c$.

The initial values of individual attitudes $y$ are sampled from a Beta distribution with a mean $\bar{y}_0$ and a SD $\sigma$, with $\bar{y}_0$ ranging from 0 to $-0.8$. A mean of $\bar{y}_0 = 0$ implies that the population on average is ambivalent between the two policies, while $\bar{y}_0 = -0.8$ represents deeply ingrained preferences for the old norm. These distributions are shown in the first column of graphs in the figures below. The initial values of second-order beliefs $\tilde{y}$ are assumed to match attitudes.

The simulations address how a population with specific psychological characteristics (cultural tightness $\tau$ and initial average internalization strength $\bar{y}_0$ of the old norm evolves when the new norm ($x = 1$) provides a benefit $b$. We will systematically vary the parameters $\tau$, $\bar{y}_0$, the cost $c$, and the benefit $b$.

We will measure the frequencies $p$ and $q$ of individuals choosing $x = -1$ and $x = 1$, respectively, with $1 - p - q$ denoting the proportion who remain silent. We also examine the distributions of attitudes $y$ and second-order beliefs $\tilde{y}$. The frequency $q$ also serves as a measure of cultural evolutionary mismatch. We will measure pluralistic ignorance by the population-average gap $I$ between actual and perceived attitudes. A positive $I$ indicates overestimation of support for the new norm, whereas a negative $I$ indicates underestimation. Preference falsification is quantified by the fraction $\Phi$ of individuals whose expressed action $x$ contradicts their private attitude $y$ (i.e., $xy < 0$), with silent individuals ($x = 0$) excluded. Table. 1summarizes the individual and population-level variables as well as its parameters.

***2.2.1. Distribution of attitudes and second-order beliefs.*** The initial attitude distributions in the simulations are unimodal, covering a range of $y$ values. Over time, they form one, two, or three sharp peaks, typically at $y = -1$, $y = 1$, and $y = 0$. The peaks at $-1$ and 1 represent individuals strongly internalizing

**Table 1. Main model variables and parameters**

| | Symbols | Their meaning |
|---|---|---|
| Variables | $x$ | Action: $x = 1$, 0, or $-1$ |
| | $y$ | Attitude: $-1 \leq y \leq 1$ |
| | $\tilde{y}$ | Second-order belief: $-1 \leq \tilde{y} \leq 1$ |
| Parameters | $b, c$ | Benefit of the new norm and the cost of expressing an opinion |
| | $k_1, k_2$ | Effects of cognitive dissonance and social influence in decision-making |
| | $\alpha_1, \beta_1, \gamma_1$ | Effects of cognitive dissonance, social influence, and benefit in attitude adjustment |
| | $\alpha_2, \beta_2, \gamma_2$ | Effects of social projection, learning, and benefit in second-order belief adjustment |
| | $\tau$ | Cultural tightness: $0 \leq \tau \leq 1$ |
| Statistics | $p$ | The frequency of people supporting the new norm ($x = 1$) |
| | $q$ | The frequency of people supporting the old norm ($x = -1$); also a measure of cultural evolutionary mismatch |
| | $1 - p - q$ | The frequency of silent people ($x = 0$) |
| | $I$ | Pluralistic ignorance: the difference between the average values of $y$ and $\tilde{y}$ |
| | $\Phi$ | Preference falsification: the proportion of people for whom $xy < 0$ |

old and new norms, while the peaks at $y = 0$ indicate indifference and silence. The distributions of $\tilde{y}$ qualitatively resemble those of $y$.

**2.2.2. Long-term behavior.** Theoretical analysis of mathematical models has traditionally focused on understanding their long-term behavior. Below is a summary of the patterns observed at $t = 25,000$. The results of numerical simulations can be accessed at https://volweb2.utk.edu/~gavrila/PI/main.html.

When the new norm offers no direct benefit ($b = 0$), it persists only if the initial internalization of the old norm is weak. The frequency of the new norm is higher in loose societies, where individuals place greater weight on their personal norms relative to conformity pressures.

When $b > 0$, three outcomes emerge (Fig. 1): full adoption if old-norm internalization is weak and $b$ is large; coexistence for intermediate internalization; and rejection of the new norm despite its benefits if internalization is strong or $b$ is small. Attitude distributions range from uni- to trimodal, and high cost ($c$) yields many silent individuals.

At equilibrium, pluralistic ignorance and preference falsification vanish, although a silent minority may persist when $c$ is large. Tight cultures take much longer to equilibrate than loose ones. In every case, the new norm attains a higher frequency in tight societies than in loose ones, because in loose societies individuals with strong old-norm internalization resist switching despite material and social incentives. The persistence of the old norm for $b > 0$ constitutes a cultural–evolutionary mismatch that can arise in both tight and loose societies. In tight cultures, it persists through peer conformity; in loose cultures, through deeply held convictions in the old norm.

**2.2.3. Short-term behavior.** The foregoing analysis describes long-term asymptotics, but transient dynamics on shorter time can differ substantially. Analyzing transient dynamics on these shorter time scales is particularly important because external conditions, such as environmental factors, economic policies, or system parameters, rarely remain constant over extended periods (56, 57). By focusing on transient dynamics, we can gain insights into how systems respond to changes, adapt to shocks, or evolve before reaching their asymptotic states, making this analysis highly relevant to real-world applications.

Fig. 2 shows the state at $t = 500$, analogous to Fig. 1. Loose societies with low initial old-norm internalization display a higher frequency of the beneficial norm early on (*Top-Left* of Fig. 2). If the society is initially unbiased in their attitudes on average (so that $\bar{y} = 0$; first row of graphs in Fig. 2), intermediate tightness maximizes adoption. Silent individuals remain prevalent in tight societies when $\bar{y}_0$ is large.

**2.2.4. Rate of change.** Behavior, attitudes, and second-order beliefs evolve faster in loose than in tight cultures (Fig. 3and *SI Appendix*, Fig. S2). In societies where the old norm is strongly internalized, the new norm spreads more slowly or may not spread at all.

**2.2.5. Transient pluralistic ignorance.** The mismatch $I = |y - \tilde{y}|$ (Fig. 4) rises to a peak then declines. Peaks occur sooner and are higher in loose societies and at low internalization $\bar{y}_0$—when beliefs update rapidly—and shrinking for large $\bar{y}_0$, where both attitudes and beliefs change slowly. PI can emerge without silence when $\tilde{y}$ lags $y$ (*SI Appendix*, Fig. S2). Preference falsification steadily decreases. *SI Appendix*, Fig. S3, presents heatmap graphs illustrating the dependencies of $I$ and the number of time steps to reach it on cultural tightness $\tau$ and the initial strength of the old norm internalization $\bar{y}_0$.

**2.2.6. Attitude estimation bias.** As $p, y \to 1$, $\bar{y}$ is underestimated; as $q, y \to -1$, it is overestimated. Tight cultures exhibit stronger bias, perceiving $\bar{y}$ closer to zero than it truly is (*SI Appendix*, Fig. S2).

**2.2.7. Overestimation of unity.** Attitude and belief distributions can be broad (Figs. 1 and 2). Both very tight and very loose societies may polarize—when $b$ is small in loose cultures, and when $b$ is large in tight ones—driven by personal norms vs. mutual influence. Beliefs $\tilde{y}$ are more concentrated around the mean than attitudes $y$, causing an underestimation of polarization, especially in tight contexts.

## 3. Discussion

Pluralistic ignorance is a pervasive problem that systematically undermines societies' ability to adopt changes that serve collective interests. When people consistently misjudge what others think—as we see in domains ranging from climate action to reproductive rights—entire communities can remain stuck with policies and norms that most people privately oppose, creating a fundamental barrier to beneficial social change. Yet despite its prevalence and importance, our understanding of the evolution of PI remains very limited.

To address this void, our mathematical analysis unifies the classical problems of spirals of silence, pluralistic ignorance (PI), preference falsification, the strength of social norms, and cultural–evolutionary mismatch within a single framework. We model the coevolution of actions $x$, private attitudes $y$, and second-order beliefs $\tilde{y}$ in a setting where silence, conformity, and cultural tightness interact dynamically. Our model makes a number of theoretical contributions.

First, by embedding material payoffs, cognitive dissonance and conformity in one utility function and coupling it with separate update rules for $y$ and $\tilde{y}$, the model reproduces several phenomena endogenously. Earlier treatments either held attitudes fixed, equated beliefs with visible behavior, or ignored abstention; consequently they were missing some important dynamic components like the lags between $y$ and $\tilde{y}$ that underlie PI.

Second, this unified framework advances our theoretical understanding of social change by revealing how psychological and cultural factors interact to produce counterintuitive dynamics. Unlike previous models that treated these phenomena in isolation, our approach demonstrates that pluralistic ignorance, preference falsification, and cultural persistence emerge from the same underlying mechanisms operating at different timescales. This insight bridges previously disconnected literatures in social psychology, cultural evolution, and political science.

Finally, in the model, we made cultural tightness-looseness an explicit, testable parameter $\tau$ which maps onto concrete psychological weights (*SI Appendix*, Eq. **S6**). This has allowed us to parse how the effects of the same intervention (e.g., a benefit $b$) propagates very differently in tight vs. loose societies: loose cultures move quickly but overshoot in early PI, while tight cultures move slowly yet can ultimately achieve higher adoption of a beneficial norm (Figs. 1–4).

Our model shows that with binary choices, internalized norms and peer pressure can pin a population to a suboptimal equilibrium indefinitely which is impossible in continuous-choice versions of the framework (17, 37), which always drift to the material optimum. This provides a formal evolutionary mechanism for the persistence of maladaptive norms such as child marriage. Mathematical models often yield different

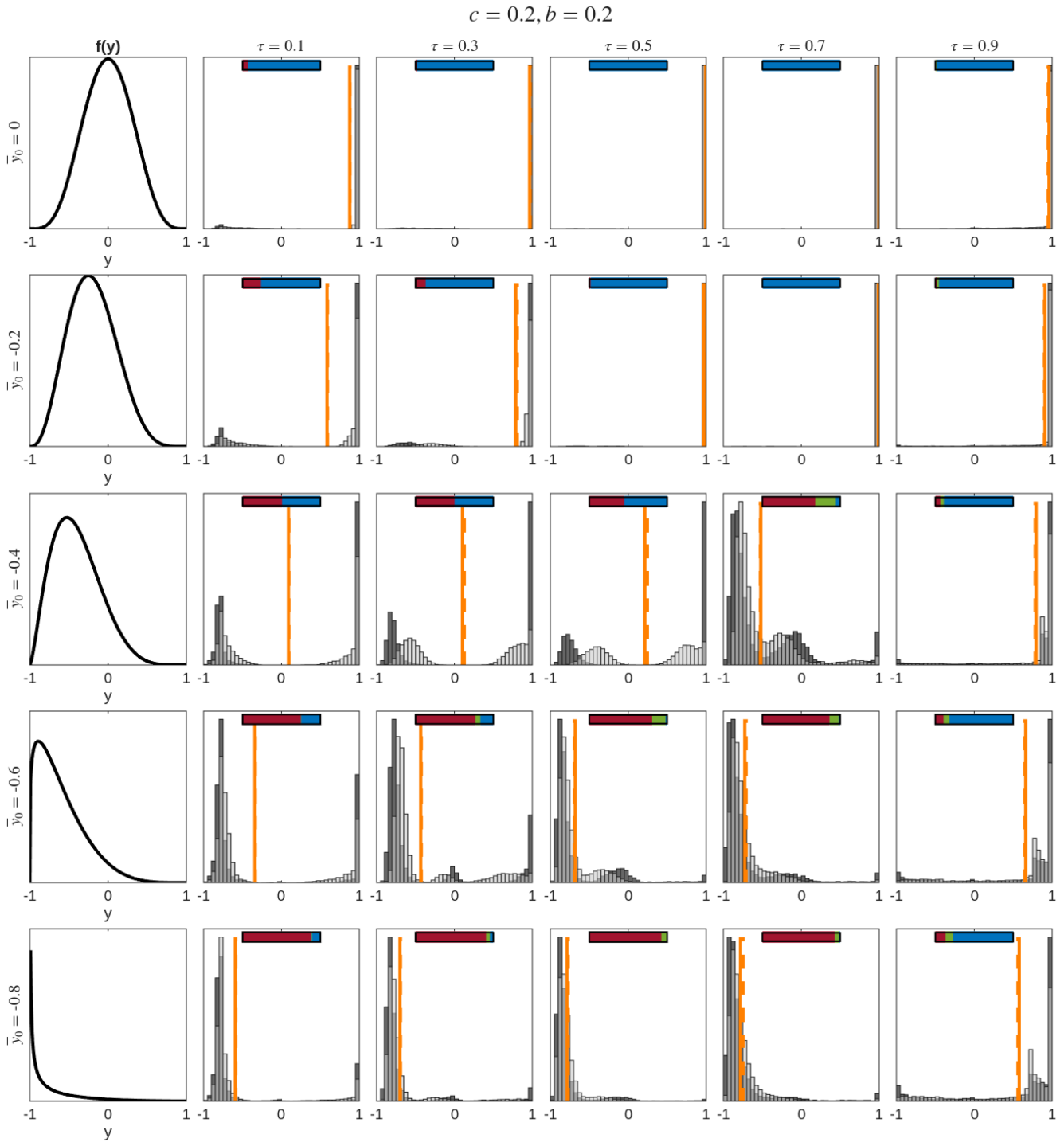


**Fig. 1.** Examples of population states after 25,000 time steps for various initial distributions of individual attitudes $f(y)$ (left column) and different levels of societal tightness $\rho$. The histograms represent the distributions of attitudes $y$ (black) and second-order beliefs $\tilde{y}$ (light gray). The horizontal bars indicate the frequency of individuals choosing $x = -1$ (blue), $x = 0$ (red), and $x = 1$ (green). The gap between the solid and dashed vertical lines, which represent the mean attitude $y$ and the mean belief $\tilde{y}$ respectively, reflects the extent of pluralistic ignorance in the population. Parameters: $n = 10,000$, $b = 0.2$, $c = 0.2$, $\sigma = 0.15$. Both $y$ and $\tilde{y}$ are updated at rate 0.01.

predictions depending on the structure of the action space (58–60) Continuous spaces permit smooth, incremental adjustments, and typically predict gradual convergence. Discrete spaces, by contrast, impose thresholds that individuals must cross before switching actions, introducing rigidity that can strongly shape collective outcomes especially under social influences such as conformity, norms, or peer pressure.

Most explanations of pluralistic ignorance trace the bias to what people see and say. If individuals stay silent, a spiral of silence hides true opinions; if they lie, preference falsification distorts the public signal. Our model shows that even when every person speaks honestly, a gap between reality and perception can still form because private attitudes $y$ and second-order beliefs $\tilde{y}$ move at different speeds.

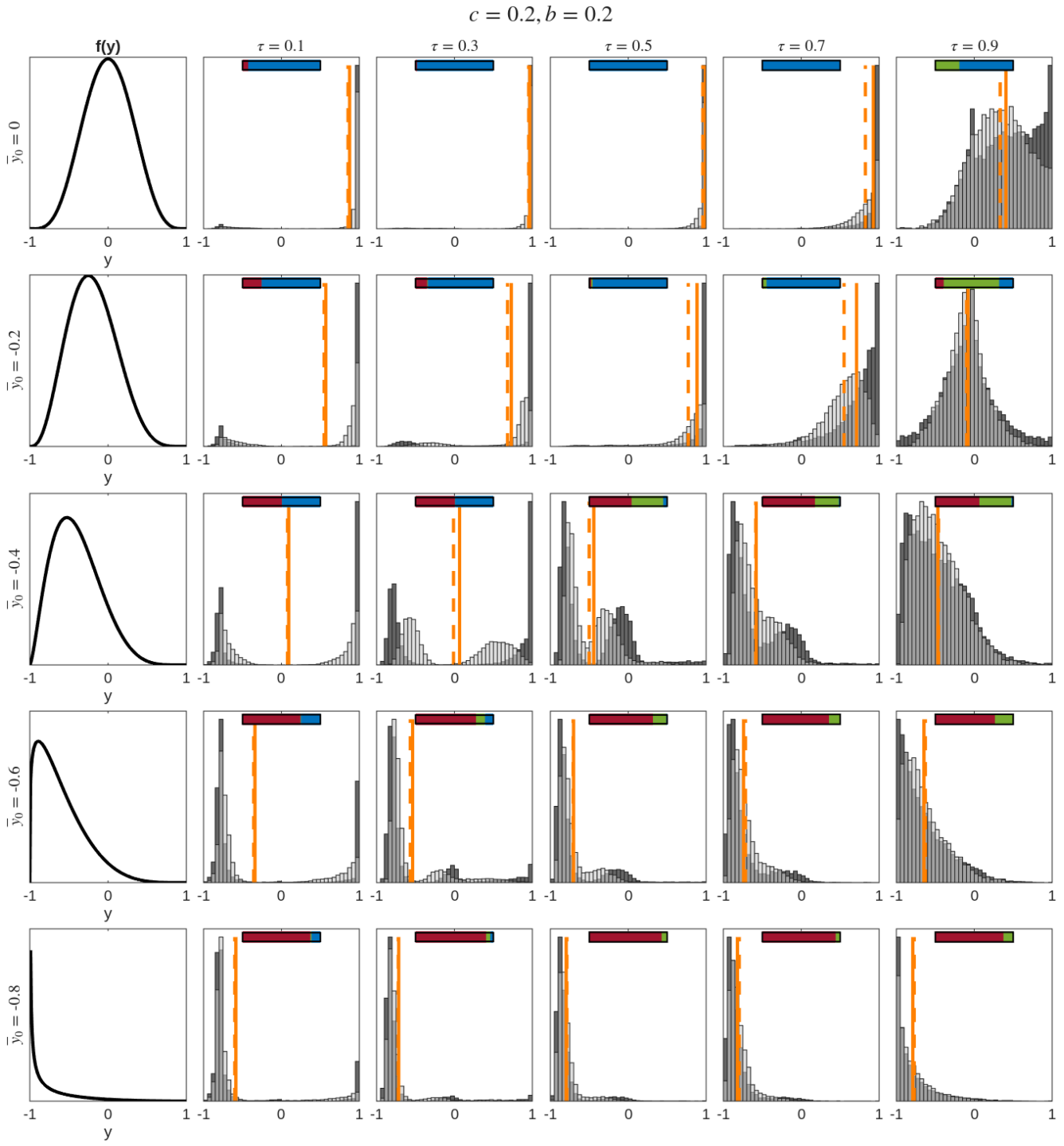


**Fig. 2.** As in Fig. 1but at $t = 500$.

Why do these rates diverge? While both personal attitudes and second-order beliefs respond to observed peer behavior and material benefits, each is also shaped by a distinct psychological process: personal attitudes are influenced by cognitive dissonance, whereas second-order beliefs rely on social projection. When someone changes their private attitude $y$ but does not change their behavior, that shift remains hidden and has no effect on others' beliefs. Once the person changes their behavior, it reinforces their attitude through cognitive dissonance and effort justification, further pulling $y$ toward the new norm. However, early on, when only a few individuals change their attitudes and behavior, these changes are too limited to noticeably affect the population-level behavior $\bar{x}$. As a result, others see no strong reason to revise their second-order beliefs. Even those who have updated both their attitudes and behavior do not immediately project their new attitudes onto others, due to cognitive inertia modeled by the assumption that $\alpha_2 < 1$. These combined effects create an initial lag between $y$ and $\tilde{y}$, which only begins to shrink as more individuals switch and the population moves closer to a new equilibrium. The lag is magnified in loose societies, where

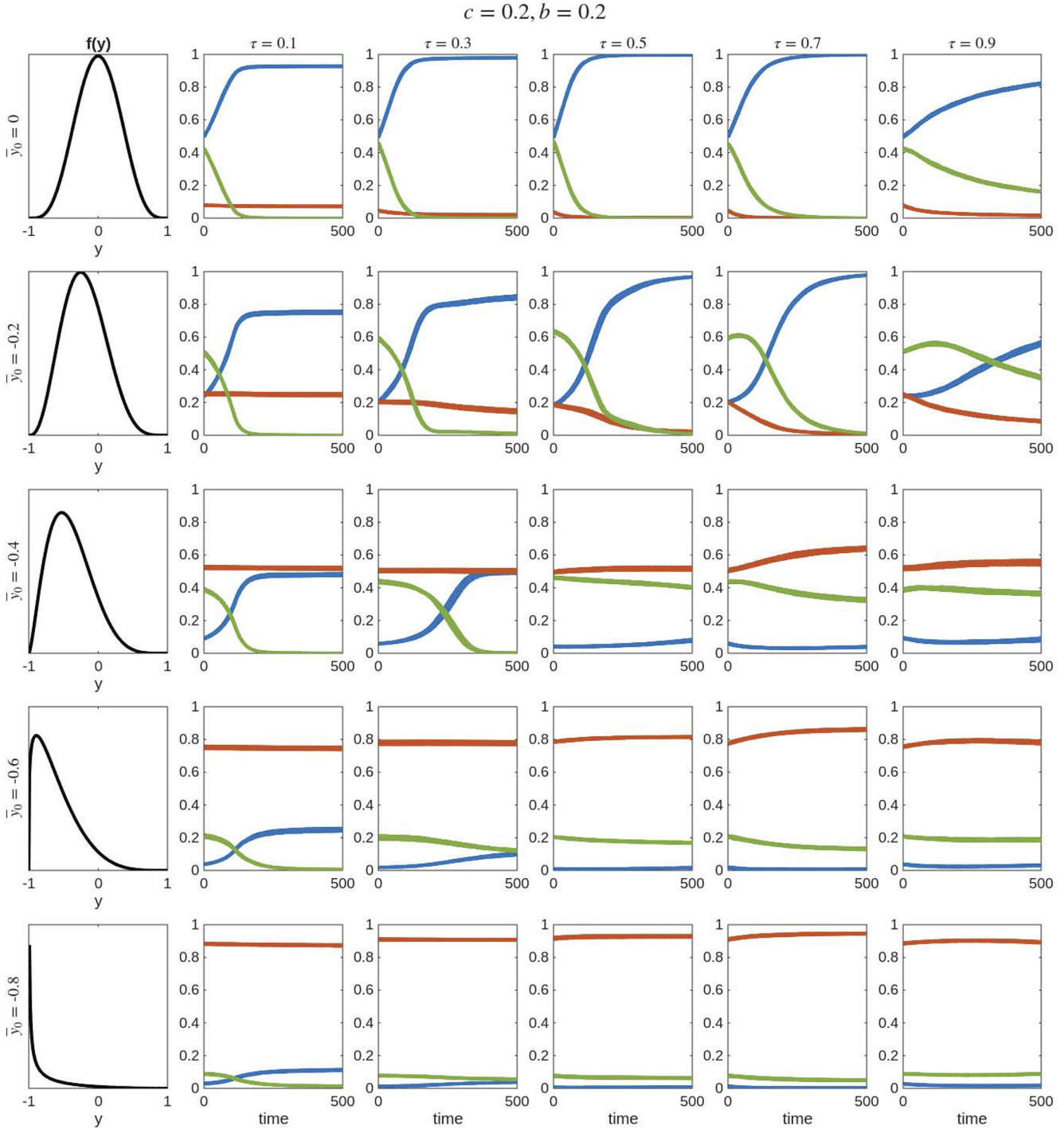


**Fig. 3.** The dynamics of the frequencies $p$ (blue; new norm supporters), $q$ (red; old norm supporters), and $1 - p - q$ (green; silent individuals) with $b = c = 0.2$ in 10 independent runs with parameter values used in Fig. 2.

individuals are more willing to express their true attitudes, feel less pressure to conform to others, and have a stronger psychological drive to align their behavior with their internal views.

The dynamic fingerprint of these processes include an initial rise in PI which then falls. Empirically this means that cross-sectional surveys taken at different moments along the trajectory can yield opposite inferences about PI's prevalence. In contrast, preference falsification falls monotonically and can be very low even when PI is still significant. Longitudinal or cohort-sequential designs are needed to separate the transient gap from the equilibrium state.

Our results show that both tight and loose societies can experience cultural–evolutionary mismatch, though for different reasons. In tight societies, the main barrier is strong conformity pressure, which makes it difficult or costly for individuals to deviate in either behavior or beliefs. In these settings, lowering the cost of expressing alternative views-for example, through anonymity or legal protections-can be a key lever for change. In contrast, loose societies are less constrained by conformity but may still be held back by deeply internalized old norms. These societies are more likely to adopt a new norm if interventions successfully shift attitudes, such as by clearly communicating the

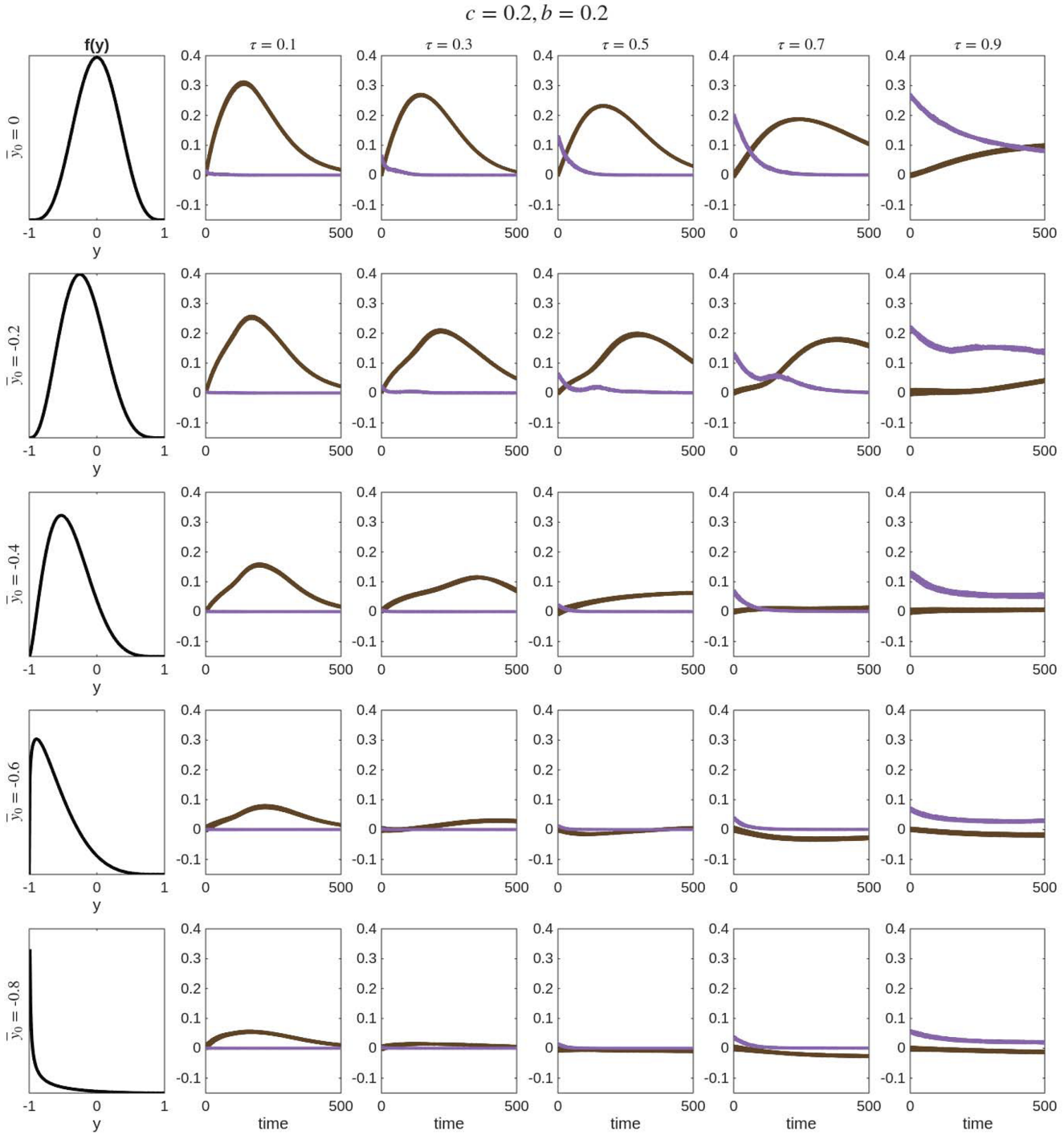


**Fig. 4.** The dynamics of the average values of pluralistic ignorance *I* (dark brown curves) and preference falsification Φ (purple) in 10 independent runs with parameter values used in Fig. 2.

benefits of the new behavior (i.e., offering a large *b*). Societies with intermediate levels of tightness show nonmonotonic patterns of adoption: because they respond to both material incentives and conformity signals, the most effective interventions combine direct benefits with visible peer support to accelerate norm change.

In the introduction, we highlighted three global surveys-on climate action, women's rights, and abortion attitudes that reveal strong mismatches between individuals' personal attitudes and their beliefs about others. The model accounts for these misperceptions as outcomes of different stages along a common adjustment path, where private attitudes shift more quickly than second-order beliefs. In the climate donation and basic women's rights data, most individuals already support the prosocial norm, but their beliefs about others lag behind, leading to systematic underestimation of public support. The model predicts that this gap will close fastest in loose, climate-vulnerable societies where belief updating is less constrained. In contrast, the affirmative-action data capture countries near a tipping point: in those where support remains below 50%, it tends to be

overestimated, while in those where it has just surpassed the majority threshold, it is underestimated. This pattern matches the model's prediction when the population average $\bar{y}$ moves through zero. The U.S. abortion data from the post-Roe period follow the same logic, with the largest lags found in tight states and among strongly "pro-life" individuals, where private attitudes are changing but beliefs about others remain anchored in the past.

Our results may help explain findings like those of ref. 61, who found that pluralistic ignorance regarding climate attitudes was stronger in culturally looser countries than in tighter ones across 55 nations. Our model suggests this pattern could reflect loose cultures being captured at a point in the trajectory when PI peaks due to faster attitude change outpacing belief updating, while tight cultures may have been measured during their slower, more gradual adjustment phase. In contrast, preference falsification falls monotonically and can be very low even when PI is still significant. Longitudinal or cohort-sequential designs are needed to separate the transient gap from the equilibrium state.

In our model, each time step $t$ consists of a sequence of three updates: action, attitude, and second-order belief. The mapping from $t$ to calendar time should be calibrated to the empirical setting. A practical strategy is to align $t$ with the observed rate of updating opportunities. Panel surveys, experience-sampling or diary methods, and digital-trace data can estimate how often individuals face decision points (for behaviors) or topic exposures (for attitudes). For example, climate-action opportunities such as recycling can occur 5 to 10 times per day, implying that $t = 500$ corresponds to roughly 50 to 100 d, whereas exposure to abortion-related content may be weekly or monthly among politically engaged individuals, placing $t = 500$ on the order of 10 to 40 y. Because opportunity rates vary across people and settings (e.g., by political interest), the effective clock can run at different speeds; smartphone sensors, app logs, social-media/news consumption traces, and conversation data could provide scalable ways to quantify these rates and thereby anchor the model's temporal dynamics to real-world updating opportunities.

Importantly, tightness-looseness (TL) is a dimension of normative constraint distinct from other dimensions of culture such as individualism-collectivism (IC) (35, 62). Cross-cultural research shows that TL and IC are only moderately correlated (35). Tight collectivist societies (e.g., Singapore, South Korea, Japan) differ from loose collectivist ones (e.g., Brazil, Spain) in norm enforcement and responses to deviance (35). Likewise, tight cultures can be individualistic (Germany) or collectivistic (China). Tightness also varies within both individualistic and collectivistic contexts: U.S. states differ markedly in tightness despite high individualism (63). Nevertheless, given that IC and TL have sometimes been shown to independently predict the same phenomena through different mechanisms (64–66) and that TL can amplify effects of IC and other cultural dimensions (67), future research should be done to examine these effects empirically.

Our results have several practical implications. First, we show how policy levers differ by culture. In loose settings, messaging that raises $b$ (material or status benefits) moves attitudes quickly, but second-order beliefs lag; thus publicizing accurate opinion polls or peer pledges is also vital. Second, in tight settings, lowering the expressive cost $c$ (e.g., via anonymity or legal shields) is more effective because conformity dominates cognition; once a visible minority forms, change can cascade. Third, monitoring the PI distance rather than the sign-mismatch flag provides an early-warning signal: a rising gap without growth in silence heralds an imminent flip. These findings have profound implications for democratic societies and social movements. The model suggests that apparent public opinion stability may mask underlying attitude shifts, potentially misleading policymakers about the true state of public sentiment. For social movements, the results indicate that early stages of change may be particularly vulnerable to backlash, as rising pluralistic ignorance can create false impressions of isolation among supporters. Understanding these dynamics becomes crucial for maintaining democratic legitimacy during periods of rapid social change.

Here we focused on a well-mixed population to keep the core mechanisms transparent and analytically tractable. In general, incorporating network structure—such as localized peer influence, clustering, or echo chamberscan substantially alter both the dynamics and equilibrium outcomes. Network structure can be introduced by replacing the global mean $\bar{x}$ in Eqs. **2** and **3** with a local, possibly visibility-weighted neighborhood average, $\bar{x}_i = \sum_j w_{ij} x_j$, where $w_{ij}$ represents the weight of interpersonal influence. *SI Appendix*, Figs. S4 and S5 illustrate that centrally positioned individuals can exert disproportionate influence on collective outcomes and that network structure alone can sustain local pluralistic ignorance, in contrast to the well-mixed population where it disappears at equilibrium. These examples further suggest that network effects are strongest in loose cultures and when the initial internalization of the new norm is low (i.e., when $\tau$ and $\bar{y}_0$ are small). A full network analysis incorporating degree heterogeneity, homophily, clustering, directed ties, and algorithmic curation would provide valuable insights but lies beyond the scope of this paper.

For simplicity, we treated the parameters $b$ (benefit of expressing a given behavior) and $c$ (cost of expression) as constant, but in reality both are likely to vary with cultural tightness. For example, $c$ may be higher in tight cultures due to stronger social sanctions, while $b$ may be lower in loose cultures where greater gender equality or norm diversity reduces the perceived benefit of signaling support for progressive change. Ignoring these relationships may confound the effects of tightness with those of incentives, so future empirical work should model $b$ and $c$ as functions of $\tau$. Moreover benefit $b$ may increase with the observed frequency $p$ of individuals supporting the new norm or with its perceived support in the population, captured by the average of $\tilde{y}$. Allowing $b$ to depend on these quantities would strengthen the feedback between individual decision-making and collective behavior, effectively increasing the population's cultural tightness.

Our approach has several additional limitations that can be removed in future work. Our cultural tightness parameter, while empirically grounded, captures only one dimension of cultural variation and may not fully represent the complexity of real cultural systems. In our framework, we assumed a binary choice and a continuous one-dimensional attitude. Alternatively, attitudes can be modeled as bidimensional, specifying approval of the old norm and approval of the new norm, where these two dimensions are not perfectly negatively correlated (68–71). For the present study, and in line with previous work (37, 40), we adopted the simpler one-dimensional representation to focus on the core dynamics of pluralistic ignorance. Conceptually, this bipolar scale can also be interpreted as the difference between the two approvals (approval of the new norm minus approval of the old norm). Extending the model to cases where both actions and beliefs are treated as either discrete or continuous variables represents an important direction for future research.

Finally, it would be valuable to estimate the model's parameters directly from surveys or experiments (40, 41), which would allow us to test the model's assumptions and evaluate whether the predicted effects of cultural tightness on attitudes, beliefs, and behavior are supported by empirical data.

By marrying utility theory with belief dynamics and embedding cultural tightness, our model clarifies when and why whole societies misread themselves and how those misreadings eventually unwind. It not only reconciles disparate empirical patterns under a common logic, but also yields actionable levers for accelerating norm change or, conversely, for stabilizing valued traditions against the noise of transient shocks.

**Data, Materials, and Software Availability.** There are no data underlying this work.

**ACKNOWLEDGMENTS.** We thank the reviewers for valuable comments and suggestions. Supported by the U.S. Army Research Office Grants W911NF-18-1-0138, the Air Force Office of Scientific Research Grant FA9550-21-1-0217, and the John Templeton Foundation.

1. P. Andre, T. Boneva, F. Chopra, A. Falk, Globally representative evidence on the actual and perceived support for climate action. *Nat. Clim. Change* **14**, 253–259 (2024).
2. L. Bursztyn, A. W. Cappelen, B. Tungodden, A. Voena, D. H. Yanagizawa-Drott, "How gender norms are perceived?" (NBER Working Papers Series, 2024), pp. 31–49.
3. N. C. Dias, J. N. Druckman, M. S. Levendusky, Unraveling a "cancel culture" dynamic: When, why, and which Americans sanction offensive speech. *J. Polit.* **87**, 588–600 (2025).
4. G. Fornaro, Conservative bias in perceptions of public opinion among citizens: Perceived social norms about abortion rights in post-roe United States. *Polit. Sci. Res. Methods*, 1–10 (2025).
5. H. J. O'Gorman, The discovery of pluralistic ignorance. *J. Hist. Behav. Sci.* **22**, 333–347 (1986).
6. D. T. Miller, C. McFarland, "When social comparison goes awry: The case of pluralistic ignorance" in *Ocial Comparison: Contemporary Theory and Research*, J. Suls, T. A. Wills, Eds. (Lawrence Erlbaum, 1991), pp. 287–313.
7. J. Shamir, M. Shamir, Pluralistic ignorance across issues and over time: Information cues and biases. *Public Opin. Q.* **61**, 227–260 (2007).
8. R. H. Sargent, L. S. Newman, Pluralistic ignorance research in psychology: A scoping review of topic and method variation and directions for future research. *Rev. Gen. Psychol.* **25**, 163–184 (2021).
9. L. Bursztyn, D. Y. Yang, Misperceptions about others. *Annu. Rev. Econ.* **14**, 425–452 (2021).
10. D. T. Miller, A century of pluralistic ignorance: What we have learned about its origins, forms, and consequences. *Front. Soc. Psychol.* **1**, 1260896 (2023).
11. E. Noelle-Neumann, The spiral of silence a theory of public opinion. *J. Commun.* **24**, 43–51 (1974).
12. G. Dixon, S. Bashian, K. Snelling, The influence of minority views on majority participation in online discourse. *J. Media Psychol. Theor. Methods Appl.* **37**, 256–268 (2025).
13. T. Kuran, Sparks and prairie fires: A theory of unanticipated political revolution. *Public Choice* **61**, 41–74 (1989).
14. R. B. Cialdini, N. J. Goldstein, Social influence: Compliance and conformity. *Annu. Rev. Psychol.* **55**, 591–621 (2004).
15. G. Song, Q. Ma, F. Wu, L. Li, The psychological explanation of conformity. *Soc. Behav. Pers.* **40**, 1365–1372 (2012).
16. P. J. Richerson, S. Gavrilets, F. B. M. de Waal, Modern theories of human evolution foreshadowed by Darwin's Descent of Man. *Science* **372**, eaba3776 (2021).
17. S. Gavrilets, *Social Influence and the Logic of Collective Action* (Princeton University Press, Princeton, NJ, 2026).
18. D. G. Taylor, Pluralistic ignorance and the spiral of silence: A formal analysis. *Public Opin. Q.* **46**, 311–335 (1982).
19. M. Granovetter, R. Soong, Threshold models of diversity: Chinese restaurants, residential segregation, and the spiral of silence. *Sociol. Methodol.* **18**, 69–104 (1988).
20. M. Granovetter, Threshold models of collective behavior. *Am. J. Sociol.* **83**, 1420–1443 (1978).
21. D. Centola, R. Willer, M. Macy, The emperor's dilemma: A computational model of self-enforcing norms. *Am. J. Sociol.* **110**, 1009–1040 (2005).
22. R. Bénabou, J. Tirole, "Laws and norms" (NBER Working Paper 17579, National Bureau of Economic Research, 2011).
23. M. Fernández-Duque, The probability of pluralistic ignorance. *J. Econ. Theory* **102**, 105449 (2022).
24. M. DeGroot, Reaching a consensus. *J. Am. Stat. Assoc.* **69**, 118–121 (1974).
25. B. D. Anderson, M. Ye, Recent advances in the modelling and analysis of opinion dynamics on influence networks. *Int. J. Autom. Comput.* **16**, 129–149 (2019).
26. M. Ye, Y. Qin, A. Govaert, B. D. Anderson, M. Cao, An influence network model to study discrepancies in expressed and private opinions. *Automatica* **107**, 371–381 (2019).
27. D. Sohn, N. Geidner, Collective dynamics of the spiral of silence: The role of ego-network size. *Int. J. Public Opin. Res.* **28**, 25–45 (2015).
28. B. Ross *et al*., Are social bots a real threat? An agent-based model of the spiral of silence to analyse the impact of manipulative actors in social networks. *Eur. J. Inf. Syst.* **28**, 394–412 (2019).
29. D. Sohn, Spiral of silence in the social media era: A simulation approach to the interplay between social networks and mass media. *Commun. Res.* **49**, 139–166 (2022).
30. B. Cabrera, B. Ross, D. Röchert, S. Stieglitz, The influence of community structure on opinion expression: An agent-based model. *J. Bus. Econ.* **9**, 1331–1356 (2021).
31. D. Vilone, E. Polizzi, Modeling opinion misperception and the emergence of silence in online social system. *PLoS ONE* **19**, e0296075 (2024).
32. N. Nunn, "On the causes and consequences of cross-cultural differences: An economic perspective" in *Handbook of Advances in Culture and Psychology*, M. J. Gelfand, C. Y. Chiu, Y. Y. Hong, Eds. (Oxford University Press, Oxford, UK, 2022), pp. 125–188.
33. N. Nunn, On the dynamics of human behavior: The past, present, and future of culture, conflict, and cooperation. *AEA Pap. Proc.* **112**, 15–37 (2022).
34. M. J. Gelfand, Cultural evolutionary mismatches in response to collective threat. *Curr. Dir. Psychol. Sci.* **30**, 5401–5409 (2021).
35. M. J. Gelfand *et al*., Differences between tight and loose cultures: A 33-nation study. *Science* **332**, 1100–1104 (2011).
36. S. Gavrilets, The dynamics of injunctive social norms. *Evol. Hum. Sci.* **2**, e60 (2020).
37. S. Gavrilets, Coevolution of actions, personal norms, and beliefs about others in social dilemmas. *Evol. Hum. Sci.* **3**, e44 (2021).
38. S. Gavrilets, P. J. Richerson, Authority matters: Propaganda and the coevolution of behaviour and attitudes. *Evol. Hum. Sci.* **4**, e51 (2022).
39. I. Alger, S. Gavrilets, P. Durkee, Proximate and ultimate drivers of norms and norm change. *Curr. Opin. Psychol.* **60**, 101916 (2024).
40. S. Gavrilets, D. Tverskoi, A. Sanchez, Modeling social norms: An integration of the norm-utility approach with beliefs dynamics. *Philos. Trans. R. Soc. Lond. B* **379**, 20230027 (2024).
41. D. Tverskoi, A. Guido, G. Andrighetto, A. Sánchez, S. Gavrilets, Disentangling material, social, and cognitive determinants of human behavior and beliefs. *Hum. Soc. Sci. Commun.* **10**, 236 (2023).
42. S. Gavrilets *et al*., Co-evolution of behaviour and beliefs in social dilemmas: Estimating material, social, cognitive and cultural determinants. *Evol. Hum. Sci.* **6**, e50 (2024).
43. C. Bicchieri, *The Grammar of Society. The Nature and Dynamics of Social Norms* (Cambridge University Press, Cambridge, UK, 2006).
44. J. I. Krueger, From social projection to social behaviour. *Eur. Rev. Soc. Psychol.* **18**, 1–35 (2007).
45. A. Elster, M. J. Gelfand, When guiding principles do not guide: The moderating effects of cultural tightness on value-behavior links. *J. Pers.* **89**, 325–337 (2021).
46. E. Dimant, M. Gelfand, A. Hochleitner, S. Sonderegger, Strategic behavior with tight, loose, and polarized norms. *Manage. Sci.* **71**, 2245–2263 (2024).
47. S. Kitayama, A. C. Snibbe, H. R. Markus, T. Suzuki, Is there any "free" choice? Self and dissonance in two cultures. *Perspect. Psychol.* **15**, 527–533 (2004).
48. E. Hoshino-Browne, A. S. Zanna, S. J. Spencer, M. P. Zanna, S. Kitayama, On the cultural guises of cognitive dissonance: The case of Easterners and Westerners. *J. Pers. Soc. Psychol.* **89**, 294–310 (2005).
49. S. J. Heine, D. R. Lehman, Culture, dissonance, and self-affirmation. *Person. Soc. Psychol. Bull.* **23**, 389–400 (1997).
50. D. R. Ames, Strategies for social inference: A similarity contingency model of projection and stereotyping in attribute prevalence estimates. *J. Pers. Soc. Psychol.* **87**, 573–585 (2004).
51. J. I. Krueger, From social projection to social behavior. *Eur. Rev. Soc. Psychol.* **18**, 1–35 (2007).
52. Y. Rosokha, X. Lyu, D. Tverskoi, S. Gavrilets, "Cooperation under the shadow of political inequality" (SSRN 4496690, 2023).
53. D. T. Miller, D. A. Prentice, Collective mistakes and misperceptions at Princeton: Revising the pluralistic ignorance explanation of alcohol use on campus. *Adv. Exp. Soc. Psychol.* **26**, 161–209 (1994).
54. S. De, D. S. Nau, M. J. Gelfand, Understanding norm change: An evolutionary game-theoretic approach. arXiv [Preprint] (2017). https://doi.org/10.48550/arXiv.1704.04720 (Accessed 22 January 2026).
55. J. C. Jackson *et al*., Ecological and cultural factors underlying the global distribution of prejudice. *PLoS ONE* **14**, e0221953 (2019).
56. A. Hastings, Transients: The key to long-term ecological understanding? *Trends Ecol. Evol.* **19**, 39–45 (2004).
57. A. Hastings *et al*., Transient phenomena in ecology. *Science* **361**, eaat6412 (2018).
58. K. Tuyls, R. Westra, "Replicator dynamics in discrete and continuous strategy spaces" in *Multi-Agent Systems. Simulation and Applications*, A. M. Uhrmacher, D. Weyns, Eds. (CRC Press, Boca Raton, FL, 2009), pp. 215–241.
59. W. H. Sandholm, *Population Games and Evolutionary Dynamics* (MIT Press, Cambridge, MA, 2010).
60. W. Zhong, S. Kokubo, J. Tanimoto, How is the equilibrium of continuous strategy game different from that of discrete strategy game? *Biosystems* **107**, 88–94 (2012).
61. S. J. Geiger *et al*., What we think others think and do about climate change: A multicountry test of pluralistic ignorance and public-consensus messaging. *Perspect. Psychol.*, 1–22 (2025).
62. H. C. Triandis, The self and social behavior in differing cultural contexts. *Psychol. Rev.* **96**, 506–520 (1989).
63. J. R. Harrington, M. J. Gelfand, Tightness-looseness across the 50 United States. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 7990–7995 (2014).
64. E. Stamkou, D. W. van Knippenberg, M. J. Gelfand, J. Homan, D. A. van Kleef, Cultural collectivism and tightness moderate responses to norm violators: A cross-cultural investigation. *Org. Behav. Hum. Decis. Process.* **147**, 67–82 (2018).
65. C. Crossland, D. C. Hambrick, Differences in managerial discretion across countries: How nation-level institutions affect the degree to which CEOs matter. *Strateg. Manag. J.* **32**, 797–819 (2011).
66. C. S. Eun, L. Wang, S. C. Xiao, Culture and R2. *J. Fin. Econ.* **115**, 283–303 (2015).
67. R. Fischer, J. A. Karl, Predicting behavioral intentions to prevent or mitigate COVID-19: A cross-cultural meta-analysis of attitudes, norms, and perceived behavioral control effects. *Soc. Psychol. Pers. Sci.* **13**, 264–276 (2021).
68. V. Taras, B. L. Kirkman, P. Steel, Examining the impact of culture's consequences: A three-decade, multilevel, meta-analytic review of Hofstede's cultural value dimensions. *J. Appl. Psychol.* **95**, 405 (2010).
69. J. T. Cacioppo, W. L. Gardner, G. G. Berntson, Beyond bipolar conceptualizations and measures: The case of attitudes and evaluative space. *Pers. Soc. Psychol. Rev.* **1**, 3–25 (1997).
70. I. K. Schneider, N. Schwarz, Mixed feelings: The case of ambivalence. *Curr. Opin. Behav. Sci.* **15**, 39–45 (2017).
71. P. Briñol, R. E. Petty, M. Stavraki, "Structure and function of attitudes" in *Oxford Research Encyclopedia of Psychology*, M. A. Hogg, Ed. (Oxford University Press, New York, NY, 2019).

# Supplementary information
# The cultural evolution of pluralistic ignorance
**Sergey Gavrilets, Johannes Karl and Michele J. Gelfand**

**A. Best response actions.** The best response actions are

$$x = 1 \text{ if } k_1 y + k_2 \tilde{y} > c - b, \tag{S1a}$$
$$x = -1 \text{ if } k_1 y + k_2 \tilde{y} < -c - b, \tag{S1b}$$
$$x = 0 \text{ otherwise.} \tag{S1c}$$

**B. Simplified model.** When individuals cannot choose the action that contradicts their private attitude, the best-response actions are as follows. If $y > 0$, choose $x = 1$ if $y > d_{\min}$, and choose $x = 0$ otherwise. If $y < 0$, choose $x = -1$ if $y < d_{\max}$, and choose $x = 0$ otherwise. The thresholds are:

$$d_{\min} = \frac{-b + c - k_2 \tilde{y}}{k_1}, \tag{S2a}$$
$$d_{\max} = \frac{-b - c - k_2 \tilde{y}}{k_1}. \tag{S2b}$$

When $d_{\min} > 0$ and $d_{\max} < 0$, individuals on both sides of the norm spectrum may choose silence.

We consider a population of individuals who differ in their attitudes $y$. Let $F(y)$ be the cumulative distribution function (cdf) of $y$ on $[-1, 1]$ with $F = 0$ for $y < -1$ and $F = 1$ for $y > 1$. The proportion of people with $y > 0$ is $P = 1 - F(0)$, and those with $y < 0$ are $Q = F(0)$.

Assume that, while the individuals differ in their attitudes $y$, there are no differences between them in parameters $c, b, k_i, \alpha_i, \beta_i, \gamma_i$. Furthermore, let individuals form their beliefs completely on the basis of observed average behavior $\bar{x}$. In this case, each individual has exactly the same second-order belief $\tilde{y} = \bar{x}$. (For example, this is the case if all $\alpha_1 = \gamma_1 = 0, \beta_1 = 1$).

Let us rescale the parameters $c$ and $b$ relative to $k_1$, so that the new parameter $c$ is the original parameter $c$ divided by $k_1$, and the new parameter $b$ is the original parameter $b$ divided by $k_1$. Define $r = k_2/k_1$, which can be viewed as a measure of cultural tightness: in loose societies, cognitive dissonance is stronger than conformity ($r < 1$), while in tight societies, conformity dominates ($r > 1$). Note that parameters $\tau$ of the main text and $k$ used here are related: $\tau = r/(r + 1)$.

Then adapting the Granovetter model (20, 73), in the next time step, the frequencies of people choosing $x = 1$ and $x = -1$ are

$$p' = 1 - F(d_p), \tag{S3a}$$
$$q' = F(d_q), \tag{S3b}$$

where, using rescaled parameters, $d_p = \max(0, -b + c - r\bar{x})$ and $d_q = \min(0, -b - c - r\bar{x})$. Because the average observed behavior in the next time step is $\bar{x}' = \frac{p'-q'}{p'+q'}$. one can combine equations (S3) into a single recurrence equation for $\bar{x}$:

$$\bar{x}' = \frac{1 - F(d_p) - F(d_q)}{1 - F(d_p) + F(d_q)}. \tag{S3c}$$

***B.1. Constant attitudes.*** Suppose that attitudes $y$ remain fixed, which occurs when $\alpha_2 = \beta_2 = \gamma_2 = 0$. Under these conditions, the frequencies $P$ and $Q$ do not change, allowing us to identify some equilibria of the dynamic system (S3).

- Silent population. The state where nobody expresses their opinions ($p = q = \bar{x} = 0$) is an equilibrium if

  $$c > 1 + b,$$

  meaning that the cost $c$ is sufficiently large.

- Old norm dominance. The state in which all individuals who prefer old norm express their opinions, while all those preferring new norm remain silent ($p = 0, q = Q, \bar{x} = -1$), is an equilibrium if

  $$r > b + \max(c, 1 - c).$$

- New norm dominance. The state in which all individuals who prefer new norm express their opinions, while all those preferring old norm remain silent ($p = P, q = 0, \bar{x} = 1$), is an equilibrium if

  $$r > -b + \max(c, 1 - c).$$

  Both these states require the society to be sufficiently tight (so that $r$ is sufficiently large), but with positive $b$, the range of parameter values for new norm prevalence is broader than that for old norm prevalence.

- Complete expression of opinions. The state where everyone expresses their opinion ($p = P, q = Q, \bar{x} = P - Q$) cannot be an equilibrium.

**Uniform distribution of $y$.** The three equilibria described above exist for any distribution $f(y)$ of attitudes $y$ in the population. To gain further insight into the model, we need to specify this distribution. The simplest case assumes $y$ is uniformly distributed between $-1$ and 1, which implies that $F(y) = \frac{1+y}{2}$.

With this choice of $f(y)$, one can identify three additional types of equilibria where both opinions are expressed, and some individuals remain silent. In one equilibrium, all individuals with $y > 0$ express their opinion ($p^* = P = 1/2$) while some individuals with $y < 0$ remain silent ($q^* < Q$). At this equilibrium

$$\bar{x}^* = \frac{1/2 - F(-b - c - r\bar{x}^*)}{1/2 + F(-b - c - r\bar{x}^*)}.$$

This equality leads to a quadratic equation for $\bar{x}^*$. At the other equilibrium the situation is reversed: $p^* < P$ and $q^* = Q$. At this equilibrium

$$\bar{x}^* = \frac{F(-b + c - r\bar{x}^*) - 1/2}{F(-b + c - r\bar{x}^*) + 1/2}.$$

This equality also leads to a quadratic equation for $\bar{x}^*$.

In the third equilibrium, some individuals of both types remain silent, so that $p^* < P$ and $q^* < Q$. This requires that $d_p > 0$ and $d_q < 0$. From recurrence equations (S3a-S3b), it can be shown that at this equilibrium:

$$p^* + q^* = 1 - c,\ p^* - q^* = \frac{b(1-c)}{1-c-r},\ \bar{x}^* = \frac{b}{1-c-r}.$$

Note that the first equality implies $c$ must be smaller than 1 and that the frequency of silent individuals is $c$.

A necessary condition for this equilibrium to be feasible is $-1 \le \bar{x}^* \le 1$. This leads to the following conditions on $r$

$$r < 1 - c - b \text{ or } r > 1 - c + b. \qquad [S4]$$

That is, the equilibrium $(p^*, q^*)$ can exist in loose societies with small $r$ (provided that $b + c < 1$) or in tight societies with large $r$ (provided that $c < 1$).

To ensure that $d_p > 0, d_q < 0$, it should also be the case that $-b - c \le r\bar{x}^* \le b - c$. Solving these inequalities for $r$ one finds the following conditions:

$$r < 1 - c - b\ \frac{1-c}{c} \text{ or } r > 1 - c + b\ \frac{1-c}{c}. \qquad [S5]$$

Conditions (S4) are stricter than conditions (S5) if $c > 1/2$ but weaker if $c < 1/2$.

This equilibrium is locally stable if $r < 1 - c$, that is, in sufficiently loose societies.

Figure S1 illustrates these results. It shows the existence of two stable equilibria at $x = -1$ and $x = 1$, in which half of the population expresses its opinion while the other half remains silent. These equilibria are stable for sufficiently large values of $r$, with the range of $r$ values resulting in an equilibrium at $x = 1$ being larger than that at $x = -1$ due to the additional benefit $b$ associated with the former. There is also a stable equilibrium at intermediate positive values of $\bar{x}$, which emerges at small values of $r$, that is, in loose cultures.

With a uniform distribution of attitudes, the population average of $y$ is zero, so the absolute value $|\bar{x}^*|$ serves as a measure of the extent of pluralistic ignorance. Figures S1 shows that $|\bar{x}^*|$ increases with cultural tightness $r$.

***B.2. Changing attitudes.*** With changing attitudes, from equation (2) one finds that at equilibrium,

$$y = \frac{x + r_y|\bar{x}|\bar{x} + \frac{\gamma_1}{\alpha_1}b}{1 + r_y|\bar{x}|},$$

where $r_y = \beta_1/\alpha_1$.

This implies that there can be no more than 3 different values of $y$ corresponding to $x = -1, 0$ and 1, so the cdf function $F(y)$ is a step function with three steps at $y = -1, 0$ and 1 of heights $1 - q, 1 - p$ and 1, respectively.

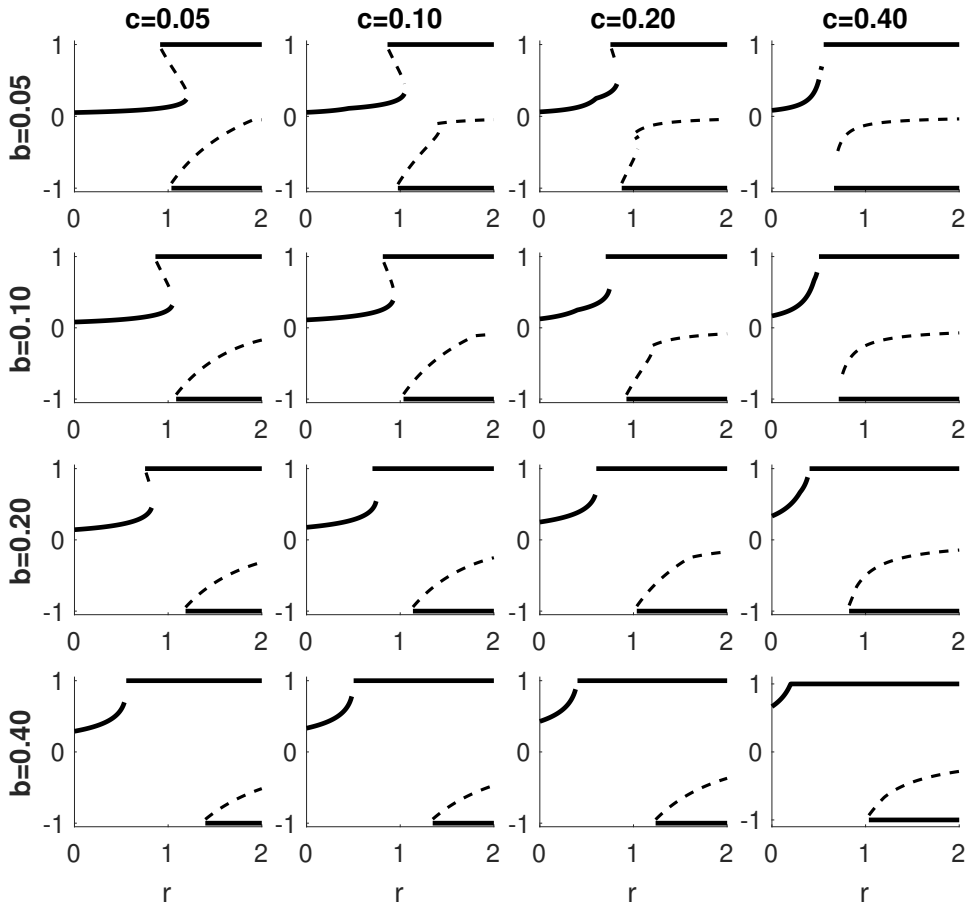


**Fig. S1.** Bifurcation diagrams in the model with a uniform and stable distribution of attitudes $y$, where second-order beliefs $\tilde{y}$ are based on the observed average behavior $\bar{x}$. Plotted are the equilibrium values of $\bar{x}$ as functions of cultural tightness $r$, which serves as the bifurcation parameter, for different combinations of the benefit $b$ of supporting the new norm and cost $c$ associated with expressing an opinion. Solid curves represent stable equilibria; dashed curves represent unstable equilibria.

**C. Effects of cultural tightness $\tau$ on model parameters.** In numerical simulations, we assume that individual values of parameters are randomly drawn from Beta distributions with constant variance $\sigma$ and the following mean values:

$$k_1 = 1 - \tau,\ k_2 = \tau, \tag{S6a}$$

$$\alpha_1 = 1 - \tau,\ \beta_1 = \tau(1 - \tau),\ \gamma_1 = 1 - \tau, \tag{S6b}$$

$$\alpha_2 = 1 - \tau,\ \beta_2 = \tau(1 - \tau),\ \gamma_2 = \tau(1 - \tau). \tag{S6c}$$

This parameterization implies that in loose cultures ($\tau < 1/2$), cognitive forces dominate social influence ($k_1 > k_2$, $\alpha_1 > \beta_1$, $\alpha_2 > \beta_2$). In contrast, in tight cultures ($\tau > 1/2$), social influence has a greater impact on decision-making than cognitive dissonance ($k_2 > k_1$). Although the effects of cognitive forces on attitudes and beliefs ($\alpha_1, \alpha_2$) are smaller in tight cultures compared to loose ones, they still exceed those of social influence ($\alpha_1 > \beta_1$, $\alpha_2 > \beta_2$). In very tight cultures ($\tau \approx 1$), the influence of both cognitive and social forces on personal attitudes and second-order beliefs becomes minimal ($\alpha_i, \beta_i \approx 0$), resulting in little change to these variables. In loose cultures, $\gamma_1 > \gamma_2$. Note that the sum $\alpha_i + \beta_i$ represents the overall speed of belief change, while the sum $\gamma_1 + \gamma_2$ gives the overall effect of messaging. Under this parameterization, both sums are equal to $1 - \tau^2$, decreasing as cultural tightness increases.

**D. The dynamics of the average values of attitude $y$ and second-order belief $\tilde{y}$.** Figure S2 shows the dynamics of the average values of $y$ and $\tilde{y}$ for parameter values used in Figures 3 and 4 of the main text.

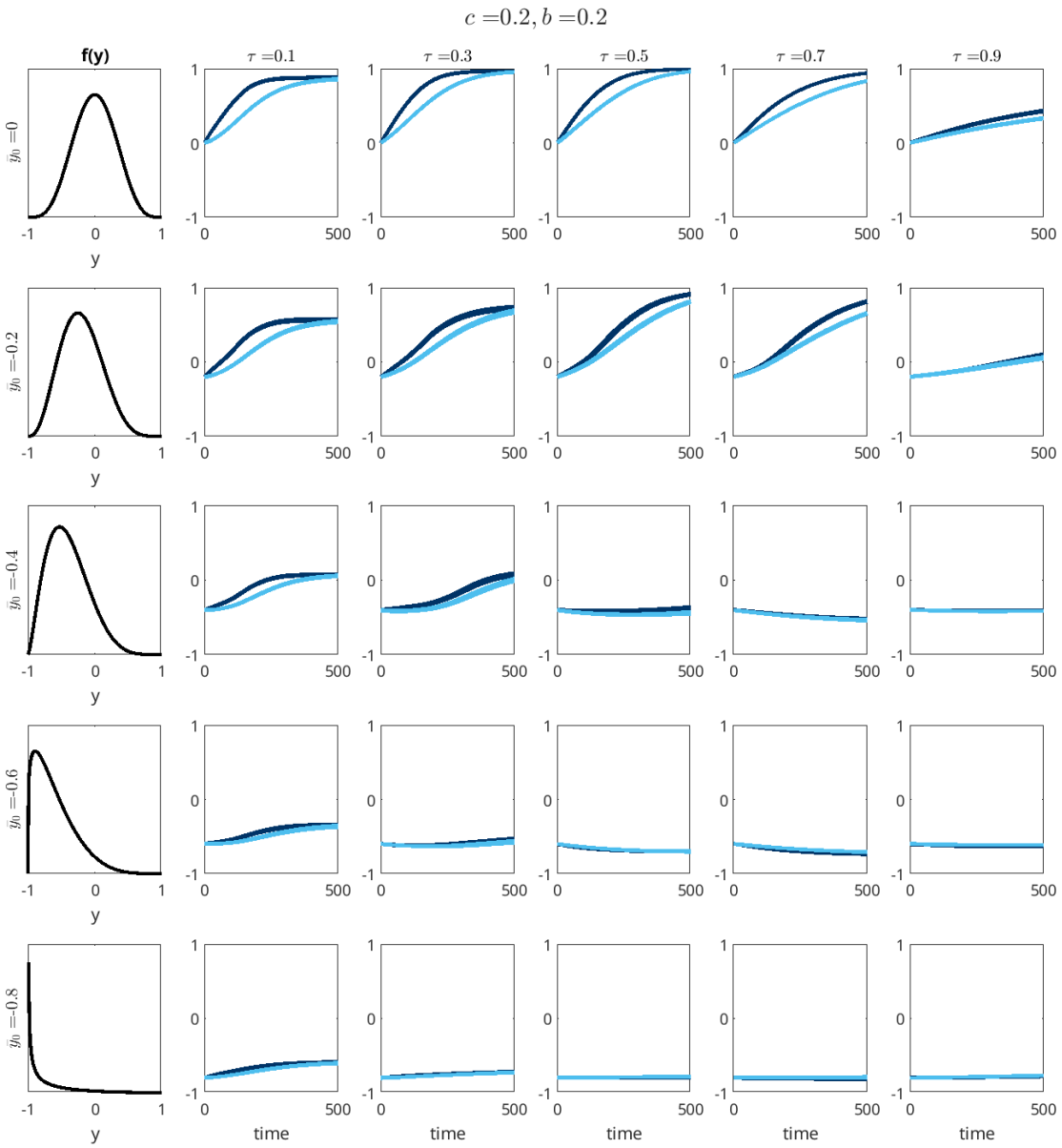


**Fig. S2.** The dynamics of the average values of attitude $y$ (dark blue) and second-order belief $\tilde{y}$ (light blues) in 10 independent runs with parameter values used in Figure 2.

Figure S3 illustrates the dependencies of the maximum level of pluralistic ignorance $I = \overline{y - \tilde{y}}$ and the number of time steps to reach it on cultural tightness $\tau$ and the initial strength of the old norm internalization $\bar{y}_0$.

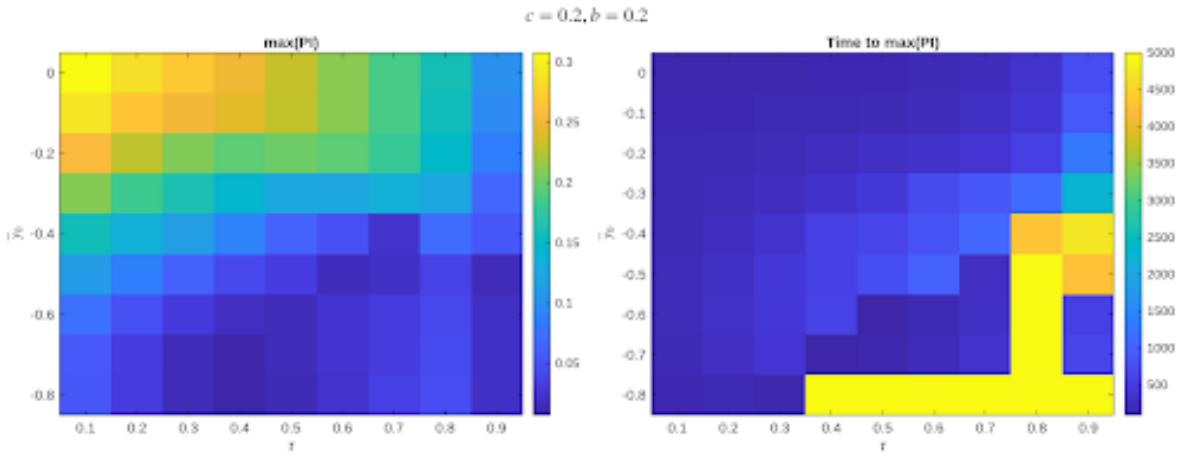


**Fig. S3.** The dependencies of maximum value of $I = \overline{y - \tilde{y}}$ (left graph) and the number of time steps to reach it (right graph) on cultural tightness $\tau$ (horizontal axes) and the initial strength of the old norm internalization $\bar{y}_0$ (vertical axes).

**Network structure and neighbor-weighted social signals.** To examine whether population structure can sustain pluralistic ignorance, we place the same individual-level decision process on simple networks and replace the global mean of expressed actions, $\bar{x}(t)$, with a *degree-biased neighbor mean* for each agent $i$:

$$\hat{x}_i(t) \;=\; \frac{\sum_{j\in N(i)} k_j^{\eta}\, x_j(t)}{\sum_{j\in N(i)} k_j^{\eta}},$$

where $N(i)$ is the set of neighbors of $i$, $k_j$ is neighbor $j$'s degree, and $\eta\varepsilon 0$ controls attention bias toward high-degree nodes ($\eta = 0$ reduces to the unweighted neighbor average).

**Toy network families and parameterization.** We use four standard synthetic graphs, each with $N = 3000$ nodes and mean degree $\approx 12$:

1. **Erdős–Rényi (ER):** $G(n, p)$ with $p = 12/n$, producing a homogeneous degree distribution without clustering.
2. **Barabási–Albert (BA):** Preferential-attachment with $(m_0, m) = (8, 6)$, yielding a heavy-tailed degree distribution with hubs.
3. **Watts–Strogatz (WS):** Small-world ring with $k = 12$ nearest neighbors and rewiring probability $\beta = 0.15$, combining high clustering with short paths.
4. **Star (STAR):** A single hub connected to all leaves, representing an extreme hub-dominated structure.

We set the attention-bias exponent to $\eta = 1.5$.

To create a "visibility minority," the top 5% of nodes by degree are initialized with the old-norm attitude ($y = -1$) and belief ($z = -1$); all other parameters match the well-mixed baseline. Because $\hat{x}_i(t)$ overweights high-degree neighbors when $\eta > 0$, a hub-dominated minority can keep the perceived social signal below the population average even as many agents privately shift their attitudes. As a result, $I$ can remain positive for long periods and the share of falsified expressions stays elevated, in contrast to the well-mixed model where these quantities peak and then decay toward zero.

Figures S4 and S5 show the dynamics of $p$ and $I$ across different network structures. The outcomes represented by curves of different colors diverge primarily in the upper-left region of the parameter space - corresponding to loose cultures (small $\tau$) and low initial internalization of the new norm (small $\bar{y}_0$).

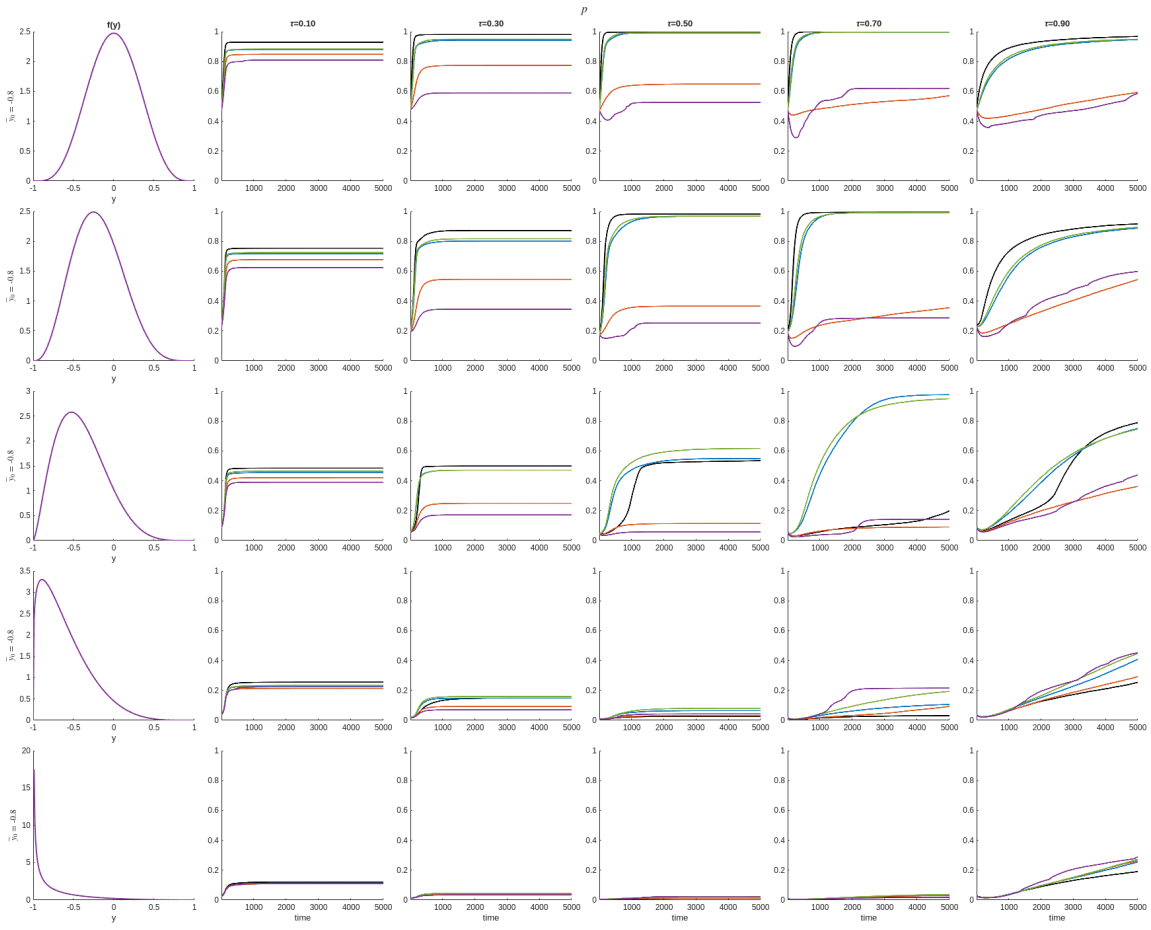


**Fig. S4.** The dynamics of $p$ in a well-mixed population and in four different networks. Five curves are overlaid in every panel: **black** = well-mixed; **blue** = Erdős–Rényi (ER, $p \approx 12/N$); **orange** = Barabási–Albert (BA, $m_0$=8, $m$=6); **green** = Watts–Strogatz (WS, $k$=12, $\beta$=0.15); **purple** = Star (single hub). Other parameters as in Figures 3,4 and S2.

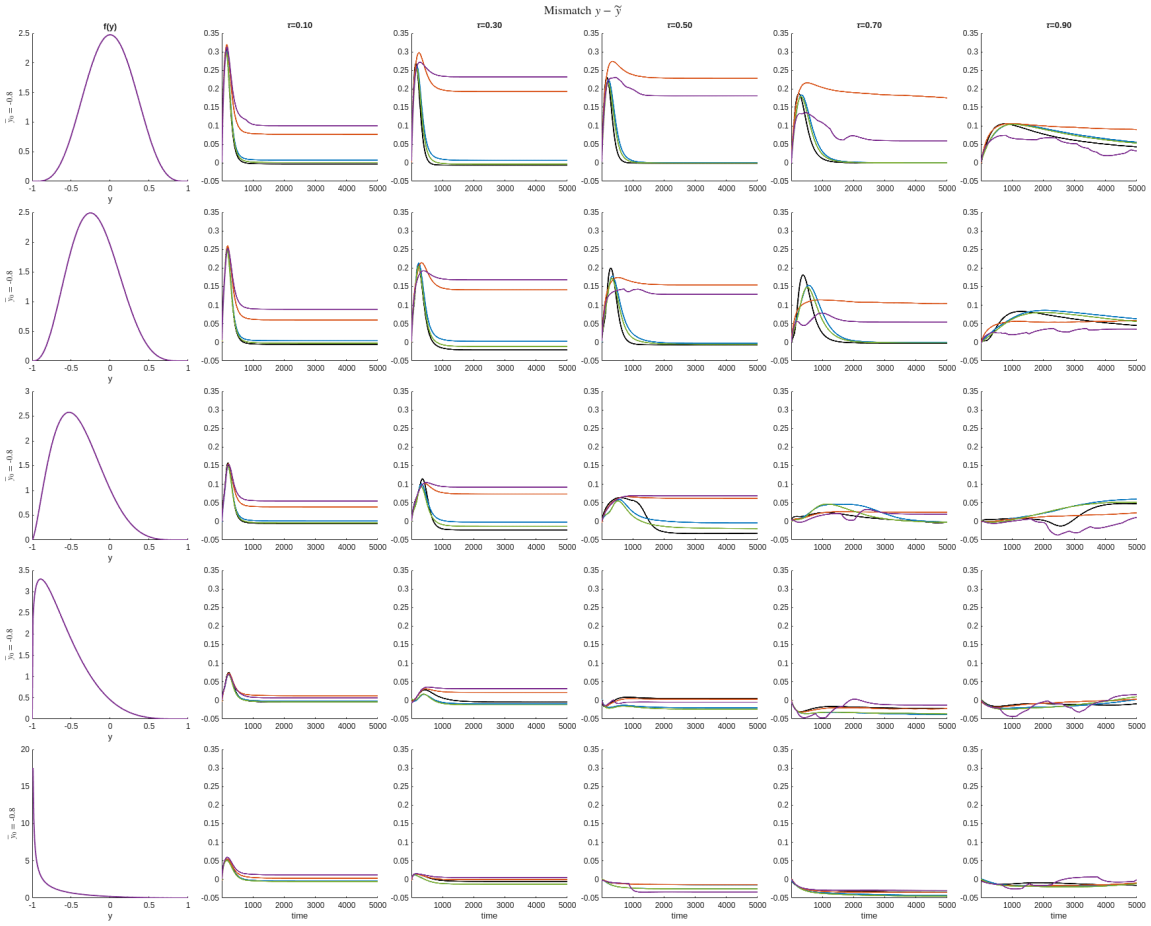


**Fig. S5.** Same as in Figure S4 but for the pluralistic ignorance $I$.