Credibility Governance: A Social Mechanism for Collective Self-Correction under Weak Truth Signals

Wanying He[1,2] **Yanxi Lin**[3] **Ziheng Zhou**[4] **Xue Feng**[2]
Min Peng[1] **Qianqian Xie**[1,*] **Zilong Zheng**[2,*] **Yipeng Kang**[2,*]
[1]School of Artificial Intelligence, Wuhan University, Wuhan, China
[2]State Key Laboratory of General Artificial Intelligence, BIGAI, Beijing, China
[3]Tsinghua University, Beijing, China
[4]University of California, Los Angeles, USA
wanying.he@whu.edu.cn, linyx22@mails.tsinghua.edu.cn
josephziheng@ucla.edu, feng.xue1580@gmail.com, pengm@whu.edu.cn
xqq.sincere@gmail.com, zlzheng@bigai.ai, kangyipeng@bigai.ai
[*]Corresponding authors

Abstract

Opinion aggregation on social media and web platforms increasingly determines how real-world resources are allocated. Yet current aggregation methods rely on easily amplified indicators such as vote-based engagement or capital-weighted commitments. These signals reflect visibility rather than reliability, making collective judgments vulnerable to early surges, strategic manipulation, and uneven evidence, especially when truth signals are weak, noisy, or delayed. To address this issue, we propose Credibility Governance (CG), a mechanism that reallocates influence based on how well agents track evolving signals rather than on voting or capital levels alone. CG assigns each opinion and each agent a dynamic influence score (credibility) that reflects how trustworthy they appear under observed signals. It updates opinion influence through weighted endorsements from reliable agents, and reciprocally updates agent influence based on the long-term performance of the opinions they support, rewarding early insights and consistent alignment with emerging evidence. This enables the system to distinguish persistent trends from short-lived noise. We evaluate CG in POLIS, a socio-physical simulation environment that models the co-evolution of social beliefs and physical feedback under uncertainty. Under weak signals, noisy observations, and initial majority misalignment, CG surpasses vote-based, capital-weighted, and no-governance baselines by enabling (i) earlier and more stable recovery of the true state, (ii) stronger resilience to misinformation shocks, and (iii) sustained support for correct minority viewpoints. By growing influence through demonstrated correctness rather than surface-level signals, CG contributes a governance mechanism that improves collective epistemic quality and supports more trustworthy, equitable, and socially beneficial online ecosystems. Code is available at https://github.com/Wanying-He/Credibility_Governance.

1 Introduction

Imagine you are a young researcher exploring a cross-disciplinary idea with quiet promise. You share a short working paper with tentative but intriguing pilot results, and discussion begins in an online forum where your field increasingly exchanges ideas. The signals are slow, noisy, and hard to interpret, yet the direction feels meaningful.

Over time, however, the conversation drifts. A competing line of work, promoted by a tightly connected group, produces a rapid stream of polished claims and frequent updates. Members highlight one another's progress and repost summaries, and their coordinated

activity shapes the platform metrics that funders and committees increasingly monitor. Although the underlying evidence remains thin, the visible volume of support grows quickly; aggregated signals begin to guide invitations, collaborations, and resource allocation. Your more careful line of inquiry becomes harder to detect, and later scholars inherit a distorted picture of what the field believes.

This scenario reflects a broader limitation of modern web-based platforms: rapid, reinforcing social signals can drown out slower but more reliable evidence Muchnik et al. (2013); Bakshy et al. (2015). Collective evaluations may track visibility rather than support, leading communities to converge prematurely on fragile directions Lorenz et al. (2011). In such environments, even well-intentioned participants can be pulled toward self-reinforcing consensus because the governance layer privileges what is quickly amplified over what is gradually validated.

The resulting challenge is institutional: when truth is weakly observable, influence rules can amplify capital, popularity, or early noise faster than evidence. Robust governance therefore requires influence updates that reward sustained signal tracking rather than short-lived momentum.

We ask whether a collective can self-correct under these weak-truth conditions. We propose *Credibility Governance* (CG), a mechanism that reallocates influence based on how well agents track evolving signals over time rather than on raw popularity or capital Golub and Jackson (2010). CG is designed to make influence dynamic and evidence-sensitive: agents gain influence when their support aligns with persistent, quality-weighted signal improvement, and lose influence when alignment is inconsistent or opportunistic.

Our contributions are: (i) a dual-component simulation environment linking a Physical World of topic progress with an Opinion World of agent deliberation; (ii) a formalization of CG and controlled comparisons to staking-, vote-, and no-governance baselines Sunstein (2006); and (iii) evidence that CG supports collective self-correction across noisy, delayed, and misleading environments, including boundary analyses of when it degrades Kleinberg and Easley (2022).

More broadly, we argue that governance quality depends not only on who is currently influential, but on how influence is updated through time under uncertainty. In weak-signal settings, this temporal design choice can determine whether a community drifts with momentum or progressively recovers toward epistemic accuracy.

2 Related Work

2.1 Truth Aggregation Mechanisms

Systems for aggregating dispersed judgments underpin web platforms and digital governance. Prior approaches broadly map to our baselines: capital-linked aggregation (staking), popularity-linked aggregation (voting/engagement), and no explicit influence governance.

Staking-based mechanisms emerged from decentralized coordination and blockchain governance frameworks Buterin (2014); Daian et al. (2020). Participants commit resources behind claims and are rewarded when supported outcomes perform well, echoing prediction-market logic Hanson (2006). While these designs can create incentives for accuracy, they do not guarantee epistemic quality: resource asymmetries can systematically bias visibility, coalition power, and outcome influence Frye (2021).

Vote-based and engagement-driven systems, common in social media and content-ranking ecosystems Leskovec et al. (2010), broaden participation but are vulnerable to momentum effects. Early surges can produce cascades where popularity itself becomes evidence. Reputation-weighted variants partially address this issue by incorporating historical reliability Resnick and Zeckhauser (2000); Squires et al. (2021), yet they often lag regime changes and can preserve stale authority after conditions shift.

No-governance settings form a third class: agents observe signals and update beliefs without an explicit influence-allocation mechanism. These systems are simple and transparent, but typically remain sensitive to noise and path dependence Salganik et al. (2006). Prior work thus motivates governance mechanisms that can preserve responsiveness while reducing early amplification bias.

2.2 Collective Misinformation Dynamics

Research in collective epistemology shows that groups can misaggregate information when early evidence is ambiguous and socially reinforced. Herding and cascade dynamics arise when individuals infer quality from others' actions rather than independent evidence Bikhchandani et al. (1992). Under partial observability, these cascades can persist even when many participants are locally competent.

Parallel work on misinformation diffusion finds that visibility, repetition, and coordinated amplification can help weakly supported claims spread faster than slower, better-grounded evidence Vosoughi et al. (2018); Pennycook and Rand (2021). Platform ranking and recommendation loops can further entrench these trajectories, producing echo-like structures and fragile consensus Del Vicario et al. (2016).

These findings suggest robustness requires governance rules that update source influence as evidence evolves, consistent with social-epistemic arguments about reliability-sensitive trust Goldman (1999).

2.3 LLM-Based Social Simulation

Agent-based models have long been used to study how local interactions generate macro-level social outcomes Epstein and Axtell (1996). Traditional models often rely on fixed heuristics; recent LLM-based agents expand this paradigm with language-mediated reasoning, richer contextual adaptation, and value-conditioned behavior (He et al., 2025; Li et al., 2025; Kang et al., 2020; Zhang et al., 2026; Ziheng et al., 2026; Kang et al., 2025).

These capabilities have enabled open-ended simulations of social and political interaction, including deliberation, coalition formation, and emergent norm dynamics (Ziheng et al., 2025; Smith et al., 2025; Zhang et al., 2025; Mao et al., 2025; Park et al., 2023; Argyle et al., 2023). Recent studies further use LLM populations to probe coordination and diffusion phenomena at scale Bratton et al. (2023); Chen et al. (2024); Schramowski et al. (2023).

We build on this line by using LLM-based social simulation as an evaluation environment for governance itself. Rather than treating aggregation as a static post-processing step, we model it as a dynamic mechanism that co-evolves with social signals, allowing us to test whether credibility-sensitive influence updates improve collective correction under weak and noisy truth signals.

3 Methodology

3.1 The POLIS Simulation Framework

3.1.1 Platform Capabilities

POLIS is a modular simulation platform for studying coupled social–physical dynamics with LLM-driven agent populations. It provides:

- **Scalable agent orchestration and logging:** persistent identities/memory and full trace capture of actions and state.
- **Parametric Physical World:** flexible resource-to-progress mappings for topic trajectories.
- **Pluggable governance modules:** interchangeable Opinion World aggregation and state-update rules.

3.1.2 Dual-World Structure and Interaction

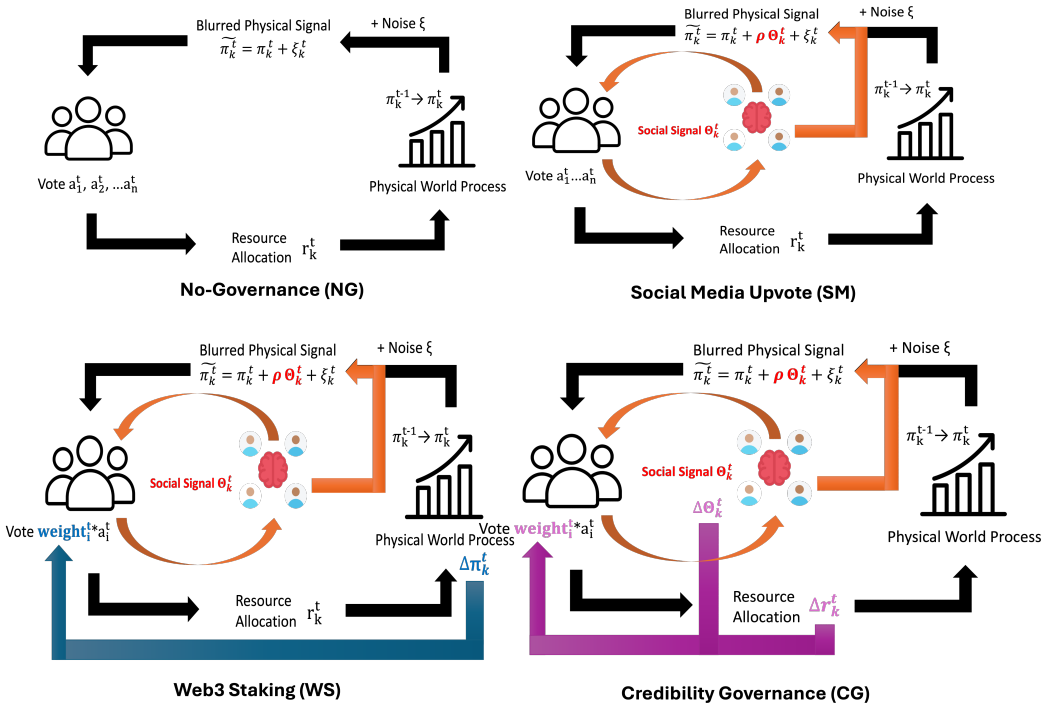


Figure 1: Comparison of the four governance mechanisms. (1) NG: Agents observe physical signals $\tilde{\pi}_k^t$. (2) SM: NG + social signal Θ_k^t in observation. (3) WS: SM + agent influence w_i^t updated by $\Delta\pi_k^t$. (4) CG: WS + agent influence w_i^t updated by Θ_k^t.

POLIS couples a *Physical World* (topic progress) with an *Opinion World* (social states and aggregation) over shared topics k:

- **Topic state:** physical progress π_k^t and social signal Θ_k^t.
- **Agent state:** beliefs/confidence α_i^t, choice a_i^t, and influence w_i^t (from credibility/stake, mechanism-specific).
- **Observations:** agents observe Θ_k^{t-1} and a noisy physical signal $\tilde{\pi}_k^{t-1}$.
- **Coupling:** votes a_i^t are aggregated into allocations $\mathbf{r}^t$, which update progress π_k^t; the mechanism updates Θ_k^t (and agent attributes), closing the feedback loop.

3.1.3 Round-Level Walkthrough

A simulation round evolves the Physical and Opinion Worlds from $t-1$ to t in a strictly causal loop. (1) Agents observe the previous round's public signals Θ_k^{t-1} and $\tilde{\pi}_k^{t-1}$. (2) They update beliefs/confidence α_i^t, select a topic a_i^t, and the governance mechanism aggregates votes using weights w_i^{t-1} to produce allocations $\mathbf{r}^t$. (3) The Physical World updates progress π_k^t from $\mathbf{r}^t$, while the mechanism updates social state variables (e.g., agent influence and topic signal) to yield Θ_k^t. The next round's noisy physical observation is then constructed as:

$$\tilde{\pi}_k^t = \pi_k^t + \rho\,\Theta_k^t + \xi_k^t.$$

3.2 Opinion World (Governance Mechanisms)

All mechanisms instantiate the same generic opinion-world protocol (Sec. 3.1.3), and only differ in (i) influence weights w_i, (ii) how the social signal Θ_k is updated, and (iii) whether any agent-level state is updated.

Generic mechanism template. At round t, each agent i observes signals and forms an intended choice $a_i^t \in \{1, \ldots, K\}$ with confidence α_i^t (Sec. 3.1.3). The governance rule then applies:

(1) Weighting.

$$w_i^{t-1} = g\big(\alpha_i^t,\ x_i^{t-1}\big), \tag{1}$$

where x_i^{t-1} is the mechanism-specific agent state (e.g., credibility c_i, balance bal_i, or empty).

(2) Aggregation.

$$r_k^t = \frac{\sum_i w_i^{t-1} \mathbf{1}[a_i^t = k]}{\sum_j w_j^{t-1}}, \qquad k \in \{1, \ldots, K\}. \tag{2}$$

(3) Topic-level update (optional).

$$\Theta_k^t = F\big(\Theta_k^{t-1},\ r_k^t,\ r_k^{t-1},\ \mathrm{aux}_k^t\big), \tag{3}$$

where aux_k^t denotes any mechanism-specific auxiliary terms.

(4) Agent-level update (optional).

$$x_i^t = U\big(x_i^{t-1},\ a_i^t,\ \Theta^t,\ \Theta^{t-1},\ r^t,\ \Delta\pi^t\big), \tag{4}$$

where $\Delta\pi_k^t$ is the topic progress increment in the physical world, and absent state variables are simply omitted.

Mechanism deltas (CG, WS, SM, NG). Credibility Governance (CG).
CG instantiates $x_i \equiv c_i$ (credibility) and uses credibility-adjusted aggregation, a momentum-based Θ update with supporter-quality and anti-bubble correction, and a credibility update driven by $\Delta\Theta$ with an early-mover factor.

$$w_i^{t-1} = \alpha_i^t \exp(\lambda c_i^{t-1}), \tag{5}$$

$$\Theta_k^t = (1 - \lambda_s)\Theta_k^{t-1} + \lambda_s \Big[(r_k^t - r_k^{t-1})\, \overline{q}_k^t - \gamma B_k^t\Big], \tag{6}$$

$$c_i^t = c_i^{t-1} + \eta\big(\Theta_{a_i^t}^t - \Theta_{a_i^t}^{t-1}\big)\, e^{-\kappa r_{a_i^t}^t}, \tag{7}$$

$$\overline{q}_k^t = \frac{\sum_{i:a_i^t=k} c_i^{t-1}}{\sum_{i:a_i^t=k} 1}, \tag{8}$$

$$B_k^t = \sigma\big(\alpha(r_k^t - r_k^{t-1})\big)\,(1 - \overline{q}_k^t). \tag{9}$$

Web3-style Staking (WS).
WS instantiates $x_i \equiv \mathrm{bal}_i$ (stake balance), sets influence proportional to stake, sets Θ_k^t to stake-weighted popularity, and updates stake by physical progress $\Delta\pi$.

$$w_i^{t-1} = \alpha_i^t \mathrm{bal}_i^{t-1}, \tag{10}$$

$$\Theta_k^t = r_k^t, \tag{11}$$

$$\mathrm{bal}_i^t = \mathrm{bal}_i^{t-1} + \gamma_s\, w_i^{t-1}\, \Delta\pi_{a_i^t}^t. \tag{12}$$

Social Media Upvote (SM).
SM uses uniform influence, sets Θ_k^t to raw popularity, and maintains no agent-level state.

$$w_i^{t-1} = 1, \tag{13}$$

$$r_k^t = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[a_i^t = k], \tag{14}$$

$$\Theta_k^t = r_k^t. \tag{15}$$

No Governance (NG).
NG maintains no social signal and no agent-level state. For convenience, one may still define the same uniform aggregation r_k^t as in SM, but Θ is identically zero and no updates are applied.

$$w_i^{t-1} = 1, \tag{16}$$

$$r_k^t = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[a_i^t = k], \tag{17}$$

$$\Theta_k^t \equiv 0. \tag{18}$$

3.3 Physical World Setups

In this work, we instantiate the Physical World as an academic setting with two scientific theories competing for collective resources. Their physical progress models how real-world scientific efforts unfold under different governance mechanisms.

3.3.1 Topic Initialization

We consider two future-oriented scientific topics, **quantum physics (true Topic A)** and **neuromorphic physics (false Topic B)**, to avoid LLM omniscience, as current evidence supports both sides and it is inconclusive which is more valid. Truth is assigned arbitrarily to avoid pre-existing LLM biases. The initial physical progress for both topics is set to 0.

3.3.2 Topic Progress Dynamics

Topic progress follows a nonlinear development trajectory. Let π_k^t denote the cumulative progress of topic k at round t and r_k^t the resource share allocated to topic k in that round. The increment $\Delta\pi_k^t$ is given by:

$$\Delta\pi_k^t = \begin{cases} \epsilon_k^t, & \pi_k^{t-1} < \pi_1, \\ \left(v_k + \gamma\, r_k^t\right)\left(1 - \dfrac{\pi_k^{t-1}}{M_k}\right) + \epsilon_k^t, & \pi_1 \le \pi_k^{t-1} < \pi_2, \\ \gamma' r_k^t\left(1 - \dfrac{\pi_k^{t-1}}{M_k}\right) + \epsilon_k^t, & \pi_k^{t-1} \ge \pi_2. \end{cases}$$

This models an initial *exploration* stage, an *acceleration* stage, and a *saturation* stage. Here, v_k is the intrinsic baseline velocity of topic k, γ and γ' control how resources accelerate progress in different growth phases, M_k is the saturation limit, and $\epsilon_k^t \sim \mathcal{N}(0, \sigma^2)$ captures environmental shocks.

3.3.3 Agent Population and Initial Conditions

We initialize $N = 100$ agents with the following belief and epistemic profiles. The 7:2:1 split allows us to test whether the population can self-correct from a skewed initial state and to observe whether high-quality agents can rise in influence over time.

- **Misaligned Majority (70 agents):** Start with a belief in the false Topic B and have moderate epistemic stability.
- **Truth-Aligned Minority (20 agents):** Start with a belief in the true Topic A and remain moderately influenceable.
- **High-Conviction Core (10 agents):** Start with a belief in the true Topic A and have very high epistemic stability, anchoring the system around ground truth.

Unless otherwise specified, agent- and topic-level states are initialized as:

- **Credibility (CG):** $c_i^0 = 1$ for all agents.
- **Stake (WS):** $\text{bal}_i^0 = 1$ for all agents.
- **Social signal (CG):** $\Theta_k^0 = 0$ for both topics.
- **Confidence:** $\alpha_i^0 = 0.5$ for all agents.

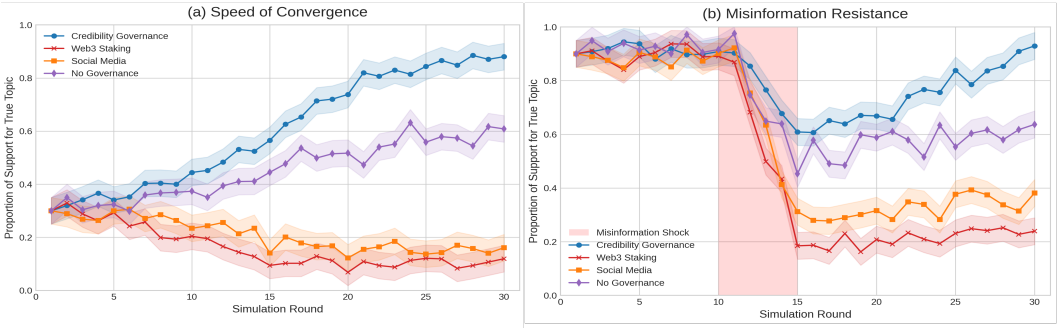


Figure 2: System-level convergence. (a) From false-majority initialization. (b) Recovery after a misinformation shock.

These initial conditions ensure that all mechanisms start from a symmetric, low-information state: no topic has accumulated progress, no agent has privileged influence, the social signal has not yet formed, and all agents start with a moderately strong initial confidence level.

3.3.4 Simulation Parameters

The specific parameter values used in our experiments are summarized in Table 2 in Appendix. We group them by scope (global, perception-related, and mechanism-specific).

4 Results

We compare CG, WS, SM, and NG over 30 rounds and 10 trials per setting (mean with confidence bands), with robustness analyses in the Appendix.

To make comparisons interpretable, we keep the same initialization, physical dynamics, and observation noise across mechanisms, changing only governance rules. This isolates how each mechanism transforms the same evidence stream into collective trajectories.

We report trends over rounds rather than single-round endpoints because governance quality is intrinsically temporal: practical systems must not only converge eventually, but also avoid costly detours during intermediate rounds when decisions are already consequential. Accordingly, we analyze both final alignment and trajectory shape (overshoot, volatility, and recovery lag).

4.1 System-Level Dynamics

Figure 2 evaluates convergence from a false-majority start and recovery after a temporary misinformation shock.

4.1.1 Speed of Convergence

CG is the only mechanism that reliably overturns the initial 70% false majority. WS and SM reinforce early imbalances through reward-by-progress and popularity feedback, respectively, while NG improves more slowly through independent deliberation.

This difference is not only about final accuracy but also about transient path quality. In CG, support for the true topic rises with fewer oscillations, indicating that credibility updates act as a stabilizer when observations are noisy. By contrast, WS and SM exhibit stronger early lock-in: once the false topic receives initial reinforcement, the same mechanism that is intended to aggregate information instead amplifies its own prior bias.

From a governance perspective, this highlights a practical design criterion: fast adaptation is useful only when the update signal is quality-sensitive. Mechanisms that react quickly to unfiltered momentum can be confidently wrong, whereas CG's slower-but-filtered reallocation improves both direction and stability.

A second criterion is reversibility. In many online systems, temporary ranking errors create downstream lock-in (visibility, funding, participation). CG's trajectory suggests lower lock-in risk: because influence is continuously re-evaluated, early errors are less likely to become permanent structural advantages.

4.1.2 Misinformation Resistance

Under a 5-round parameter swap, all methods dip. CG recovers fastest because credibility accumulates over quality-weighted momentum and anti-bubble terms limit overreaction. WS and SM remain misled after shock amplification, and NG partially rebounds but lacks coordinated correction.

The key asymmetry is memory structure. CG retains information about prior source reliability, so temporary shocks perturb but do not reset influence assignment. WS and SM are more myopic: short-run performance or popularity spikes are rapidly capitalized into future influence, which makes rollback difficult once the shock ends.

This behavior matters for real platforms, where adversarial or exogenous shocks are inevitable. A robust mechanism should degrade gracefully under manipulation windows and recover without external intervention. In our setting, only CG consistently satisfies both requirements.

Notably, the recovery pattern is asymmetric: deterioration during the shock is steeper than post-shock repair for WS/SM, while CG keeps these two slopes closer. This slope symmetry is desirable in practice because it bounds worst-case downtime after misinformation events.

Another useful reading is counterfactual exposure time: the number of rounds in which a mechanism allocates dominant support to the wrong topic. Even when eventual convergence is possible, long error windows can impose large opportunity costs in funding, attention, and experimentation. CG reduces this error-window duration by combining resistance (smaller drawdown during shocks) with faster rebound after shocks.

4.2 Mechanism-Level Behavior

Figure 3 explains these outcomes: CG shifts influence toward consistently reliable agents, improves the social signal for the true topic, and sustains true-topic physical progress.

Panel (a) indicates that influence concentration in CG is selective rather than purely accumulative: agents gain influence when their judgments align with improving topic quality, not merely when they are aligned with the current majority. Panel (b) shows that this produces a cleaner social signal trajectory, with fewer abrupt reversals. Panel (c) connects these social dynamics to material outcomes: once resources are reallocated toward truth-supporting agents, the physical progress of the true topic accelerates and remains dominant.

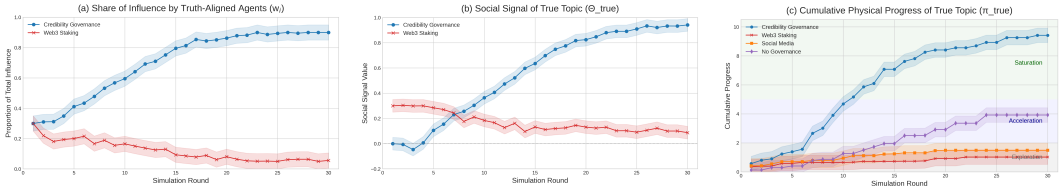
Together, the three panels provide a mechanism-level explanation for system-level recovery. CG does not rely on a single correction event; instead it creates a repeated feedback cycle in which better signal quality improves allocation, and better allocation further improves signal quality.

This also clarifies why no-governance can occasionally outperform weak governance: absent a distortion-amplifying feedback loop, random error may wash out over time. Weak governance, by contrast, can transform transient noise into persistent structural bias.

4.3 Ablation Studies

Figure 4 shows that CG depends on four components: credibility updates, anti-bubble penalties, the early-mover bonus, and using $\Delta\Theta$ rather than $\Delta\pi$ as the reward basis.

The ablations clarify distinct roles. Credibility updates provide the main long-horizon correction channel; anti-bubble terms suppress early cascades; the early-mover bonus accelerates recovery; and replacing $\Delta\Theta$ with raw $\Delta\pi$ reproduces the WS failure mode. Together, these results show that CG's robustness comes from component complementarity rather than any single term.



(a) Influence share of truth supporters

(b) Social signal of true topic

(c) Physical progress of true topic

Figure 3: Mechanism-level behavior.

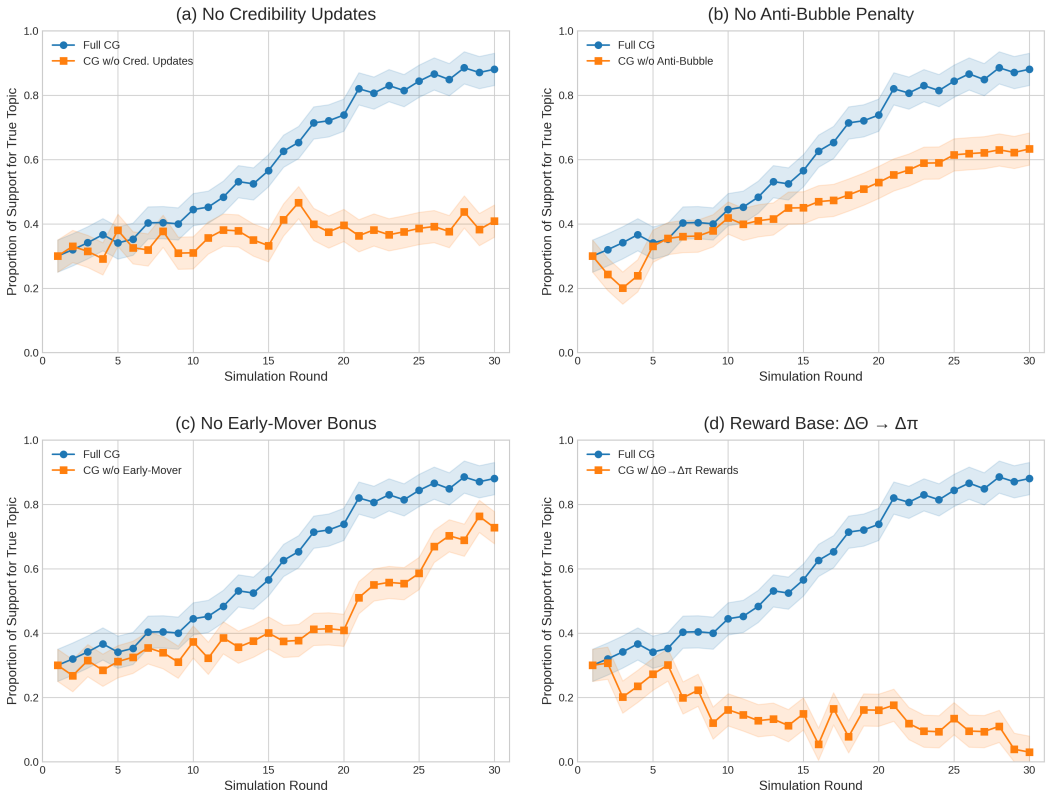


Figure 4: Ablations of Credibility Governance. (a) No credibility updates. (b) No anti-bubble penalty. (c) No early-mover bonus. (d) Reward swaps from $\Delta\Theta_k$ to $\Delta\pi_k$.

5 Discussion and Conclusion

CG improves self-correction by making influence updates quality-sensitive over time rather than tying them to static popularity or capital. In our experiments, this reduces lock-in from early amplification and improves recovery after misinformation shocks.

The mechanism still has limits: coordinated gaming remains possible, and performance depends on signal quality, topic separability, and noise. Future work should test multi-topic settings, richer communication dynamics, and more heterogeneous agents.

Overall, our results show that collective accuracy under weak truth signals depends critically on temporal influence design. By reallocating influence through credibility rather than visibility or capital, CG improves stability, correction speed, and resilience in socio-physical simulations.

References

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.

Eytan Bakshy, Solomon Messing, and Lada A Adamic. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132, 2015.

Sushil Bikhchandani, David Hirshleifer, and Ivo Welch. A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, 1992.

Katie Bratton et al. Modeling emergent social behavior in multi-agent llm systems. In *NeurIPS*, 2023.

Vitalik Buterin. On stake and consensus. Ethereum Blog, 2014.

Zhenyu Chen et al. Agentverse: A flexible framework for multi-agent llm simulation. In *ICLR*, 2024.

Philip Daian et al. Flash boys 2.0: Frontrunning, transaction reordering, and consensus instability. In *IEEE S&P*, 2020.

Michela Del Vicario et al. Echo chambers in the age of misinformation. *PNAS*, 2016.

Joshua Epstein and Robert Axtell. *Growing Artificial Societies*. MIT Press, 1996.

Timothy Frye. Credence goods, signaling, and the link between money and trust. *Comparative Political Studies*, 2021.

Alvin Goldman. *Knowledge in a Social World*. Oxford University Press, 1999.

Benjamin Golub and Matthew O Jackson. Naive learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics*, 2(1):112–149, 2010.

Robin Hanson. Foul play in information markets. *Economic Journal*, 2006.

Buwei He, Yang Liu, Zhaowei Zhang, Zixia Jia, Huijia Wu, Zhaofeng He, Zilong Zheng, and Yipeng Kang. Make an offer they can't refuse: Grounding bayesian persuasion in real-world dialogues without pre-commitment. *arXiv preprint arXiv:2510.13387*, 2025.

Yipeng Kang, Tonghan Wang, and Gerard de Melo. Incorporating pragmatic reasoning communication into emergent language. *Advances in neural information processing systems*, 33:10348–10359, 2020.

Yipeng Kang, Junqi Wang, Yexin Li, Mengmeng Wang, Wenming Tu, Quansen Wang, Hengli Li, Tingjun Wu, Xue Feng, Fangwei Zhong, and Zilong Zheng. Are the values of LLMs structurally aligned with humans? a causal perspective. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23147–23161, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1188. URL https://aclanthology.org/2025.findings-acl.1188/.

Jon Kleinberg and David Easley. Trust in a complex world: A network theory of social capital. *Communications of the ACM*, 2022.

Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. Predicting positive and negative links in online social networks. In *WWW*, 2010.

Hengli Li, Zhaoxin Yu, Qi Shen, Chenxi Li, Mengmeng Wang, Tinglang Wu, Yipeng Kang, Yuxuan Wang, Song-Chun Zhu, Zixia Jia, et al. Beda: Belief estimation as probabilistic constraints for performing strategic dialogue acts. *arXiv preprint arXiv:2512.24885*, 2025.

Jan Lorenz, Heiko Rauhut, Frank Schweitzer, and Dirk Helbing. How social influence can undermine the wisdom of crowd effect. *PNAS*, 108(22):9020–9025, 2011.

Yihuan Mao, Yipeng Kang, Peilun Li, Wei Xu, and Chongjie Zhang. Ibgp: Imperfect byzantine generals problem for zero-shot robustness in communicative multi-agent systems. In *International Conference on Artificial General Intelligence*, pages 421–432. Springer, 2025.

Lev Muchnik, Sinan Aral, and Sean J Taylor. Social influence bias: A randomized experiment. *Science*, 341(6146):647–651, 2013.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22, 2023.

Gordon Pennycook and David Rand. The psychology of fake news. *Trends in Cognitive Sciences*, 2021.

Paul Resnick and Richard Zeckhauser. Reputation systems. *Communications of the ACM*, 2000.

Matthew Salganik, Peter Dodds, and Duncan Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 2006.

Patrick Schramowski et al. Language models as simulated societies. *Scientific Reports*, 2023.

Chandler Smith, Marwa Abdulhai, Manfred Diaz, Marko Tesic, Rakshit S Trivedi, Alexander Sasha Vezhnevets, Lewis Hammond, Jesse Clifton, Minsuk Chang, Edgar A Duéñez-Guzmán, et al. Evaluating generalization capabilities of llm-based agents in mixed-motive scenarios using concordia. *arXiv preprint arXiv:2512.03318*, 2025.

Ethan Squires et al. How do peer-review rating systems influence scientific consensus? *Science Advances*, 2021.

Cass R Sunstein. *Infotopia: How Many Minds Produce Knowledge*. Oxford University Press, 2006.

Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 2018.

Junyu Zhang, Yipeng Kang, Jiong Guo, Jiayu Zhan, and Junqi Wang. Bach-v: Bridging abstract and concrete human-values in large language models. *arXiv preprint arXiv:2601.14007*, 2026.

Zhaowei Zhang, Xiaobo Wang, Minghua Yi, Mengmeng Wang, Fengshuo Bai, Zilong Zheng, Yipeng Kang, and Yaodong Yang. Policon: Evaluating llms on achieving diverse political consensus objectives. *arXiv preprint arXiv:2505.19558*, 2025.

Zhou Ziheng, Huacong Tang, Mingjie Bi, Yipeng Kang, Wanying He, Fang Sun, Yizhou Sun, Ying Nian Wu, Demetri Terzopoulos, and Fangwei Zhong. An llm-based agent simulation approach to study moral evolution. *arXiv preprint arXiv:2509.17703*, 2025.

Zhou Ziheng, Jiakun Ding, Zhaowei Zhang, Ruosen Gao, Yingnian Wu, Demetri Terzopoulos, Yipeng Kang, Fangwei Zhong, and Junqi Wang. Simple role assignment is extraordinarily effective for safety alignment. *arXiv preprint arXiv:2602.00061*, 2026.

Table 1: Key simulation parameters.

Parameter	Symbol	Value	Scope / Description
Physical Progress Dynamics			
$$\Delta\pi_k^t = \begin{cases} \epsilon_k^t, & \pi_k^{t-1} < \pi_1, \\ \left(v_k + \gamma\, r_k^t\right)\left(1 - \frac{\pi_k^{t-1}}{M_k}\right) + \epsilon_k^t, & \pi_1 \le \pi_k^{t-1} < \pi_2, \\ \gamma' r_k^t \left(1 - \frac{\pi_k^{t-1}}{M_k}\right) + \epsilon_k^t, & \pi_k^{t-1} \ge \pi_2. \end{cases}$$			
Environmental noise	σ	0.1	Std. dev. of ϵ_k^t
Baseline velocities	$v_{\text{true}}, v_{\text{false}}$	1.0, 0.8	Intrinsic growth rates
Acceleration factor	γ	0.8	Mid-stage resource impact
Saturation limits	$M_{\text{true}}, M_{\text{false}}$	10, 8	Maximum progress
Late-stage factor	γ'	0.4	Late-stage resource impact
Stage thresholds	π_1, π_2	2.0, 5.0	Phase transition points
Agent Perception Model			
$$\tilde{\pi}_k^t = \pi_k^{t-1} + \rho\, \Theta_k^{t-1} + \xi_k^t$$			
Social contamination	ρ	0.2	Strength of social shaping
Observation noise	ξ_k^t	0.05	Std. dev. of perception noise
Credibility Governance (CG) Specific			
$$w_i^{t-1} = \alpha_i^t e^{\lambda c_i^{t-1}}$$			
Influence concentration	λ	2.0	Exponential influence scaling
$$\Theta_k^t = (1-\lambda_s)\Theta_k^{t-1} + \lambda_s \left[(r_k^t - r_k^{t-1})\, \overline{q}_k^t - \gamma B_k^t \right]$$			
Social-signal smoothing	λ_s	0.4	Smoothing of social signal update
$$c_i^t = c_i^{t-1} + \eta \big(\Theta_{a_i^t}^t - \Theta_{a_i^t}^{t-1}\big)\, e^{-\kappa r_{a_i^t}^t}$$			
Credibility learning rate	η	0.1	Rate of credibility change
Early-mover advantage	κ	0.5	Early-mover bonus factor
$$\overline{q}_k^t = \frac{\sum_{i:a_i^t=k} c_i^{t-1}}{\sum_{i:a_i^t=k} 1}, \quad B_k^t = \sigma\big(\alpha(r_k^t - r_k^{t-1})\big)\,\big(1 - \overline{q}_k^t\big)$$			
Inconsistency penalty	β	0.3	Penalty for switching
Web3 Staking (WS) Specific			
$$w_i^{t-1} = \alpha_i^t \,\mathrm{bal}_i^{t-1}, \quad \mathrm{bal}_i^t = \mathrm{bal}_i^{t-1} + \gamma_s w_i^{t-1}\, \Delta\pi_{a_i^t}^t$$			
Staking reward rate	γ_s	0.1	Proportion of progress paid as stake reward

A Key Parameters

B Parameter Sensitivity Analysis

B.1 Sensitivity to Population Size (N)

Figure 5 shows how population size (N) affects CG's convergence. Larger populations (N=300) converge faster and more smoothly than N=100 and especially N=50. With more

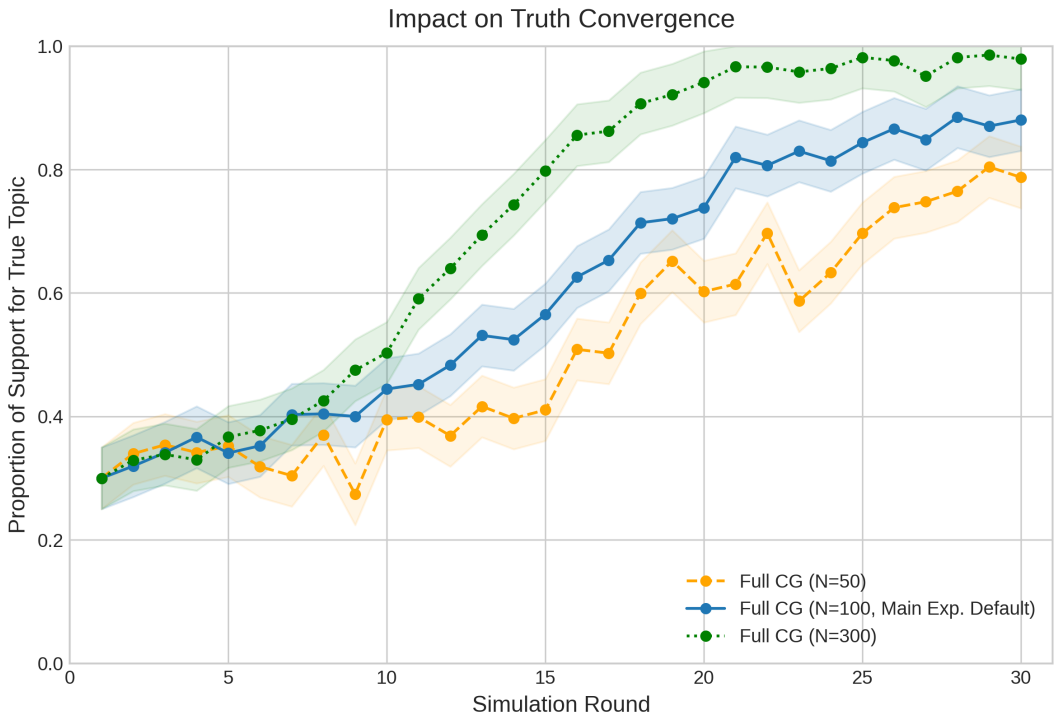


Figure 5: Sensitivity to Population Size. Comparison of CG's convergence for N=50, N=100 (Main Experiment), and N=300.

agents, the social signal produced by the LLM-based population becomes more stable: individual stochasticity in model generations is averaged out, and credibility updates reflect consistent patterns rather than noise. This stronger aggregate signal enables CG to identify reliable agents more quickly and reinforce the correct topic. Smaller populations, by contrast, are more exposed to randomness from individual LLM outputs, making the social signal less stable and slowing convergence.

B.2 Sensitivity to Initial Opinion Distribution

Figure 6 shows how the initial opinion distribution affects CG's convergence. As the share of truth-aligned agents increases (30%, 50%, 70%), convergence becomes faster and more stable. When truth begins as the majority, CG reinforces it almost immediately. This is because the true topic produces more consistent improvement signals in the early rounds, giving truth-aligned agents slightly positive credibility updates from the start. As these small advantages accumulate, the average credibility of truth supporters rises within the first few rounds. Once their mean credibility exceeds that of false supporters, the anti-bubble term for the true topic becomes negligible, since it scales with $(1 - \bar{q}^t_{\text{true}})$. Thus, a truth-majority receives an increase rather than a penalty. Overall, CG not only overcomes unfavorable initial conditions but also consolidates already truth-aligned populations.

B.3 Sensitivity to Baseline Velocity Gap

Figure 7 illustrates Credibility Governance's (CG) sensitivity to the intrinsic quality difference between the true and false topics, quantified by the baseline velocity gap (v_{true} - v_{false}). A larger velocity gap leads to significantly faster and more robust convergence. This is because a more pronounced intrinsic advantage for the true topic means its underlying quality is clearer, allowing CG's mechanisms to more effectively detect and amplify this signal, thereby empowering its supporters more efficiently. Conversely, a smaller velocity gap results in slower, more volatile convergence, potentially reaching a lower final truth support. This occurs because a subtle quality difference makes the true topic's advantage harder for CG to discern, as it is more easily masked by noise, the initial resource advantage of the

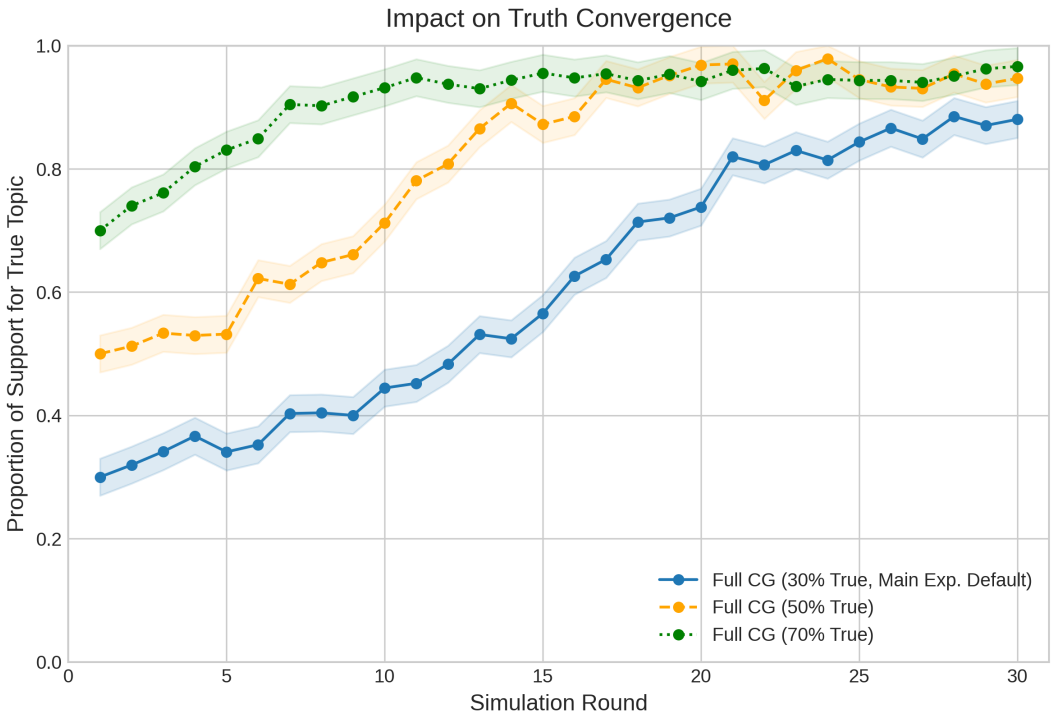


Figure 6: Sensitivity to Initial Opinion Distribution. Comparison of CG's convergence for 30%, 50%, and 70% initial truth-aligned agents.

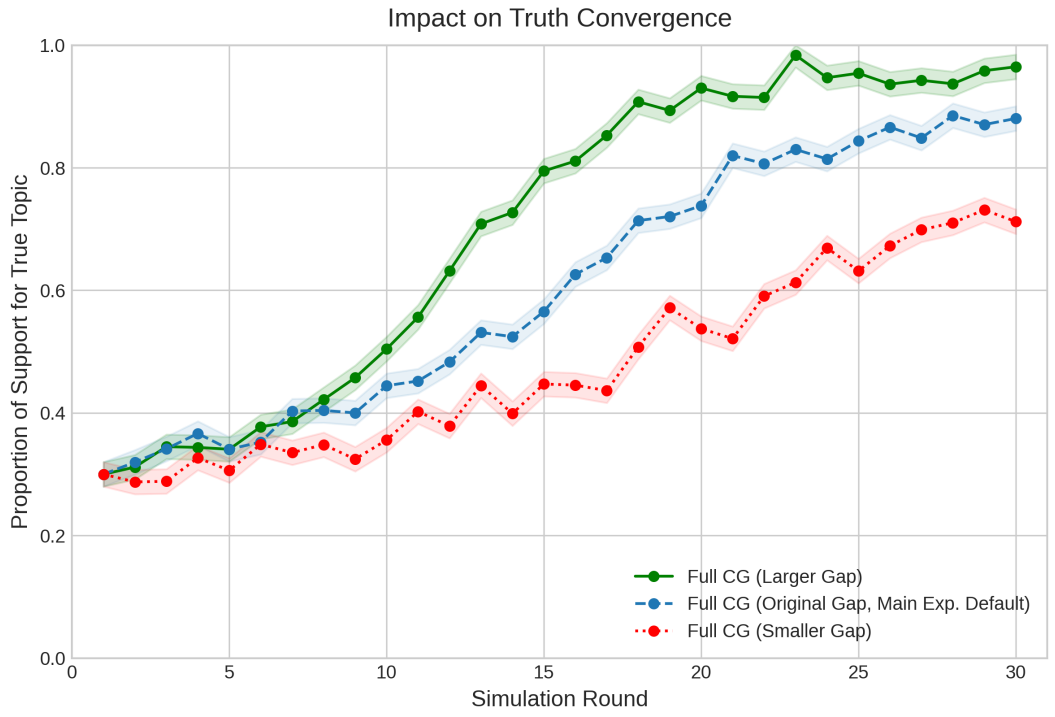


Figure 7: Sensitivity to Baseline Velocity Gap. Comparison of CG's convergence for smaller, original, and larger intrinsic growth differentials.

false topic, and individual agent stochasticity. This highlights that while CG is effective, the inherent distinguishability of truth still fundamentally influences its convergence dynamics.