# Belief-Based Behavioral Diffusion and the Limits of Incentive Design

## *Interpretability as a Structural Constraint*

Lydia Li, Simona Fabrizi and Steffen Lippert

The University of Auckland
Department of Economics & Centre for Mathematical Social Science

29th Annual SIOE Conference
UNSW, Sydney, 24–26 August 2025

Waipapa Taumata Rau University of Auckland

# Background

Digital platforms increasingly serve as the infrastructure through which individuals encounter, interpret, and respond to social information. In this context, behavior is not simply reactive—it reflects evolving internal beliefs shaped by social signals, peer actions, and platform incentives. Despite the complexity of online discourse, user behavior is structured around a limited repertoire of expressive actions—replying, quoting, retweeting, sharing, reporting, blocking.

**Traditional models of information diffusion often neglect the cognitive bottlenecks of real-world environments: users do not observe beliefs directly, but infer them from behavior—an inference process subject to distortion, ambiguity, and feedback.**

# Motivation

- How do agents behave when they infer from peer actions rather than observe the truth?
- What if actions reflect both belief and incentive—introducing ambiguity?
- Can platforms still steer behavior under filtered or misaligned inference?
- What structural conditions restore interpretability of behavioral signals?
- How to design incentive paths that prevent lock-in under distortion?

# Closely Related Literature

- **Threshold and cascade models:** Adoption spreads via peer exposure (Granovetter (1978); Watts (2002); Acemoglu et al. (2011)), but assume transparent inference.
- **Discrete choice, softmax, social learning:** Account for probabilistic actions and belief heterogeneity (McFadden (1974); Train (2009); Goeree et al. (2016); DeGroot (1974)), yet presume signals are interpretable.
- **Filtering and inference bottlenecks:** Recent advances highlight how endogenous filtering can disrupt collective learning (Bénabou and Tirole (2016); Duffie et al. (2017); Miller (2019)).
- **Latest advances:** Acemoglu et al. (2024) show that platform interventions may collapse under feedback-driven misinterpretation, demanding new mechanism design paradigms.

→ How to systematically incorporate interpretability as a structural constraint in dynamic incentive design remains an open question.

Waipapa Taumata Rau University of Auckland

# Our Contribution

- **Unified framework:** We propose a belief-driven behavioral diffusion model with endogenous filtering, making interpretability an explicit design constraint.
- **Theoretical advance:** We formalize the collapse of incentive effectiveness and show how filtering-induced irreversibility locks in misinformation.
- **Mechanism innovation:** We design, analyze, and simulate filtering-aware interventions—compression, anchoring, and recursive repair—that restore robust system control even under epistemic fragility.

# Some Preliminaries

**Utility Function:**

$$U_i(B) = Q_i^t \cdot B + PI^t \cdot S_i^t + \epsilon_i,$$

**Belief Update via Social Learning:**

$$\theta_i^{t+1} = \theta_i^t + \lambda \left( \frac{1}{|N_i|} \sum_{j \in N_i} B_j^t - \theta_i^t \right)$$

- $Q_i^t = \alpha + \beta\theta_i^t + \epsilon_i'$ represents the agent's internal assessment of the information, determined by a belief-dependent signal $\theta_i^t$, a baseline bias $\alpha$, belief sensitivity $\beta$, and idiosyncratic heterogeneity $\epsilon_i'$.
- $n$ agents on a network, each with belief $\theta_i^t \in [0, 1]$ at time $t$.
- $B_j^t \in \{-1, 0, 1\}$ is the behavior of neighbor $j$ at time $t$, indicating rejection, neutrality, or support;
- Platform incentive $PI^t$ and alignment label $S_i^t \in \{-1, 1\}$.
- $N_i$ is the set of agent $i$'s neighbors in the network;
- $\lambda \in (0, 1)$ is a learning rate reflecting responsiveness to social signals.

# Latent Utility and Softmax Choice Rule

The latent utility of agent $i$ taking action $b \in \{-1, 0, 1\}$ at time $t$ is given by:

$$U_i^t(b) = Q_i^t \cdot b + PI_t S_i^t \cdot b + \varepsilon_i(b)$$

Here, **the action variable $b$ plays a directional role**: it encodes the polarity and intensity of behavioral choice.
Since both $\varepsilon_i' \cdot b$ and $\varepsilon_i(b)$ represent unobserved heterogeneity across actions, we aggregate them into a single Gumbel-distributed shock for tractability:

$$U_i^t(b) = (\alpha + \beta\theta_i^t) \cdot b + PI_t S_i^t \cdot b + \tilde{\varepsilon}_i(b),$$

with $\tilde{\varepsilon}_i(b) = \varepsilon_i' \cdot b + \varepsilon_i(b)$.
This leads to the **softmax (multinomial logit)** choice rule:

$$P(B_i^t = b) = \frac{\exp(U_i^t(b))}{\sum_{b' \in \{-1,0,1\}} \exp(U_i^t(b'))}$$

Waipapa Taumata Rau University of Auckland

# Platform Problem

**Platform Objective and Error Metrics**

The platform's central concern is to avoid amplifying false information. We define:

$$\mathrm{FPR}_t = \frac{\sum_i \mathbb{I}[\text{info is false} \wedge B_i^t = 1]}{\sum_i \mathbb{I}[B_i^t = 1]}$$

$$\mathrm{FNR}_t = \frac{\sum_i \mathbb{I}[\text{info is true} \wedge B_i^t = -1]}{\sum_i \mathbb{I}[B_i^t = -1]}$$

- False Positive (Type I error): Supporting false content.
- False Negative (Type II error): Rejecting true content.

**Legitimacy Guarantee:**

- Incentive Compatibility (IC): Agents should find it utility-maximizing to act in accordance with their beliefs.
- Individual Rationality (IR): Expected utility for participation must meet a minimum threshold: $\mathbb{E}[U_i(B_i^t)] \geq \underline{U}_i$.

# Some Results

## Theorem 1: Belief Monotonicity and Comparative Statics

Let $B_i^t \in \{-1, 0, 1\}$ be chosen by agent $i$ via softmax rule with utility $U_i^t(b) = (\alpha + \beta\theta_i^t)b + PI^t S_i^t b$. Then, for $\beta > 0$:

$$\frac{\partial P(B_i^t = 1)}{\partial \theta_i^t} > 0, \qquad \frac{\partial P(B_i^t = -1)}{\partial \theta_i^t} < 0$$

The effect on neutrality satisfies: $\frac{\partial P(B_i^t = 0)}{\partial \theta_i^t}$ may be positive or negative, depending on the distribution of actions.

**Interpretation:**

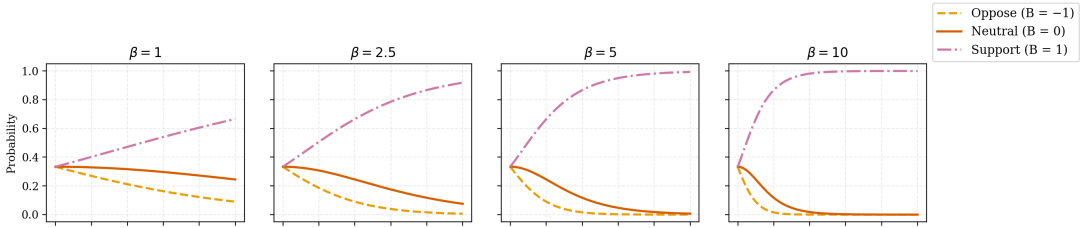
- Increasing belief strictly raises the likelihood of support, reduces rejection, and reshapes the neutrality region.
- This micro-level monotonicity is the analytic foundation for using observed actions as interpretable signals of latent belief in all subsequent inference and mechanism design.

# Some Results (Cont'd)

**Theorem 2: Degeneration of Neutral Behavior**

Let $B_i^t \in \{-1, 0, 1\}$ follow the softmax rule with utility $U_i^t(b)$. Then for any $\theta_i^t \neq \theta^*$, where $\theta^*$ is the belief value at which $U_i^t(1) = U_i^t(-1)$, $\lim_{\beta \to \infty} P(B_i^t = 0) = 0$.

- As belief sensitivity $\beta$ increases, the probability mass for neutral actions is compressed into an increasingly narrow belief interval around $\theta^*$.
- Outside this critical interval, agents overwhelmingly choose polarized actions, resulting in the disappearance of robust neutral zones and amplifying system fragility.



# Some Results (Cont'd)

**Theorem 3: Incentive Substitution with Behavioral Alignment**

Let $b^* \in \{-1, 1\}$, $S_i^t = \text{sign}(b^*)$. For any $\theta_i^t \in [0, 1]$, there exists $PI^*(b^*) > 0$ such that if $PI^t > PI^*(b^*)$,

$$P(B_i^t = b^*) > \delta \quad (\delta > 0).$$

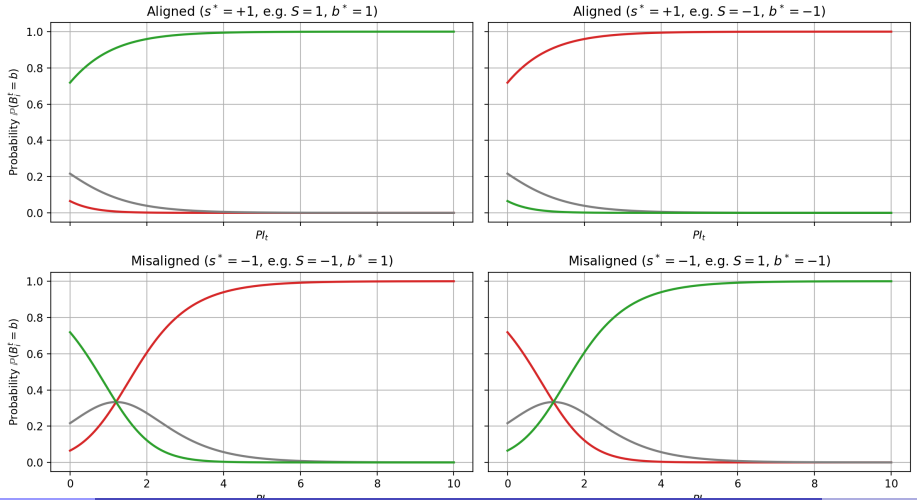
**Theorem 4: Directional Deterrence for Misaligned Incentive**

Suppose $S_i^t = -\text{sign}(b^*)$. Then $\exists\, PI^\dagger > 0$, for any $\varepsilon > 0$, if $PI^t > PI^\dagger$,

$$P(B_i^t = b^*) < \varepsilon.$$

- By adjusting the sign and magnitude of incentives, the platform can selectively amplify aligned behaviors or suppress misaligned ones for any belief profile.
- This directional control mechanism highlights the capacity—and potential cost—of overriding belief heterogeneity through strong intervention, laying the foundation for subsequent analysis of interpretability and system robustness.

# Figure: Action Distributions under Varying Incentives



# Behavioral Cascades

## Theorem 5: Misinformation-Induced Behavioral Cascades

Suppose the true state is *false*. Let agent beliefs $\theta_i^t \sim F(\theta)$ be i.i.d. across agents, and define the critical belief threshold $\underline{\theta}$ as the minimal belief at which rejection becomes utility-optimal:

$$\underline{\theta} := \inf \{\theta \in [0,1] \mid U_i(-1) > \max\{U_i(0), U_i(1)\}\}.$$

Assume the average belief $\mathbb{E}[\theta_i^t] < \underline{\theta}$, and let

$$\underline{PI} := \inf \left\{ PI^t \;\middle|\; \mathbb{P}(B_i^t = -1) > \delta \right\}$$

be the minimal incentive to sustain consistent rejection.

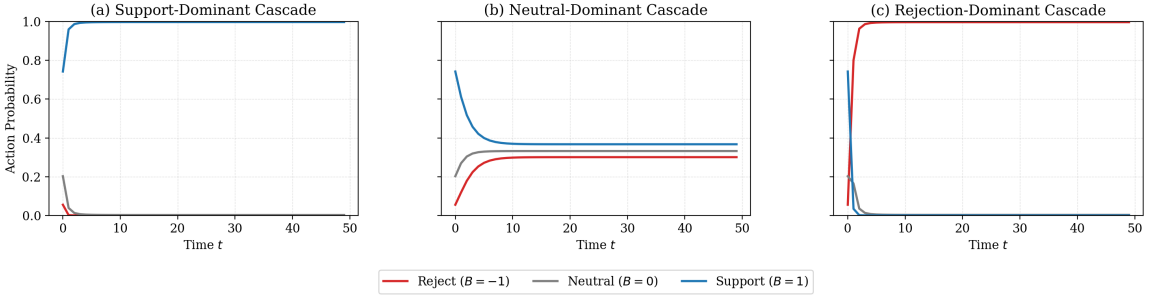
If $PI^t < \underline{PI}$ and agents update beliefs via observed neighbor behavior, then the system may exhibit locally reinforced behavioral cascades.

# Behavioral Cascades (Cont'd)

(a) **Support-dominant:** Early supportive behavior boosts peer signals, raises beliefs, and reinforces pro-false cascades—even when the true state is negative.

(b) **Neutral-dominant:** Ambiguous actions and weak incentives sustain indecision, resulting in prolonged stagnation and inaction.

(c) **Rejection-dominant:** Low initial beliefs and visible opposition reduce confidence, triggering collective rejection.



# Corollary: Path Dependence Near Critical Incentives

## Directional Sensitivity to Incentive Level

Let $\underline{PI}$ be the minimum incentive to trigger rejection dominance. When initial beliefs $\theta_i^0 < \underline{\theta}$, long-run system behavior depends sharply on how $PI^t$ compares to $\underline{PI}$:

(i) $PI^t < \underline{PI}$: Support dominates via rising belief.

(ii) $PI^t \approx \underline{PI}$: Neutrality and stagnation persist.

(iii) $PI^t > \underline{PI}$: Rejection dominates as belief decays.

*Small changes in incentive near the threshold can lead to dramatically different outcomes—highlighting strong path dependence and the critical importance of incentive calibration.*

Waipapa Taumata Rau University of Auckland

# Bang-Bang Strategy and Equilibrium

## Theorem 6: Existence of Nash Equilibrium

In the belief-driven threshold utility model with discrete actions ($B_i^t \in \{-1, 0, 1\}$) and continuous, belief-dependent softmax utilities, a mixed-strategy Nash equilibrium exists for any finite agent network with bounded heterogeneity.

## Definition: Bang-Bang Incentive Strategy

Assume the platform only observes the average belief $\bar{\theta}^t = \frac{1}{n}\sum_i \theta_i^t$. Then the optimal incentive policy reduces to a threshold rule:

$$PI_t^* = \begin{cases} PI_{\max}, & \bar{\theta}^t < \bar{\theta}^* \\ 0, & \bar{\theta}^t \geq \bar{\theta}^* \end{cases} \quad \text{for some } \bar{\theta}^* \in (0, 1).$$

*Interpretation:* The platform intervenes only when average belief falls below a threshold, yielding a clear and interpretable control boundary.

# Belief Lock-In

## Theorem 7: Belief Lock-In under Network Reinforcement

Let the average belief evolve as

$$\bar{\theta}^{t+1} = (1-\lambda)\bar{\theta}^t + \lambda \cdot \rho(PI^t, \bar{\theta}^t),$$

where $\rho(PI^t, \bar{\theta}^t)$ is continuously differentiable in $PI^t$. **Let $\hat{\theta}$ be the unique fixed point under maximal incentive $PI_{\max}$:**

$$\hat{\theta} = (1-\lambda)\hat{\theta} + \lambda \cdot \rho(PI_{\max}, \hat{\theta}).$$

Then for any $\epsilon > 0$, if $\bar{\theta}^t \in [\hat{\theta} - \epsilon, \hat{\theta} + \epsilon]$,

$$\frac{\partial \bar{\theta}^{t+1}}{\partial PI^t} = \lambda \frac{\partial \rho(PI^t, \bar{\theta}^t)}{\partial PI^t} \to 0 \quad \text{as } \bar{\theta}^t \to \hat{\theta},$$

i.e., incentives lose marginal effect near belief lock-in.

Incentives lose effect near lock-in, causing structural irreversibility.

# Optimal Incentive Path

## Theorem 8: Structure of the Optimal Incentive Path

The platform's discounted objective $[\mathcal{L}(PI^t, \bar{\theta}^t) := \mathrm{FPR}_{\mathrm{exp}}(PI^t, \bar{\theta}^t) + \eta\, \mathrm{FNR}_{\mathrm{exp}}(PI^t, \bar{\theta}^t)$, with $c > 0$ marginal cost of increasing incentive intensity.]

$$\min_{\{PI^t\}} J = \sum_{t=0}^{T} \gamma^t \left[\mathcal{L}(PI^t, \bar{\theta}^t) + c\, PI_t^2\right], \quad \text{s.t.} \quad \bar{\theta}^{t+1} = (1-\lambda)\bar{\theta}^t + \lambda \cdot \rho(PI^t, \bar{\theta}^t).$$

**The optimal incentive path $\{PI^{t*}\}$ satisfies:**

- Starts positive if initial belief is low: $PI_0^* > 0$;
- Is weakly decreasing over time;
- Vanishes as beliefs approach the target: $PI_t^* \to 0$ as $\bar{\theta}^t \to \hat{\theta}$.

*Strong early intervention should taper as belief stabilizes, but pure smooth paths risk fragility near thresholds—motivating more responsive, trigger-based designs.*

# Filtering-Based Update Rule and Filtering Barrier

Agent $i$ observes peer $j$'s action $B_j^t$ and infers latent belief via

$$\hat{\theta}_{ij}^t = \phi_i(B_j^t, PI^t, S_j^t),$$

then updates her own belief by

$$\theta_i^{t+1} = \theta_i^t + \lambda \left( \frac{1}{|N_i|} \sum_{j \in N_i} \hat{\theta}_{ij}^t - \theta_i^t \right).$$

**Filtering Barrier:** For agent $i$ at time $t$,

$$\Delta_i^t := \mathbb{E}\left[ \left| \hat{\theta}_{ij}^t - \theta_j^t \right| \mid B_j^t \right],$$

measuring the expected distortion between inferred and true belief. High $\Delta_i^t$ erodes **interpretability**, breaking the link between observed behavior and actual belief.

# Filtering-Induced Irreversibility

## Theorem 9: Filtering-Induced Irreversibility

Let beliefs evolve as

$$\theta_i^{t+1} = \theta_i^t + \lambda \left( \frac{1}{|N_i|} \sum_{j \in N_i} \phi_i(B_j^t, PI^t, S_j^t) - \theta_i^t \right).$$

Suppose that for all $i$ in some subset $\mathcal{G}$, the filtering barrier $\Delta_i^t > \delta$ for all $t \in [t_0, t_0 + k]$. Then:

$$\left| \frac{\partial \mathbb{E}[\theta_i^t]}{\partial PI^t} \right| \to 0 \quad \text{as } t \to t_0 + k.$$

Once distortion exceeds a structural threshold, agents' beliefs no longer respond to incentives—**platform control collapses**.

# Total Control Collapse

## Theorem 10: Filtering Expansion Causes Incentive Irrelevance

Suppose filtering barriers evolve as:

$$\Delta_i^{t+1} = \Delta_i^t + \kappa \cdot \frac{1}{|N_i|} \sum_{j \in N_i} \Delta_j^t + \gamma \cdot \frac{1}{|N_i|} \sum_{j \in N_i} \left| \hat{\theta}_{ij}^t - \theta_j^t \right|,$$

with $\Delta_i^{t+1} - \Delta_i^t \geq \gamma > 0$ for all $i$ and $t > t_0$.
Define the Effective Incentive Reach (EIR) as:

$$\text{EIR}_t := \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left( \left| \frac{\partial \theta_i^t}{\partial PI^t} \right| > \epsilon \right).$$

Then there exists $t^* < T$ such that for all $t > t^*$: $\text{EIR}_t = 0$ (for any $\varepsilon \in (0, 1)$).

As filtering distortion compounds, incentive effectiveness vanishes system-wide—**no amount of intervention can steer beliefs**. This marks a point of total control collapse.

# Filtering-Aware Mechanism Design

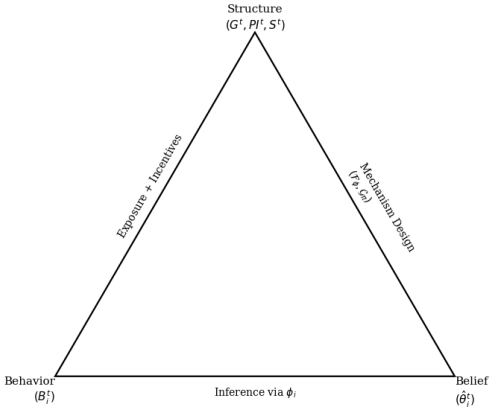



Fig. 4. Mechanism Design under Structure–Behavior–Belief Triangle Coherence
Design Constraint: $\Delta_i^t < \delta$ ensures triangle coherence

**Platform Objective:** Control belief while preserving interpretability.

$$\max_{\{PI^t, G^t\}} U(\theta^T) \quad \text{s.t. } \Delta_i^t < \delta, \ \forall i, t$$

Filtering-Aware Design: Structure–Behavior–Belief Triangle Coherence

The triangle $(G^t, B_j^t, \phi_i)$ is *coherent* if:

$$\mathbb{E}\left[\left|\phi_i(B_j^t) - \theta_j^t\right| \mid G^t\right] < \delta \quad \forall i, j, t.$$

Filtering-aware mechanisms must actively preserve epistemic viability—keeping signals interpretable throughout the feedback loop.

Waipapa Taumata Rau University of Auckland

# Filtering-Aware Mechanism Design (Cont'd)

**Goal:** Preserve **epistemic viability**—i.e., keep behavioral signals interpretable.

**Design Objective:** $\max_{\{Pl^t,\, G^t,\, \mathcal{I}^t\}} U(\theta^T) \quad \text{s.t.} \quad \mathbb{E}\left[\left|\phi_i(B_j^t) - \theta_j^t\right|\right] < \delta$

**Triangle Failure Channels:**

- **Structure → Behavior:** *Compression* – reduce action entropy via quantized incentives.
- **Behavior → Belief:** *Anchoring* – inject labeled agents or interpretive tags.
- **Belief → Structure:** *Recursive Repair* – periodic feedback or synchronization.

Interpretability is not given—it must be constructed and sustained.

**Insight:** Mechanism design becomes a structural challenge: preserving coherence across the *structure–behavior–belief triangle* to ensure control under distortion.

# EIR and Distortion: Filtering-Aware vs Naive

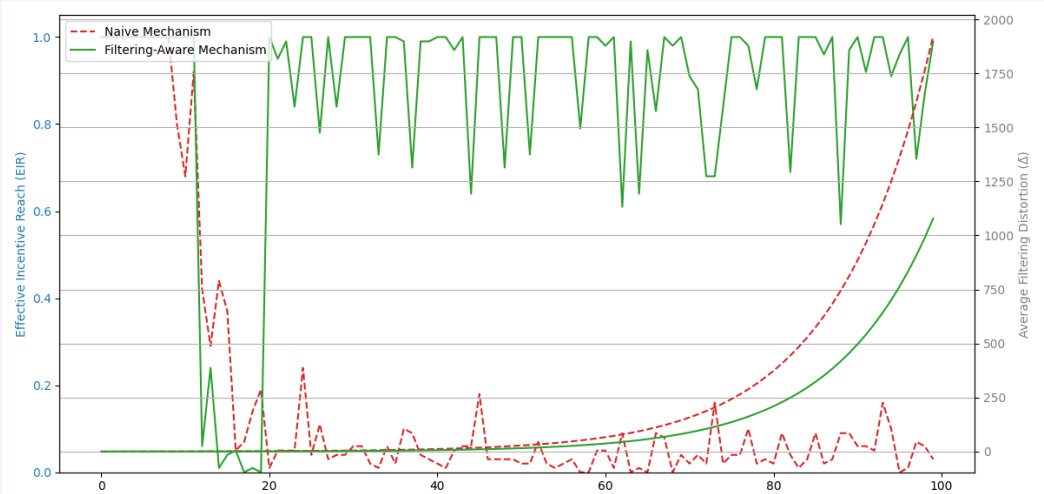



Fig.5. Filtering-Aware vs Naive Mechanism

# Belief Trajectories (Heatmap) Comparison

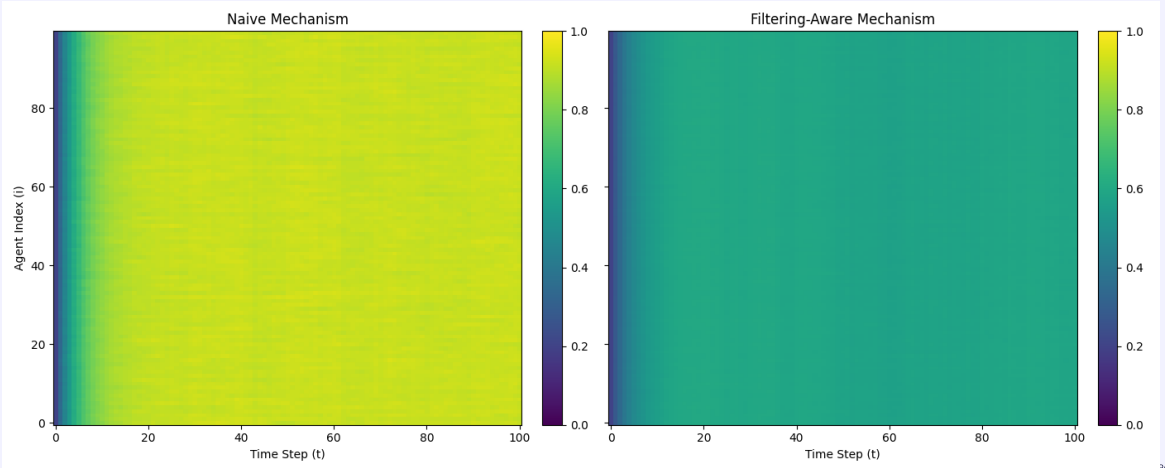
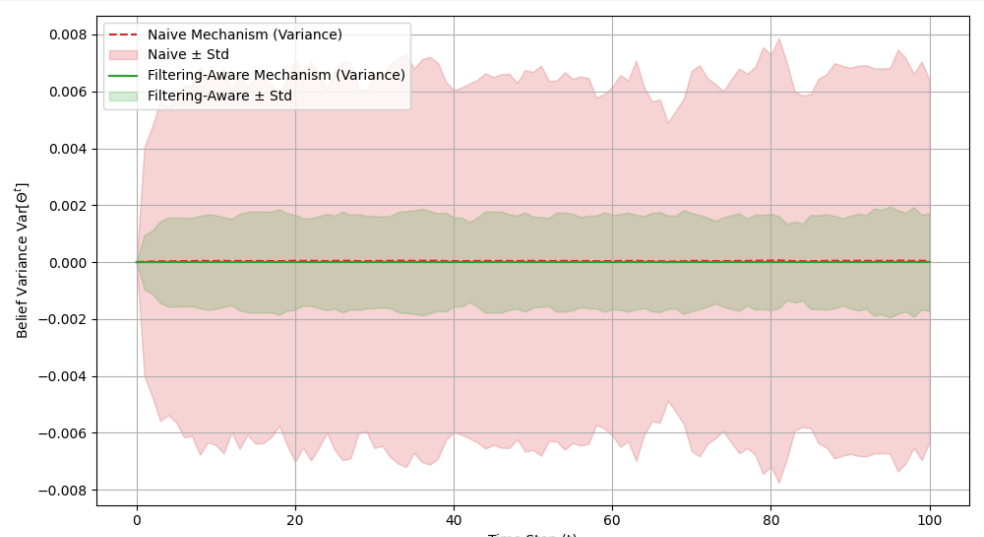



Figure 6: Belief Trajectories (θ) under Naive vs Filtering-Aware Mechanisms

# Evolution of Belief Variance



# Bang-Bang vs. Hybrid Mechanism Performance

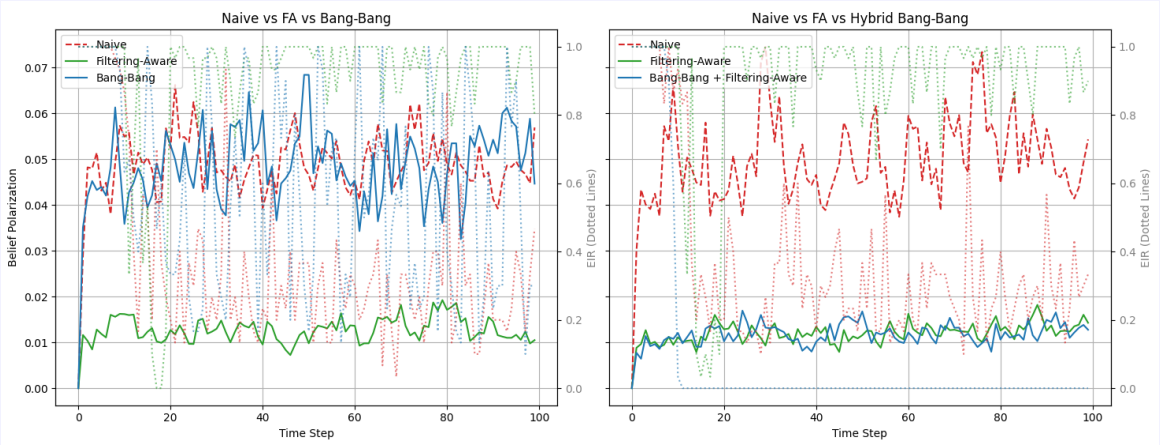



Fig.8. Comparison of EIR (dotted) and Polarization under Two Bang-Bang Variants

# Comparative EIR and Distortion across Six Mechanisms

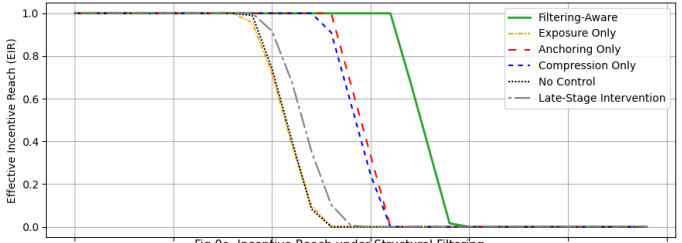



Fig.9a. Incentive Reach under Structural Filtering

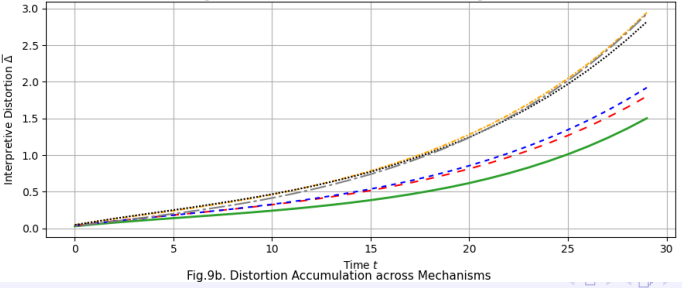



Fig.9b. Distortion Accumulation across Mechanisms

# Network Snapshots: Six Mechanism Interventions


Fig.10. Misinformation Diffusion & Repair under Six Mechanisms | Snapshots: $t = 0$, $t = 8$, $t = 20$, $t = 38$

Waipapa Taumata Rau University of Auckland

# Conclusion

- **Core insight:** Incentives only work if behavior remains interpretable—**interpretability is a structural constraint**.
- **Filtering collapses control:** When inference is distorted, belief-action coherence dissolves, and incentives fail.
- **Our solution:** We propose **filtering-aware mechanisms**—compression, anchoring, and recursive repair—to preserve *triangle coherence* between structure, behavior, and belief.

# Implications & Future Work

- **Governance $\neq$ Incentive tuning:** Platform control is a problem of cognitive infrastructure design.
- **Beyond smooth strategies:** Bang-bang fails near belief thresholds; robust control needs responsive, interpretable interventions.
- **Future directions:**
  - Endogenize filtering and belief heterogeneity.
  - Extend to dynamic networks and hybrid incentive schemes.
  - Explore negative incentives and strategic reversibility.