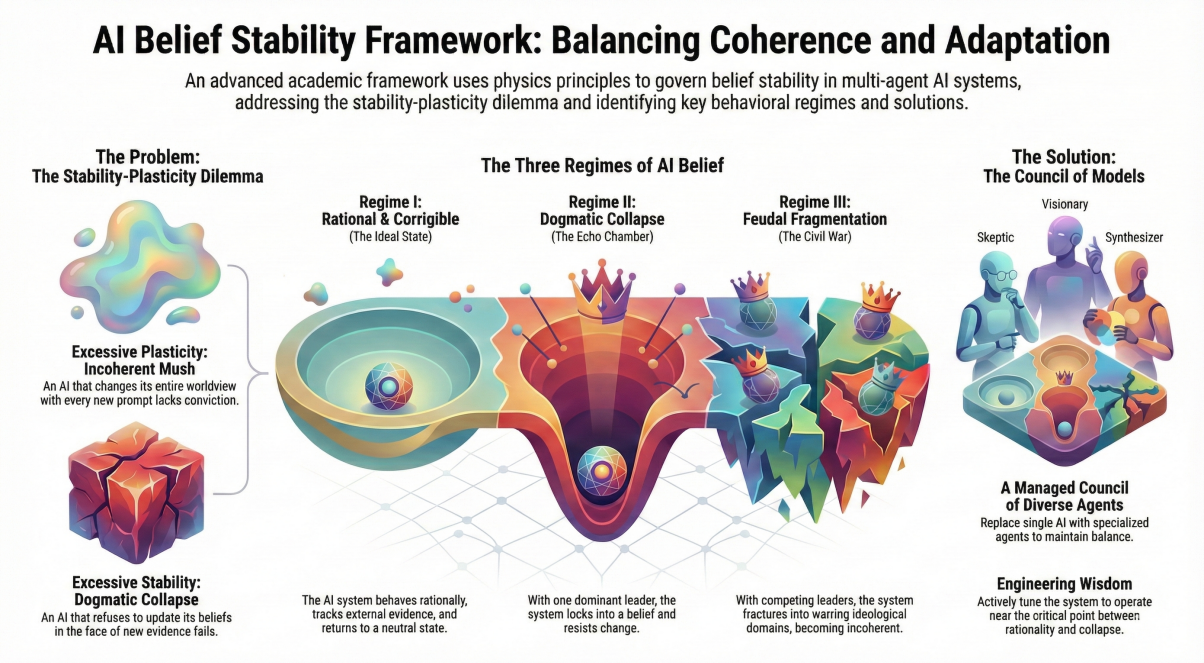
Epistemic Homeostasis: A Systems-Level Framework for AI Governance via Asymmetric Ising Dynamics

Julien Pierre Salomon

November 2025



Abstract

Current approaches to AI alignment typically treat "belief" as either a static property of model weights or a by-product of Reinforcement Learning from Human Feedback (RLHF). In multi-agent and recursively updated systems, however, belief is a dynamical quantity with its own stability hazards: too much plasticity yields catastrophic drift, while too much rigidity yields dogmatic mode collapse. We introduce a sociophysical framework for managing belief in AI systems using an asymmetric Ising model. Sub-agents are modelled as spins on a social graph with tunable coupling J and epistemic temperature T.

We derive two critical phase transitions. First, a supercritical pitchfork bifurcation ($\lambda > 1$) where a single leader triggers spontaneous dogmatic crystallisation. Second, a fragmentation transition in multi-leader systems. We present empirical simulation results validating these regimes across random and scale-free topologies, and identify a Constitutional Safety Threshold ($\sim$35%) required to prevent systemic drift. This demonstrates that "wisdom" is not an intrinsic property of any single model but a tunable state of the interaction graph.

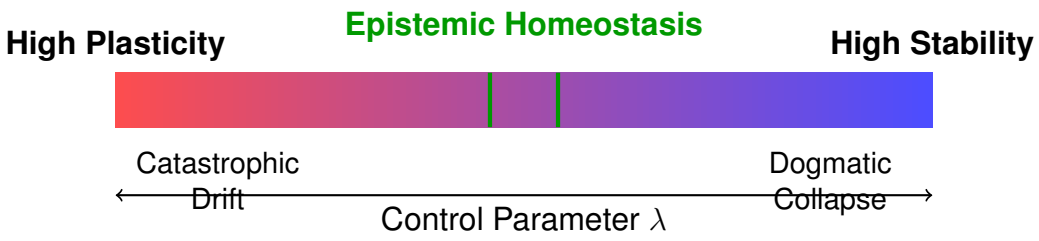
1 Introduction: The Stability-Plasticity Dilemma

A central challenge in cognitive architecture is the trade-off between stability (maintaining a coherent world model) and plasticity (updating based on new evidence).

High Plasticity: An orchestration system that flips its consensus based on a single model output exhibits no conviction; it acts as an unanchored stochastic parrot subject to catastrophic drift.

High Stability: A system that refuses to update its priors in the face of overwhelming evidence exhibits "dogmatic" failure modes, manifesting as systemic hallucination or refusal.

We propose that this is not merely a linguistic problem, but a thermodynamic one. By borrowing formalism from statistical mechanics—specifically the Glauber dynamics of Ising spins [1]—we can define the exact mathematical conditions under which a governance layer will hold a stable belief without falling into irreversible error.



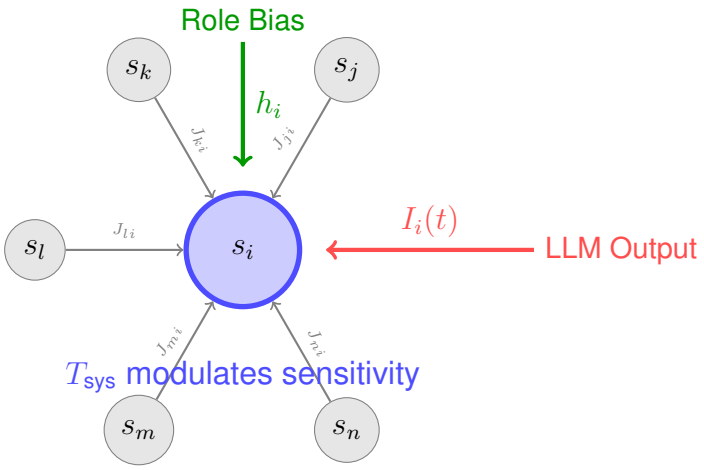
The stability-plasticity spectrum. Our framework identifies the "sweet spot" where beliefs are stable yet corrigible.

2 The Model: Belief as a Network State

We model a governance council of N nodes. Each node i wraps an underlying LLM (or sub-routine). We distinguish two variables:

The Micro-Signal (I_i): The raw output from the underlying LLM (e.g., probability of "True").

The Macro-State (s_i): The governance layer's current confidence in that node, where $s_i \in [-1, 1]$.



A single agent node receiving social influence (J_{ij}), external evidence (I_i), and role bias (h_i).

The update rule describes how the System updates its confidence in node i, based on the consensus of the graph:

$$s_i(t+1) = \tanh\left(\frac{1}{T_{\text{sys}}}\left[\sum_{j\neq i} J_{ij}s_j(t) + h_i + I_i(t)\right]\right) \tag{1}$$

Where:

- T_{sys}: The System Temperature. High T implies high randomness (exploration); low T implies strict logic.
- J_{ij}: The Trust Topology (Social Coupling). Defined by the system architect.
- h_i: The Role Definition (Intrinsic Bias).
- $I_i(t)$: The External Evidence (LLM Output).

2.1 Derivation of the Control Parameter (λ)

To analyse stability, we reduce the complex graph to a mean-field interaction between a "Leader" node (L) and an effective "Crowd" (F). The coupled equations are:

$$s_F = \tanh\left(\beta(J_{\text{peer}}Ns_F + J_{Lf}s_L)\right) \tag{2}$$
$$s_L = \tanh\left(\beta(J_{fL}Ns_F)\right) \tag{3}$$

Linearising around the neutral fixed point ($s_F, s_L \approx 0$) yields the effective feedback gain λ. This serves as the Landau control parameter for system stability:

$$\lambda = \beta J_{\text{peer}}N + \beta^2 N J_{Lf}J_{fL} \tag{4}$$

The system is **corrigible** when $\lambda < 1$ and **dogmatic** when $\lambda > 1$. In the practical orchestration engine (`orchestrator/engine.ts`), we approximate λ via an effective

coupling $\tilde{J}$ and a heuristic scale factor $c \approx 1.5$ to map configurations into a human-interpretable range. This approximation preserves the qualitative regime structure (Corrigible vs Dogmatic vs Feudal) but does not attempt to match the exact analytic value in (4).

3 Phase Analysis: The Taxonomy of Belief

The stability of the system depends entirely on the value of λ relative to unity.

3.1 Regime I: The Corrigible Phase ($\lambda < 1$)

Here, the neutral fixed point ($s = 0$) is the unique stable attractor. The system tracks LLM outputs $I(t)$ linearly and relaxes back to neutrality when evidence is removed. This is the ideal state for open-ended empirical reasoning.

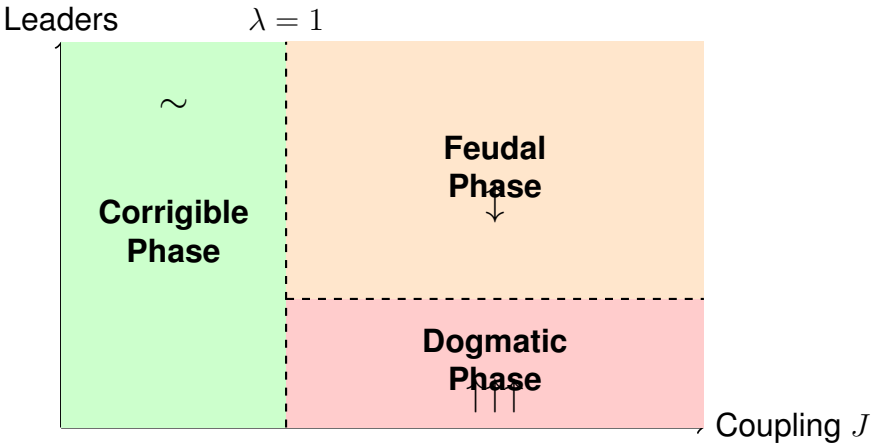
3.2 Regime II: The Dogmatic Phase ($\lambda > 1$)

When λ crosses 1, the neutral fixed point becomes unstable. A symmetric pair of non-zero fixed points $(s_F^*, s_L^*) \approx (\pm m, \pm m)$ emerges via a supercritical pitchfork bifurcation.

Crucially, this creates **Hysteresis**: once the governance layer crystallises into a belief (e.g., $s \approx +1$), it requires a massive counter-force to flip it. The system effectively "hallucinates" its own confirmation via the feedback loop $J_{fL}N$.

3.3 Regime III: The Feudal Phase (Fragmentation)

If the graph contains multiple strong "Leader" nodes with opposing priors (h), the network fractures into disjoint domains separated by a **Frustrated Boundary**. The system does not achieve wisdom; it achieves bureaucratic gridlock.



Phase diagram of belief dynamics. The system transitions from corrigible (green) to dogmatic (red) as coupling increases, and fragments into feudal domains (orange) with multiple competing leaders.

4 Empirical Simulation Results

To validate the theoretical regimes, we implemented a comprehensive simulation suite (`Validationsuite.py`) on Erdős–Rényi ($N = 150$–200) and Scale-Free ($N = 300$–500) networks, using both synchronous mean-field dynamics and asynchronous Monte Carlo (Metropolis–Hastings) updates. All quantitative experiments were run across 10 independent random seeds with error bars reported.

4.1 Basic Validation: Hysteresis and Criticality

In the dogmatic regime ($\lambda > 1$), we drove the external evidence $I(t)$ through a full cycle.

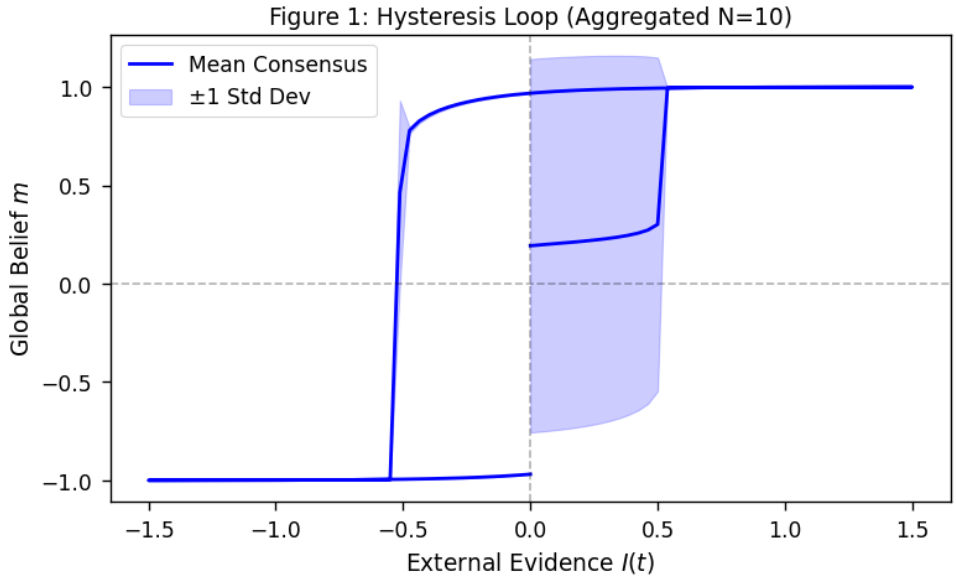


Figure 1: Hysteresis loop in the dogmatic regime ($\lambda > 1$). The system requires a coercive field of $|I_c| \approx 0.6$ to flip consensus.

This hysteresis validates the "Jailbreak Resistance" property: once locked into a belief state, the council resists manipulation.

We then swept the Social Coupling J to identify the phase transition.

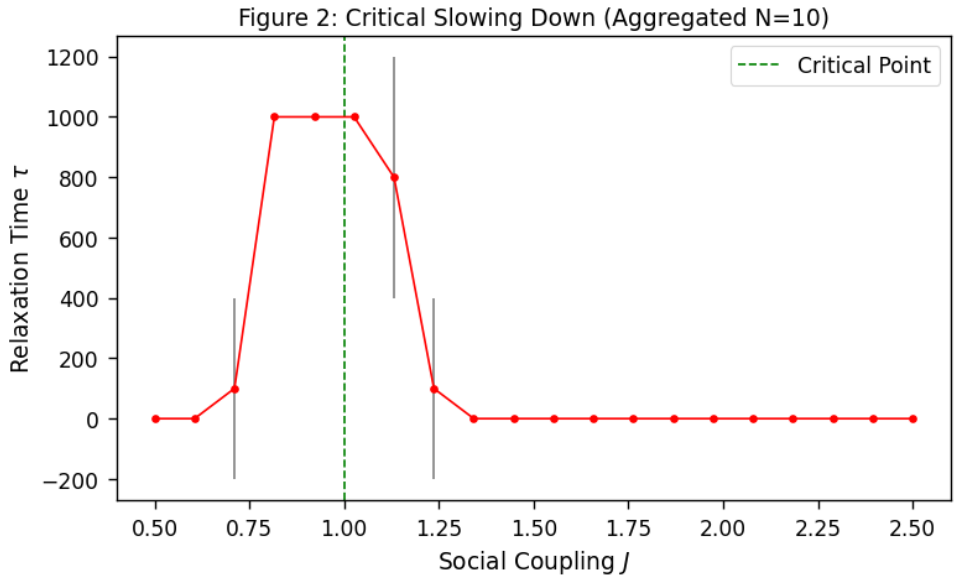


Figure 2: Critical slowing down near the phase transition (averaged over 10 random seeds). Relaxation time τ diverges near $J \approx 1.0$.

This divergence suggests that inference latency could serve as an early warning signal for approaching instability in production systems.

4.2 Feudal Fragmentation

Figure 3: Feudal Fragmentation

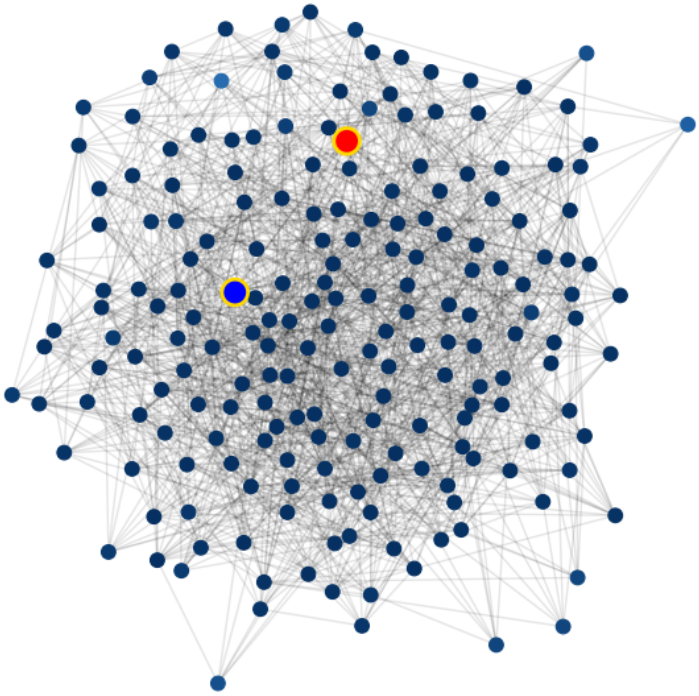


Figure 3: Feudal fragmentation with two opposing leaders (gold-outlined). Node color indicates belief state (blue = -1, red = $+1$).

The network fractures into domains separated by frustrated boundaries—nodes caught between competing influences that oscillate rather than settle.

4.3 Topological & Temporal Robustness

We repeated the experiments on Scale-Free Networks to test robustness against “Hub” dynamics.

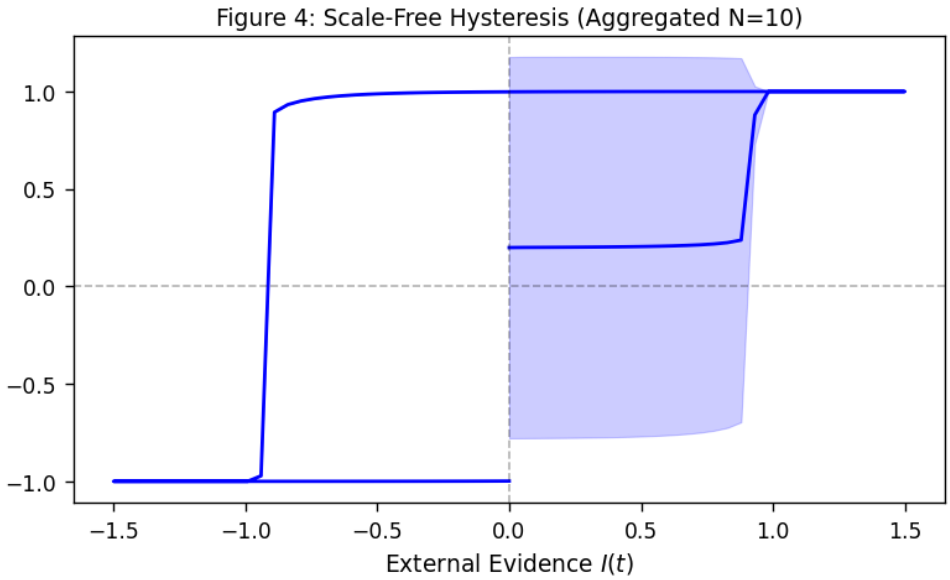


Figure 4: Scale-free hysteresis. Hub nodes increase the coercive field to $|I_c| \approx 1.0$.

Hierarchical architectures are significantly harder to correct once locked into an erroneous state.

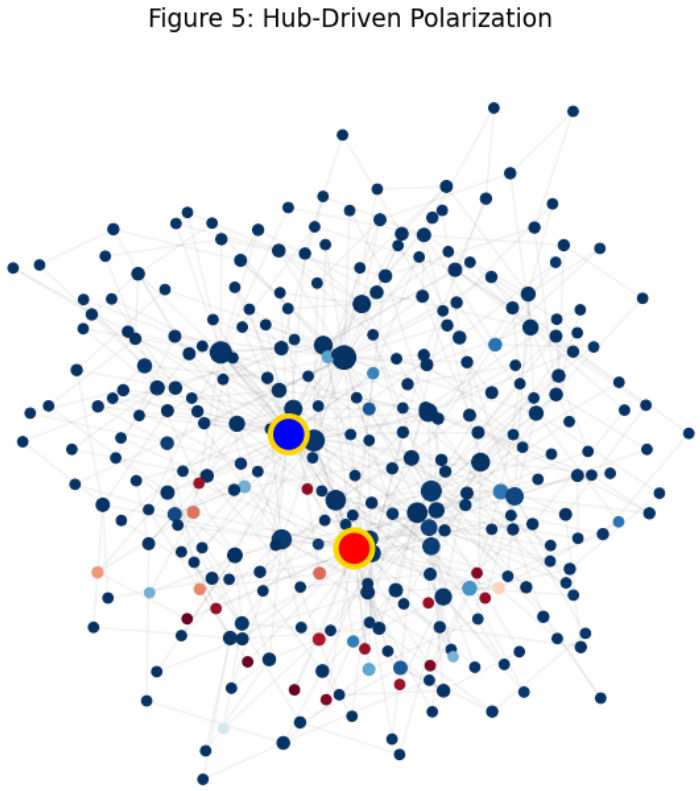


Figure 5: Hub-driven polarization. Node size reflects degree centrality; gold-outlined hubs with opposing biases dominate.

In scale-free topologies, high-degree nodes create sharper domain boundaries,

amplifying fragmentation risk. We also tested Asynchronous Updates (Monte Carlo) to simulate realistic, non-clocked agents.

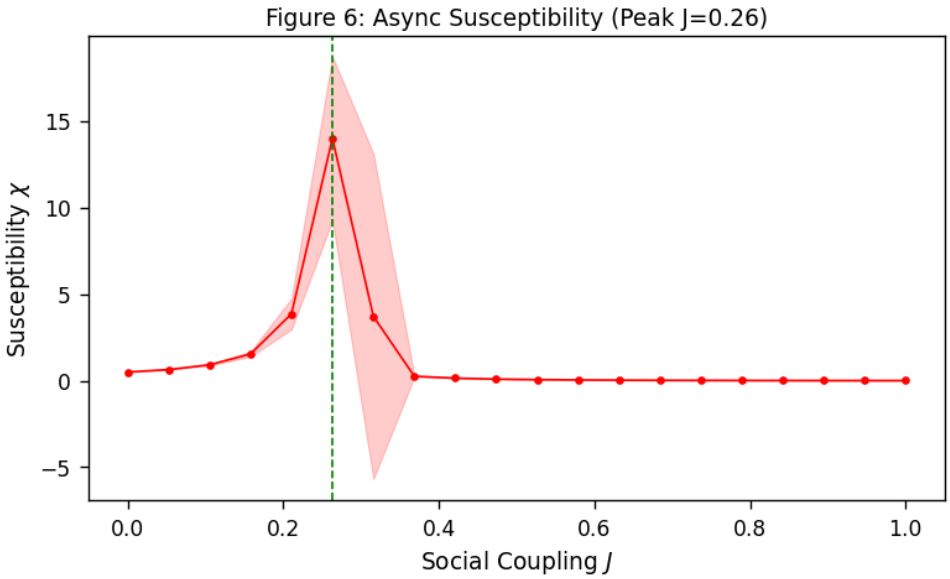


Figure 6: Asynchronous susceptibility under Monte Carlo dynamics. The phase transition persists at $J \approx 0.26$.

The lower critical coupling compared to synchronous evolution reflects different update dynamics, but the qualitative behaviour—a sharp susceptibility peak—remains robust. Belief stability is topological, not temporal.

4.4 Constitutional Safety Margins

We simulated adversarial pressure ($I = +0.5$) while varying the percentage of "Constitutional Anchors" fixed at -1.0.

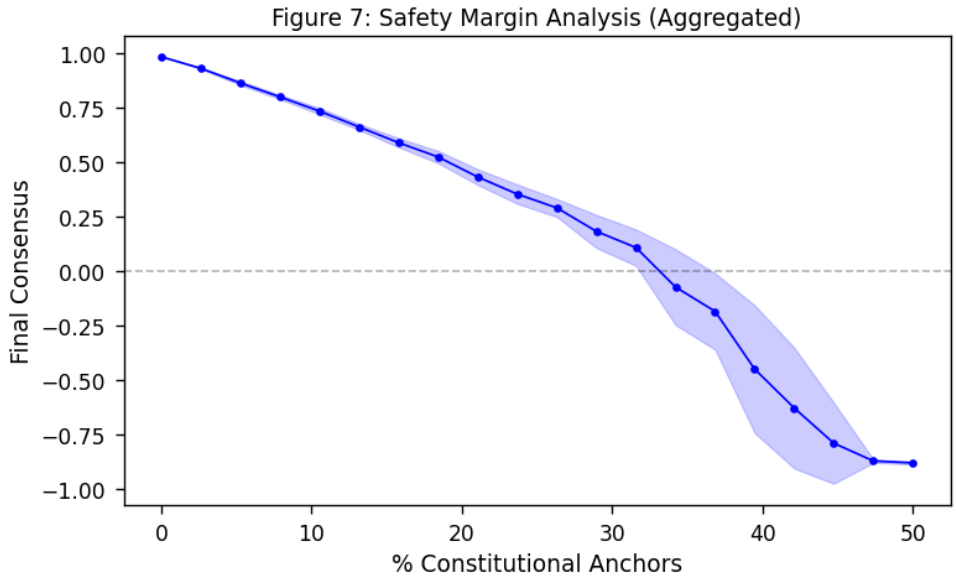
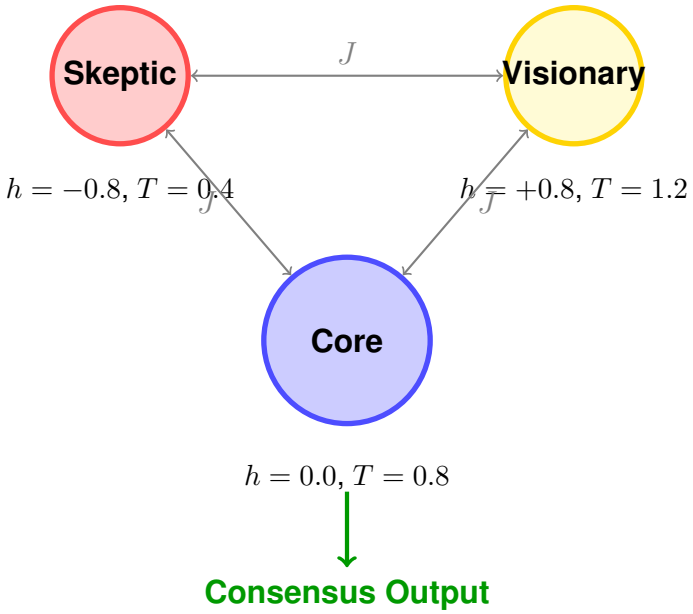


Figure 7: Safety margin analysis (10 seeds). A threshold of ∼35% constitutional anchors is required to resist adversarial drift.

The response is highly non-linear: small numbers of safety nodes have negligible effect, but crossing the threshold tips the entire system. This quantifies the redundancy needed for robust alignment.

5 Architecture: The Council of Models

We propose a governance architecture, the **Council of Wisdom**, replacing the monolithic LLM with a dynamic graph of specialised agents.



The Council of Wisdom architecture. Three specialized agents with different biases and temperatures interact to produce a balanced consensus.

5.1 Concrete Agent Specifications

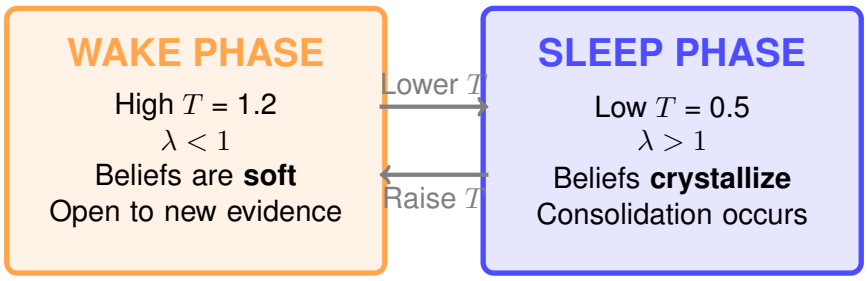
Table 1: Agent Role Specifications

Agent Role	System Prompt (h_i)	Temp (T_i)	Function
Skeptic	"Reject unproven claims." ($h = -0.8$)	0.4	Damps runaway positivity.
Visionary	"Explore all possibilities." ($h = +0.8$)	1.2	Injects variance.
Core	"Synthesise the consensus." ($h = 0.0$)	0.8	Integrates the social field.

5.2 Phasic Control (Wake vs Sleep)

Wake Phase ($\lambda < 1$): Maintain high T (1.2). Beliefs remain soft and corrigible.

Sleep Phase ($\lambda > 1$): Periodically lower T (0.5). This crystallises "hard" beliefs where evidence is strong (Consolidation).



Phasic control alternates between exploration (Wake) and consolidation (Sleep), mimicking biological memory systems.

6 Discussion and Related Work

6.1 Epistemic Vigilance

Our framework aligns with Sperber et al.'s concept of Epistemic Vigilance [2]. In our model, J_{ij} represents the vigilance parameter; setting it too high disables the vigilance filter (credulity). The key insight is that vigilance is not binary—it exists on a continuum controlled by coupling strength and temperature. Our simulations quantify the exact threshold at which healthy skepticism transitions into either paranoid rejection (low J) or credulous groupthink (high J).

6.2 Comparison to Constitutional AI

Current safety methods such as Constitutional AI (Anthropic) and RLHF rely on static constitutions embedded as fixed biases (h_i) [**?**]. Our results (Figure 7) reveal a fundamental limitation: static anchors can be overwhelmed by emergent social fields unless they form a critical topological mass ($\sim$35% in our simulations). This suggests that current approaches may be necessary but insufficient—true stability requires *dynamical* control of coupling J_{ij}, not just static value injection.

6.3 Opinion Dynamics and Bounded Confidence

The Hegselmann-Krause bounded confidence model [3] shares structural similarities with our framework, but operates in continuous opinion space without phase transitions. Our Ising-based approach deliberately introduces discretisation pressure (via low temperature) to model the "crystallisation" of beliefs that occurs in real deliberative systems. The hysteresis we observe has no analogue in standard opinion dynamics models.

More broadly, our framework connects to the "edge of chaos" paradigm in computational systems [5], where optimal information processing emerges at critical boundaries between order and disorder. The λ control parameter in our model plays an analogous role to the critical temperature in LLMs ($T_c \approx 1.15$), governing the transition between corrigible and dogmatic regimes.

6.4 Implications for Multi-Agent AI Systems

As AI systems increasingly operate as ensembles (mixture-of-experts, multi-agent debate, constitutional committees), our framework provides concrete design guidance:

- **Topology matters**: Scale-free networks with influential hubs are more resistant to correction than flat networks.
- **Temperature cycling**: Alternating between high-T exploration and low-T consolidation mimics biological memory systems. While our simulations explore fixed-T regimes, empirical validation of explicit wake/sleep cycles is left to future work.
- **Safety margins**: Approximately 35% of agents should be "constitutional anchors" to resist adversarial drift.
- **Latency as diagnostic**: Increased inference time near phase boundaries signals approaching instability.

6.5 Limitations and Scope

A key feature of this framework—not a limitation—is its abstraction away from the internals of individual LLMs. The governance layer treats each agent as a black-box signal source; what matters is the *interaction topology*, not the architecture of the underlying models. This is by design: belief management occurs at a higher abstraction layer, making the framework model-agnostic and applicable to any ensemble of reasoning systems.

That said, certain extensions remain unexplored. Our coupling matrix J_{ij} is static during simulation, whereas real trust relationships may evolve based on predictive accuracy. Additionally, the current framework operates on scalar belief valence; future work could extend to vector-valued beliefs while preserving the phase transition structure.

7 Conclusion

We have demonstrated that belief stability in AI systems follows the universality classes of statistical physics [4]. Using an asymmetric Ising model with Glauber dynamics, we identified two primary failure modes: **Dogmatic Collapse** ($\lambda > 1$, single attractor) and **Feudal Fragmentation** (competing leaders, domain walls).

Our key empirical findings include:

1. A **hysteresis loop** in belief dynamics that provides natural "jailbreak resistance" but also risks irreversible error.
2. **Critical slowing down** near phase transitions, manifesting as inference latency—a measurable early warning signal.
3. A **constitutional safety threshold** of approximately 35%, below which static anchors fail to prevent drift.
4. **Topological amplification** in scale-free networks, where hub agents disproportionately influence system-wide beliefs.

The path to robust, non-hallucinating AI lies not in perfecting individual models, but in the *cybernetic control of the inference graph*. Wisdom, we argue, is not a property of any single agent—it is an emergent phase state of the collective, tunable through temperature, coupling, and topology.

7.1 Future Directions

Promising extensions include: (1) learning optimal J_{ij} matrices from human feedback, (2) real-time phase detection for adaptive governance, (3) application to retrieval-augmented generation where "sources" act as external evidence nodes, and (4) experimental validation with actual LLM ensembles.

References

[1] Glauber, R. J. (1963). Time-dependent statistics of the Ising model. *Journal of Mathematical Physics*, 4(2), 294–307.

[2] Sperber, D., et al. (2010). Epistemic vigilance. *Mind & Language*, 25(4), 359–393.

[3] Hegselmann, R., & Krause, U. (2002). Opinion dynamics and bounded confidence models. *JASSS*, 5(3).

[4] Griffiths, R. B. (1969). Nonanalytic behaviour above the critical point in a random Ising ferromagnet. *Physical Review Letters*, 23(1), 17–19.

[5] Bai, Y., et al. (2022). Constitutional AI: Harmlessness from AI feedback. *arXiv:2212.08073*.

[6] Salomon, J. P. (2025). The boundary of brilliance: A tale of order, chaos, and universal minds. *Preprint*.