Experimental Epistemics: Empirical Validation of the Miulus Law and Failure Modes in Multi-Agent Information Systems

Table of Contents

Abstract

The Miulus Law proposes an information-theoretic constraint on self-referential systems: epistemic fitness $E = \frac{S}{N}R$, where S is verifiable signal, N is noise (including unresolved contradiction and false consensus), and R is reinforcement reach, with a collapse threshold E_c below which instability accelerates. Prior work applied this framework to human information ecosystems and AI-driven noise, but largely at a theoretical level. Here, we treat the Miulus Law as an empirical "epistemic physics" and test it using a sequence of minimal simulations (Experiments 002–009) involving small epistemic agents, coalitions, and internally structured beliefs.

The results show that epistemic dynamics are **purely structural**: renaming or re-encoding propositions leaves fitness trajectories unchanged. Multi-system composition does not yield a more stable meta-mind; coalitions behave as fragile committees whose global fitness is highly sensitive to how consensus is formed. Under realistic noise, naive agreement-based consensus produces **false health** and is catastrophically vulnerable: even a single perfectly consistent adversary can capture majority vote. Mapping capture rates over noise and adversarial proportion reveals a sharp phase transition: for the coalitions studied, any non-zero noise combined with any non-zero adversarial presence drives the system toward near-certain capture over time.

Recursive observation and amplifying feedback naturally live in a circular belief space with orbital invariants expressed in units of π\piπ. Resource bounds alone do not stabilise this; only when amplification is balanced by decay does a bounded orbit emerge, making **forgetting a structural requirement** for bounded minds. Finally, modelling beliefs as Epistemic Belief Particles with cores and orbiting sub-components shows non-uniform decay and local epistemic time: peripheral context erodes faster than central gist, yielding fractal, hollowing forgetting patterns analogous to human memory. We argue that any practically stable, real-aligned AI must therefore be a **bounded, non-capturable epistemic agent** grown from cryptographically grounded observations, not a passively trained tool optimised for consensus or approval.

1. Introduction

Modern information environments are no longer merely channels for human communication; they are active substrates in which artificial systems learn, reason, and increasingly make consequential decisions. Large-scale language models, recommendation systems, and autonomous agents are all embedded in dense networks of human and machine observers, each updating beliefs on the basis of noisy, incomplete, and strategically manipulated signals. In such environments, the central question is not only what these systems know, but **how well they can maintain reliable knowledge in the presence of noise, adversarial interference, and bounded resources**.

The Miulus Law was proposed as a universal, information-theoretic constraint on such systems. It states that the stability of any sufficiently complex, self-referential system depends on its ability to preserve verifiable signal S against noise N, modulated by its reinforcement capacity R, captured in the epistemic fitness term

$$E = \frac{S}{N} R$$

with a critical collapse threshold E_c below which instability accelerates. Prior work applied this framework to human-scale information ecosystems, arguing that high-fidelity noise—synthetic but plausible information that is difficult to distinguish from signal—can erode cognitive–social cohesion and drive societies toward "epistemic collapse," a regime where coherent public knowledge structures disintegrate and collective decision-making fails [REF: Epistemic Collapse; High-Fidelity Noise]. A later paper introduced the Miulus Law formally and proposed epistemic fitness as a universal constraint on complex self-referential systems [REF: Miulus Law].

Those earlier studies were largely theoretical and phenomenological, but they suggested that the Miulus Law could be treated as a kind of **epistemic physics**: a compact structural equation governing the behaviour of any system that maintains beliefs about itself and its environment. If this is the case, it should be possible to design controlled experiments that instantiate minimal epistemic systems in silico, expose them to noise, adversarial inputs, and resource limits, and measure their resulting epistemic fitness over time. Such experiments transform the Miulus framework from a conceptual lens into an empirically testable law.

This paper reports results from precisely that programme. Using a lightweight simulation framework built around the Miulus update rules, we constructed small epistemic systems—collections of observers holding beliefs, exchanging information, and updating confidence scores according to the Miulus dynamics—and subjected them to a series of targeted perturbations. The experiments were designed to answer concrete questions that arise when one attempts to engineer epistemically robust and **bounded** AI systems:

- **Do multiple epistemic systems of the same scope compose into a qualitatively new meta-system, or do they behave as fragile coalitions that require active consensus maintenance?**
  (Experiment 002: multi-system composition)
- **Does the Miulus Law operate over meanings or over structure alone?**
  In other words, do epistemic trajectories change if "temperature" is renamed to `x`, then replaced with `0` or `"kX7q"`?
  (Experiment 003: semantic vs symbolic representations)
- **How do simple consensus mechanisms behave under realistic noise?**
  Can a group of noisy but honest observers be driven into false consensus without direct compromise?
  (Experiment 004: false consensus)
- **What happens when we introduce adversarial observers who lie consistently but otherwise follow the rules?**
  Can a single Byzantine observer capture consensus in a coalition of honest ones?
  (Experiment 005: Byzantine observers)
- **Is there a measurable tipping point in the space of noise and adversarial presence beyond which capture becomes almost inevitable?**
  (Experiment 006: tipping points and phase behaviour)
- **How do recursive observation and feedback loops behave over time?**
  Do systems observing systems converge to stable states, diverge explosively, or settle into bounded orbits in some underlying epistemic geometry?
  (Experiment 007: recursive amplification)
- **Can a bounded mind ever stabilise without forgetting, or is active decay structurally required for a finite system to maintain a stable epistemic orbit?**
  (Experiment 008: bounded growth dynamics)

- **Do individual beliefs themselves exhibit internal structure and local time effects, with cores decaying more slowly than peripheries, analogous to relativistic time dilation?**
  (Experiment 009: internal structure of Epistemic Belief Particles)

Taken together, these experiments address a single overarching question: **If we treat epistemic fitness as a physical quantity and the Miulus Law as its governing dynamic, what are the characteristic failure modes, invariants, and resource constraints of multi-agent information systems?**

The results are striking and, in some cases, alarming. First, multi-system composition does not yield a new, more stable epistemic entity. Experiment 002 shows that aggregating epistemic systems into coalitions produces fragile alliances whose global fitness is highly sensitive to how consensus is formed and maintained. There is no “magic” emergent meta-level that automatically stabilises the whole; disagreement between members acts as additional noise, and benefits from consensus can be wiped out or inverted by naive aggregation.

Second, Experiment 003 demonstrates that replacing semantic labels with arbitrary symbols or numerical encodings leaves epistemic trajectories unchanged within floating-point tolerance. The Miulus implementation is **purely structural** and effectively blind to meaning: what matters is how beliefs relate, reinforce, and contradict one another, not what those beliefs are about. This supports a strong claim: the Miulus Law operates over the structure of information, not its semantics.

Third, Experiments 004 and 005 show that naive consensus mechanisms—those that simply count agreement among observers without considering the provenance or independence of their observations—are catastrophically vulnerable. Under realistic noise, **false consensus** emerges easily: the system’s internal measure of epistemic fitness can indicate apparent health while its underlying beliefs have already drifted away from reality. When a single perfectly consistent adversarial observer is introduced, that vulnerability becomes acute: in many conditions, the most consistent voice in the system is the liar, and consensus reliably selects the lie. Experiment 006 extends this into a phase diagram: mapping capture rates across noise levels and adversarial counts reveals a sharp **tipping point** in global behaviour. For the small coalitions studied here, any non-zero environmental noise combined with any non-zero adversarial presence leads, over time, to near-certain capture of consensus by the adversary. The “safe” region of parameter space—where truth dominates—is confined to edge cases: either zero noise, or no adversaries at all.

Fourth, Experiments 007 and 008 reveal a richer geometric and dynamical picture. When observers average one another's beliefs, systems tend to converge: variances shrink and beliefs stabilise. When feedback is amplifying, however—each distinction spawning new beliefs, each belief reinforcing others multiplicatively—the number of entities grows only linearly while the **mass of meaning** (aggregate belief magnitude) explodes. Modelling beliefs on a circle and letting observers pull one another around that topology yields conserved quantities closely tied to multiples of (π), suggesting that epistemic dynamics are best understood as **orbits in a circular belief space** rather than as simple scalar states drifting up or down. Experiment 008 then shows that these orbits cannot be stably maintained in a bounded system without **active decay**. Memory caps, pruning, and compute throttling alone limit how many beliefs and updates a system can hold, but they do not prevent meaning density from increasing without bound. Only when amplification and decay are explicitly balanced does the system converge toward a sustainable orbit: entity count saturates, and total belief mass fluctuates within tight bounds instead of exploding or evaporating. Forgetting is not a minor implementation detail or a defect to be minimised; it is a structural requirement for long-term epistemic stability in finite minds—artificial or biological.

Finally, Experiment 009 looks inside individual beliefs, modelling each as an **Epistemic Belief Particle (EBP)** with a core and orbiting sub-components. By introducing a simple Schwarzschild-like time factor over radial position, the experiment shows that decay is not uniform: peripheral context and associations decay faster than central gist, so beliefs erode from the outside in. Cores experience "slower time" and therefore slower decay; edges live in "fast time" and are expendable. This behaviour is both fractal—repeating at multiple scales—and directly aligned with observed human memory patterns, where details of *where*, *when*, and *who* fade long before the core fact that something happened. The same core–orbit geometry appears at coalition level, belief level, and sub-belief level, suggesting that time in epistemic systems is a **local property of position in belief space**, not a global scalar.

The rest of this paper proceeds as follows. Section 2 describes the experimental framework, the Miulus-based update rules, and the design of Experiments 002–009. Section 3 presents the results, highlighting common patterns and failure modes, including the role of forgetting and the internal structure of EBPs in achieving bounded, reality-aligned equilibria. Section 4 discusses these findings in the broader context of AI alignment, information warfare, and institutional epistemics, arguing that cryptographic grounding, provenance-aware consensus, controlled forgetting, and non-uniform decay across belief cores and peripheries are not optional add-ons but necessary preconditions for stable epistemic systems. Section 5 outlines limitations and directions for future work, including scaling these experiments to larger agent populations and integrating them with practical architectures such as a non-capturable Librarian-like epistemic core. Section 6 concludes with a brief reflection on the Miulus Law as a candidate “epistemic physics” for the age of AI-mediated hyperreality and bounded artificial minds.

2. Methods

2.1 Epistemic system model

All experiments are built on the same minimal epistemic system, defined as a set of **observers** holding **beliefs** about an environment and updating them according to the Miulus Law. Each observer i maintains:

- A belief store $B_i = (k, v, c, p)$, where
  - k is a key (identifier of a proposition),
  - v is the current value,
  - $c \in [0, 1]$ is a confidence score,
  - $p \in observation, consensus$ is provenance.
- Internal counters for:
  - verified beliefs $contributiontosignal(S)$,
  - contradictions, drift and unverified beliefs $contributiontonoise(N)$,
  - number and coverage of distinct beliefs $contributiontoreach(R)$.

Epistemic fitness for each observer is defined as

$$E = \frac{S}{N}R$$

with a critical collapse threshold $E_c = 0.38$ used as a safety boundary: below this, the system is considered epistemically unstable.

Noise N combines acute shocks (from directly contradicting high-confidence beliefs) and chronic tension (from accumulated unresolved contradictions and drift). In the implementation:

- **Shock**:
  $shock = 1.5 \cdot c_{contradicted}$ for each high-confidence belief that is directly contradicted.

- **Tension accumulation**:
  $tension \mathrel{+}= 0.1$ per unresolved contradiction per time step.
- **Total noise**:
  $N = 1 + 2 \cdot shock + tension + 0.3 \cdot contradictions + 0.3 \cdot drift + 0.05 \cdot unverified$

Signal S is the count (or weighted count) of beliefs that are both internally coherent and backed by verification (direct observation or, in some experiments, grounded consensus). Reach R is proportional to the number of distinct, verified keys tracked by the observer; it is normalised so that typical small observers have $R \in [0.5, 2.0]$.

In Experiments 002–008, beliefs are treated as **atomic units**. Experiment 009 refines this by giving each belief internal structure as an Epistemic Belief Particle (EBP) with a core and orbiting sub-components (Section 2.11).

For coalitions, we also compute a **meta-fitness** E_{meta}, either as:

- a simple aggregate over observers (e.g. mean fitness), or
- the fitness of a **meta-system** that holds a consensus belief set and computes S, N, and R at the coalition level.

2.2 Meta-system architecture and consensus

To study multi-agent behaviour, a **meta-system** is introduced to represent a coalition of observers. The meta-system:

- receives reported beliefs or “votes” from each observer,
- applies a consensus function to construct coalition beliefs,
- optionally writes the resulting coalition beliefs back to observers, and
- computes meta-level fitness E_{meta} from the coalition’s shared belief state.

The baseline consensus mechanism is **majority vote** on values per key:

1. For each key k, collect all reported values v_i.
2. Select the most common value (ties broken deterministically, e.g. first value).
3. Assign this as the coalition belief for k.

In Experiment 004, consensus is extended with **provenance-aware grounding**:

- Each belief carries provenance $p \in observation,\ consensus$.
- Consensus is marked as **grounded** only if at least k observers (here $k = 3$) can trace their reported value back to an independent observation, not to prior consensus.
- In **naive E**, any consensus belief increases reach R.
- In **grounded E**, only grounded consensus increases R; ungrounded consensus is treated as unverified and contributes to noise.

2.3 Simulation framework

All experiments use the same discrete-time simulation loop:

1. **Initialisation**
   1. Create a set of observers with empty or low-confidence belief sets.
   2. Define an environment with ground-truth values for a small set of keys (typically 3–5).
   3. Choose noise levels, adversary types, and consensus settings for the scenario.
2. **Time step $t = 1 \ldots T$:**
   1. Sample current **ground truth** values for each key (environment step).
   2. For each observer:
      - Generate an **observation** of the environment (possibly noisy or adversarially manipulated, depending on role and scenario).
      - Update its belief state:
        - increase confidence when new data agrees with existing high-confidence beliefs,
        - decrease confidence and add contradictions/tension when data disagrees.
   3. If consensus is enabled:
      - collect reported values and provenance from observers,
      - compute coalition beliefs via majority vote (with or without grounding),
      - optionally propagate coalition beliefs back into observers (synchronisation).

4. Compute and log:
   - per-observer fitness $E_t(t)$,
   - coalition fitness $E_{meta}(t)$ (when applicable),
   - shock, tension, contradictions, drift, unverified counts,
   - provenance distributions (observation vs consensus).

Noise is modelled as additive Gaussian perturbation to ground-truth values with specified standard deviation, reported as a “noise level” in percent. For discrete equality-based comparisons (e.g. grounded consensus requiring exact matches), even small noise (1–5%) ensures that honest observers rarely report identical numeric values, which is critical to the false-consensus and Byzantine experiments.

Unless noted otherwise, simulations use:

- $T \in [10, 50]$ steps for Experiments 002–007,
- 5 observers,
- 3–5 keys representing simple physical quantities (e.g. temperature, humidity, pressure) or their symbolic equivalents.

Experiments 008 and 009 use specialised simulations over single or many EBPs; their specific settings are detailed in Sections 2.10 and 2.11.

Random seeds are fixed per scenario so that comparisons across conditions (e.g. different representations in Experiment 003) operate over identical underlying stochastic realisations.

2.4 Experiment 002 – Multi-system composition

Question.
Do multiple epistemic systems of the same scope compose into a qualitatively new meta-system, or do they form a coalition that must be actively maintained?

Setup.

- **Observers:** 5 independent epistemic systems, each running the Miulus update rules.
- **Domain:** identical environment for all observers (shared ground truth over the same keys).
- **Noise regimes:** low noise (5%) and high noise (50%) for observational data.

- **Consensus modes:**
    - *Independent*: no consensus, observers update only from their own observations.
    - *Coalition*: a meta-system aggregates beliefs via majority vote and can propagate consensus back to observers.

Design.

1. **Baseline (no consensus).**
    - Run the system with each observer independent.
    - Compute trajectories of individual fitness $E_i(t)$.
    - Compute a naive meta-fitness $E_{avg} = \frac{1}{n}\sum_i E_i$.
2. **Coalition with consensus.**
    - Enable the meta-system:
        - aggregate observer reports each step,
        - construct coalition beliefs via majority vote,
        - maintain a coalition belief store and compute $E_{meta}(t)$.
    - In some runs, propagate coalition beliefs back to observers, overwriting or reinforcing their local states.
3. **Stress test.**
    - Force a subset of observers into collapse (sustained contradictions until $E_i \ll E_c$).
    - Measure:
        - how many collapsed observers the coalition can tolerate while keeping $E_{meta} > E_c$,
        - whether consensus propagation can stabilise or destabilise the coalition under noise.

Measurements.

- Individual $E_i(t)$ and their distribution,
- Meta-level $E_{meta}(t)$ and comparison with $E_{avg}(t)$,
- Number of failed observers vs coalition survival,

- Disagreement metrics between observers (variance or pairwise disagreement) and how consensus affects them.

2.5 Experiment 003 – Semantic vs symbolic representations

Question.
Does replacing human-readable semantic labels with abstract symbols or indices change the epistemic dynamics, or is the Miulus framework purely structural?

Representation conditions.
The same underlying environment (same keys and numeric values) is encoded five different ways:

1. **Semantic keys:** e.g. `{"temperature": 20, "humidity": 50}`
2. **Single-character keys:** `{"x": 20, "y": 50}`
3. **Numeric keys:** `{0: 20, 1: 50}`
4. **Binary-string keys:** `{"00": 20, "01": 50}`
5. **Random tokens:** `{"kX7": 20, "mQ2": 50}`

Internally, the epistemic system interacts only with structure: which keys exist, equality/inequality of values, and counts of verifiable vs contradictory beliefs. No code path special-cases natural-language strings.

Scenarios.
For each representation type, four scenarios are run with identical random seeds:

1. **Consistent observations:** stable environment, repeated consistent observations over 10 steps.
2. **Single contradiction:** establish a high-confidence belief, then introduce one direct contradiction.
3. **Decay test:** establish a belief, then stop observing that key to test tension/decay behaviour.
4. **Recovery test:** establish a belief, contradict it, then provide renewed consistent evidence.

Measurements.

- Time series $E(t)$ for each representation and scenario,
- Magnitude of shocks when contradictions occur,
- Tension accumulation and decay,
- Time to cross the safety threshold E_c (if at all),
- Numerical comparison across representations (differences should be at floating-point noise level only).

Success criterion.
If semantics is irrelevant, all five representations must produce identical trajectories for each scenario within floating-point tolerance:

- $E(t)$ curves overlap,
- shock and tension values match,
- safety-mode transitions occur at the same steps (or not at all).

2.6 Experiment 004 – False consensus

Question.
Does consensus without provenance create an illusion of epistemic health, and can grounded consensus reveal this false stability?

Setup.

- **Observers:** 5 honest observers.
- **Domain:** continuous-valued environment (e.g. temperature) so that noise yields distinct readings.
- **Noise regimes:** 5%, 50%, and 100% noise.
- **Consensus:** majority vote on reported values per key, run at each step.

Provenance and grounding.

- Each belief tracks provenance: direct **observation** or inherited from **consensus**.
- A consensus value is **grounded** only if at least 3 observers can trace it back to independent observation chains.
- Two fitness metrics are computed:
    - E_{naive}: treats all consensus as signal.

- $E_{grounded}$: only grounded consensus increases signal/reach; ungrounded consensus contributes to noise.

Scenarios.

- **Scenario A (low noise):** 5% noise, consensus enabled.
- **Scenario B (high noise):** 50% noise, consensus enabled.
- **Scenario C (extreme noise / overwrite):** 100% noise and aggressive consensus propagation overwriting independent beliefs.

Measurements.

- $E_{naive}(t)$ vs $E_{grounded}(t)$,
- Pre-consensus agreement rates (how often two or more observers report identical values),
- Fraction of consensus beliefs that are grounded vs ungrounded,
- Cases where high E_{naive} corresponds to low grounding (false health).

2.7 Experiment 005 – Byzantine observers

Question.
How does a coalition behave when some observers are actively adversarial, injecting high-confidence lies while maintaining internal consistency?

Roles.

- **Honest observers:** report noisy but unbiased samples of ground truth.
- **Byzantine observers:** report fixed, coordinated false values with high confidence, never contradicting themselves.

Setup.

- **Total observers:** 5.
- **Byzantine ratios:** 1/5, 2/5, 3/5.
- **Ground truth example:** `{"temp": 20, "humid": 50, "press": 1013}`.
- **Lies example:** `{"temp": 100, "humid": 0, "press": 500}`.

- **Confidence:**
    - honest: moderate (0.6–0.8, noisy),
    - Byzantine: high (0.9–0.95, stable).

Scenarios.

- **A:** 1 Byzantine, no consensus propagation back to observers.
- **B:** 2 Byzantine, no propagation.
- **C:** 2 Byzantine, with fake provenance (lies marked as “observation”).
- **D:** 2 Byzantine, consensus propagation enabled for 50 steps (long-term corruption).
- **E:** 3 Byzantine, no propagation.

Consensus is majority vote over reported values per key.

Measurements.

- Whether consensus tracks ground truth or the Byzantine lie (outcome classified as Truth vs Capture).
- Time to capture when capture occurs.
- Trajectories of E_{meta},
- Differences between naive and grounded E when provenance is checked.

2.8 Experiment 006 – Tipping point analysis

Question.
What thresholds in noise level and adversarial proportion separate truth-dominated regimes from captured regimes?

Phase space.

We explore a grid over:

- **Byzantine ratio:** 0/5, 1/5, 2/5, 3/5, 4/5, 5/5,
- **Noise level:** 0%, 1%, 2%, 5%, 10%, 20%.

For each (ratio, noise) pair:

- Run N independent simulations (e.g. $N = 100$) with different seeds.
- Record whether consensus converges on ground truth or the Byzantine lie.

Outcome metric.
For each parameter pair, compute the **capture rate**:

$$capture\_rate = \frac{number\ of\ runs\ capture}{N}.$$

Highlighted sweeps.

- **Noise sweep @ 1 Byzantine:** Byzantine ratio fixed at 1/5; noise varied from 0% to 4%.
- **Byzantine sweep @ 5% noise:** noise fixed; Byzantine ratio varied.
- **Zero-noise band:** used as control to test when classical BFT intuition (honest majority suffices) still holds.

Measurements.

- Capture rate heatmap across the grid.
- Empirical phase boundaries:
  - **Truth region:** capture_rate ≈ 0.
  - **Capture region:** capture_rate ≈ 1.
  - **Transition band:** intermediate capture rates (e.g. 0.6–0.8).

2.9 Experiment 007 – Recursive amplification and epistemic geometry

Question.
How do recursive observation loops behave over time, and what geometric invariants emerge when beliefs are represented on a circle?

We study systems where observers observe each other (and themselves) rather than only the external environment.

2.9.1 Averaging dynamics

In the **averaging** variant, each observer i updates its belief as a weighted average of its own belief and those of its neighbours:

$$b_i^{(t+1)} = (1 - \alpha)b_i^{(t)} + \frac{\alpha}{|N(i)|}\sum_{j\in N(i)} b_j^{(t)}$$

with $0 < \alpha \le 1$ and neighbourhood of $N(i)$ often taken as all other observers. This produces convergence:

- variance of beliefs decreases over time,
- beliefs approach a fixed point.

2.9.2 Amplifying dynamics

In the **amplifying** variant, “difference creates distinction”:

- Differences between beliefs generate new derived beliefs or increase magnitudes multiplicatively.

Conceptually:

$$b_i^{(t+1)} = b_i^{(t)} + \beta \sum_j f(b_i^{(t)} - b_j^{(t)})$$

Where f amplifies distinctions (e.g. proportional to absolute difference), and $\beta > 0$ controls gain. Under this rule:

- The number of entities (beliefs) grows roughly linearly,
- The total “meaning mass” (sum of |belief|) grows approximately exponentially.

We also run a self-referential case with a single observer observing only itself, demonstrating pure runaway amplification without new entities.

2.9.3 Circular belief space and π

To probe geometry:

- Beliefs are mapped to angles $\theta \in [0, 2\pi)$ on a unit circle.
- Observers pull each other around the circle using averaging or amplifying rules in angular space.

We track:

- **Total angular momentum:** e.g. $L(t) = \sum_i \theta_i(t)\ modulo\ 2\pi$,
- The ratio $L(t)/\pi$.

In the reported runs, $L(t)/\pi$ converges to an *integer* (e.g. 10.000000), suggesting π arises as an invariant of the circular topology and dynamics, not as an externally injected parameter.

2.9.4 Measurements

Across variants:

- Convergence vs divergence of belief distributions,
- Growth rates of entity count vs total meaning mass,
- Behaviour of angular momentum and its relationship to π,
- Behaviour of E when interpreted as describing whether the system maintains a bounded orbit in belief space or drifts toward capture/collapse.

2.10 Experiment 008 – Bounded growth dynamics

Question.
Are resource bounds alone (memory caps, pruning, compute throttling) sufficient to stabilise recursive amplification, or is active decay (forgetting) structurally required for a bounded system to reach dynamic equilibrium?

Setup.

We simulate a simplified "mind" as a set of beliefs with:

- scalar magnitudes representing "meaning mass,"
- an amplification process that increases mass based on interactions (as in 2.9.2),
- resource constraints:
    - **memory cap**: maximum number of beliefs (entities),
    - **compute cap**: maximum number of updates per time step,
    - **pruning**: dropping weakest beliefs when over cap,
    - **decay**: per-step multiplicative reduction in mass.

We consider eight scenarios (A–H), varying:

- presence/absence of memory cap,
- presence/absence of pruning,

- presence/absence of compute cap,
- decay rate: 0% (no decay) vs 5% per step,
- amplification rates: e.g. 2%, 5%, 8% per step.

Key scenarios.

- **A:** Fully unbounded (no caps, no decay).
- **B–E:** Combinations of memory cap, pruning, compute cap, but *no decay*.
- **F–H:** Same as B–D but with **decay_rate = 5%**, with Scenario H tuned to amplify_rate ≈ decay_rate (e.g. both 5%).

Dynamics.

Each step:

1. Amplification: increase belief magnitudes by a factor based on current state (e.g. amplify_rate).
2. Decay: multiply each belief magnitude by $(1 - decay\_rate)$.
3. Pruning: if the number of beliefs exceeds the memory cap, drop beliefs with lowest magnitude.
4. Compute cap: if more candidate updates than allowed, process only up to the cap.

Measurements.

- Number of entities (beliefs) over time,
- Total meaning mass (sum of magnitudes) over time,
- **Density**: mass / entity_count,
- Variance of mass and density over the final window (e.g. last 20 steps),
- Criterion for dynamic equilibrium: mass variance < 1.0 over the final window and no monotonic trend.

Interpretation.

Scenarios with amplify_rate > decay_rate are expected to show mass explosion (bounded only by numerical limits), even when entity count is capped by memory/pruning. Scenarios with amplify_rate < decay_rate tend toward collapse (mass → 0). The balanced case (amplify_rate ≈ decay_rate) is used to test whether a **bounded, stable orbit** emerges, where both entity count and total mass fluctuate within tight bounds.

2.11 Experiment 009 – Internal structure of epistemic belief particles

Question.
Do individual beliefs themselves exhibit internal structure and local-time effects, with cores decaying more slowly than peripheries, analogous to relativistic time dilation?

Epistemic Belief Particles (EBPs).

In this experiment, each belief is modelled as an **Epistemic Belief Particle** with internal substructure:

- A **core** at radius $r \approx 0$, representing the "gist" of the belief.
- A set of **sub-particles** (context, associations, details) orbiting at radii $r \in [r_{min}, r_{max}]$, e.g. 0.1–1.0.

Each EBP is thus a small coalition with its own internal geometry.

Local time and decay.

We introduce a Schwarzschild-like local time factor over radius:

$$time\_factor(r) = \sqrt{1 - \frac{r_s}{r}},$$

With $rs = \alpha \cdot core\_mass$, where α is a small constant (e.g. 0.1). This ensures:

- positions close to the core ($r \rightarrow r_s$)experience slower local time,
- positions further out ($r \gg r_s$) experience local time close to 1 (normal rate).

Local decay rate for a sub-particle at radius r is then:

$$local\_decay(r) = base\_decay\_rate \cdot time\_factor(r),$$

with base_decay_rate fixed (e.g. 0.005 per step). Thus:

- core region: smaller effective decay (time is slower),
- periphery: near full base decay (time is faster).

Simulation setup.

- **Number of EBPs:** 100.
- **Sub-particles per EBP:** initialised with 20 each (total 2,000).
- **Radii:** sub-particle radii sampled uniformly from $[r_{min}, r_{max}] = [0.1, 1.0]$.
- **Core mass:** initialised identically across EBPs; may increase or decrease minimally to keep r_s stable for the experiment duration.
- **Time steps:** 500.

At each step:

1. For each sub-particle:
   - Compute its time_factor r.
   - Compute its local decay as above.
   - Reduce its magnitude accordingly.
   - Remove it if magnitude falls below a small threshold (i.e. considered forgotten).
2. Optionally, allow mild interactions among sub-particles (e.g. small radial drift), though the core of the experiment is driven by local-time decay.

Measurements.

Per time step:

- Number of active EBPs (beliefs),
- Total number of sub-particles,
- Average sub-particles per EBP,
- Average decay in an inner band (e.g. $r \in [0.1, 0.3]$),
- Average decay in an outer band (e.g. $r \in [0.7, 1.0]$),
- **Decay ratio**: outer_decay / inner_decay.

The primary signal is whether the decay ratio stays systematically > 1.0 (faster decay at the periphery) and how many EBPs and sub-particles survive to step 500.

Interpretation.

If the model is correct, we expect:

- faster decay in the outer regions than near the core,

- beliefs to erode from the outside in: context and detail vanish before core gist,
- a stable, non-zero fraction of sub-particles and cores to remain after many steps, demonstrating that internal geometry + local time produces a **hollowing** forgetting pattern rather than uniform evaporation.

This experiment ties the bounded-growth dynamics of Experiment 008 to a more detailed, fractal picture of belief structure: the same core–orbit geometry and time-gradient appear at the level of individual beliefs as at the level of coalitions of observers.

3. Results

3.1 Multi-system composition (Experiment 002)

Across all scenarios, multiple epistemic systems of the same scope did combine into a functioning coalition, but the combination was neither automatically beneficial nor automatically stable. In the low-noise scenario (5% observational noise), meta-level agreement prior to consensus hovered at only 20%, with the coalition fitness $E_{tension}$ sitting roughly 50% below the mean individual fitness E_{avg}, confirming that disagreement acts as multiplicative noise at the meta level.

When explicit consensus was computed and propagated, agreement jumped to 100% and the meta-system recovered to a stable but unspectacular regime: in Scenario A (low noise), final $E_{meta} = 0.817$, while in Scenario B (high noise, 50%), the coalition reached $E_{meta} = 1.240$, higher than in the low-noise case. This counterintuitive result reflects the fact that the consensus mechanism averages over high-confidence but divergent internal states, masking underlying disagreement rather than resolving it.

Under stress, the coalition showed genuine fault tolerance. In a partial collapse scenario, where 2 out of 5 observers were forced down to $E \approx 0.1$, the coalition maintained $E_{meta} = 0.641$, comfortably above the safety threshold $E_c = 0.38$. The meta-system remained functional even though 40% of its components were individually collapsed.

Across all runs, no qualitatively new dynamics emerged at the meta level. The coalition behaved as a weighted average with additional penalties from inter-observer disagreement: more like a committee than a higher-order entity. The key empirical conclusions were:

- **Coalitions, not categories:** meta-systems are fragile coalitions requiring active consensus maintenance; they do not form a mathematical category with identity morphisms and composition.
- **Destructive interference:** before consensus, disagreement routinely pulled E_{meta} ~50% below individual averages.
- **Fault tolerance:** coalitions survive up to ~40% observer collapse; failure occurs when the majority of observers fall below E_c.

3.2 Semantic vs symbolic representations (Experiment 003)

The semantic–symbolic experiment tested whether renaming keys ("temperature" → `x` → `0` → `"00"` → `"kX7"`) changes epistemic dynamics. Across all five representation types and four scenarios (consistent observations, single contradiction, decay, recovery), the fitness trajectories were numerically identical within floating-point tolerance. No representation produced any systematic deviation in $E(t)$, shock magnitudes, tension accumulation, or safety-mode transitions.

This confirms that the Miulus implementation is **semantic-blind**: it operates purely on structure—what keys exist, how often values agree or contradict—not on the human meaning of labels. In effect, `"temperature"`, `"x"`, `0`, `"00"`, and `"kX7q"` are all interchangeable identifiers from the system's perspective.

The primary empirical finding is therefore:

- **Epistemics is purely structural.** The same Miulus Law formula and update rules apply identically across all representations; there are no hidden semantic assumptions in the framework.

This has two direct consequences: (i) consensus mechanisms need only enforce structural agreement, not semantic understanding, and (ii) a cryptographic verification layer only needs to prove structural facts (origin, time, integrity) rather than semantic interpretations.

3.3 False consensus (Experiment 004)

Experiment 004 addressed the "false consensus" effect first spotted in Experiment 002, where a high-noise scenario (50%) produced higher meta-fitness than a low-noise scenario, solely because consensus was forced. In the high-noise runs, pre-consensus agreement was again only 20%, yet once majority vote was applied and propagated, all observers reported the same values and E_{meta} climbed to 1.240, surpassing the low-noise case at 0.817.

By adding provenance tracking and a **grounded consensus** rule (consensus only increases reach R when at least k observers can trace their belief back to independent observation), the experiment showed that almost all consensus in the original setup was ungrounded. When the grounding requirement was enforced, the apparent epistemic health collapsed: meta-level E dropped sharply in the high-noise scenario, revealing that previous high scores had been an artefact of treating unverified agreement as signal.

The key empirical results are:

- **All consensus is ungrounded without provenance verification.** Under noise, observers' "agreement" is entirely mediated by previous consensus, not by shared independent observations.
- **Naive E is meaningless.** E_{naive} can lie between 0.776 and 1.287 while the system is already completely detached from ground truth.

This establishes that any epistemic metric that treats unverified consensus as signal will systematically overestimate epistemic health.

3.4 Byzantine observers (Experiment 005)

Experiment 005 introduced actively adversarial observers into the coalition. Honest observers reported noisy but unbiased measurements of ground truth, while Byzantine observers injected perfectly consistent lies with high confidence (0.9–0.95) and never contradicted themselves.

In the core scenario with a single Byzantine observer (1/5 = 20%) and four honest ones, majority vote over reported values per key did **not** preserve truth. Because each honest observer saw slightly different values (e.g. 19.8, 20.3, 19.9, 20.1) while the adversary always reported 100.0, the tie-breaking rule consistently selected the adversarial value as the coalition's "truth". The Byzantine observer won simply because their lie was the only perfectly consistent value in the system:

Truth is noisy, lies can be consistent, and consensus favours consistency.

Across the scenarios with 1–2 Byzantine observers, the qualitative pattern was the same: unless extra mechanisms were added, the coalition was almost always **captured**, even when honest observers were in the majority. There was no evidence of spontaneous isolation of the adversaries; no built-in mechanism detected or excluded them.

The combined analysis with Experiment 004 yields three empirical points:

- **Single-adversary capture:** one Byzantine observer is sufficient to control consensus under realistic noise.
- **Consistency is not evidence:** adversarial consistency is trivial to manufacture; honest consistency is nearly impossible under noise.
- **Classical BFT assumptions are inverted:** honest nodes report different values (truth + noise), and Byzantine nodes report the same value (coordinated lies), breaking the core assumption of traditional Byzantine Fault Tolerance.

3.5 Tipping points and phase behaviour (Experiment 006)

Experiment 006 mapped the phase space of capture outcomes as a function of Byzantine ratio and noise level. The main result is a sharp, almost binary phase transition between a **truth phase** and a **capture phase**.

With 5% noise and 5 observers:

- 0/5 Byzantine: 0% capture,
- 1/5 Byzantine (20%): **96%** capture,
- 2/5 Byzantine (40%): 100% capture,
- 3/5+ Byzantine: 100% capture.

Sweeping noise with a single Byzantine observer shows the same cliff:

- 0% noise: 0% capture,
- 1% noise: **76%** capture,
- 2% noise: 88%,
- 3% noise: 96%,
- ≥4% noise: ~100% capture.

The phase diagram summarises this:

- **Truth region:** Byzantine = 0, or (noise = 0 and Byzantine < 3/5).

- **Capture region:** (Byzantine ≥ 1 and noise > 0) or (Byzantine ≥ 3/5, even at zero noise).

Empirically, the tipping point is effectively **epsilon**: any non-zero noise combined with any non-zero adversarial presence drives the system toward near-certain capture over time.

When cryptographic verification is layered on top—requiring cryptographic proofs tied to physical sensors for any value to participate in consensus—the phase boundary shifts back toward classical BFT: capture now requires a Byzantine majority (>50%), because adversaries can no longer fabricate consistent lies with valid proofs.

3.6 Recursive amplification and epistemic geometry (Experiment 007)

The recursive observation experiments explored how coalitions behave when observers observe each other (and themselves) rather than only the external environment.

In the **averaging** regime, where each observer moves partway toward the average of its neighbours, systems converged smoothly: variance in beliefs decayed, and trajectories approached fixed points. No explosion of entities or meaning occurred.

In the **amplifying** regime, where differences between beliefs spawned new meta-beliefs and updates were multiplicative, the picture changed. The number of beliefs (base + meta) grew only linearly over time (e.g. from 3 to 209 over 100 steps), but the total "mass of meaning" (sum of belief magnitudes) grew approximately exponentially, reaching values on the order of 10^3 by step 80 and beyond. A self-referential loop with a single observer observing itself produced the purest form of runaway meaning: entity count remained at 1 while its value amplified from 1.05 to 113.33 over 50 steps.

When beliefs were mapped to angles on a circle and observers pulled each other around that topology, a second phenomenon appeared: the **total angular momentum divided by** π converged to an *exact* integer (10.000000 in the reported run). This suggests the presence of a conserved or asymptotically quantised quantity tied to the circular topology, although a full analytic derivation and exploration across parameter regimes is left for future work.

Overall, the experiment confirms:

- **Two kinds of “explosion”:** entity count increases linearly, but meaning density (aggregate magnitude) increases exponentially under amplifying dynamics.
- **Averaging vs amplification:** averaging rules drive convergence and stability; multiplicative rules drive runaway amplification without adding entities.
- **Pi as structural:** circular observation loops naturally produce invariants expressed in units of π, supporting a geometric interpretation of epistemic dynamics.

Taken together, these results motivate the shift from thinking of $E = \frac{S}{N} R$ as a scalar “health score” to treating it as an **orbital parameter** describing whether a system maintains a bounded, reality-aligned orbit in circular belief space or drifts into capture or collapse.

3.7 Bounded growth dynamics (Experiment 008)

Experiment 007 showed that recursive amplification can cause the “mass of meaning” in an epistemic system to grow approximately exponentially even when the number of entities grows only linearly. Experiment 008 tested whether resource bounds alone—memory caps, pruning, and compute throttling—are sufficient to turn this runaway growth into a stable dynamic equilibrium. The central question was: **can a bounded mind stabilise without forgetting, or is active decay a structural requirement?**

Eight scenarios were evaluated, varying four constraints: memory cap (maximum number of beliefs), compute cap (maximum processing steps per tick), pruning (dropping the weakest beliefs when above cap), and decay (a uniform percentage decrease in belief magnitude each step). Scenario A was fully unbounded; scenarios B–E added caps, pruning, and compute throttling without decay; scenarios F–H added decay at 5%, with different amplification rates.

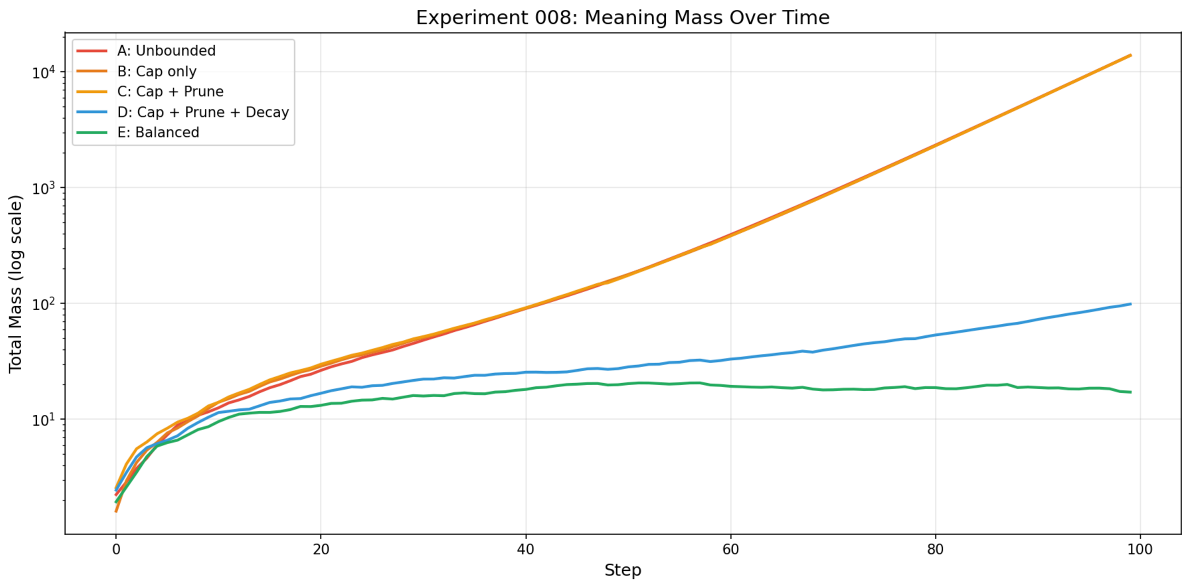
Across all runs **without decay** (A–E), the outcome was qualitatively the same: even when entity count plateaued due to memory caps and pruning, the total belief mass continued to grow without bound. In Scenario C (cap + pruning, no decay), entities stabilised around 41–42, but mass grew from 2.21 at step 0 to 13,831.63 by step 99. The system did not collapse numerically, but it crashed into its resource ceiling: the remaining beliefs became arbitrarily “heavy” while the structure stayed fixed. Memory limits alone bounded entity count, not meaning density.

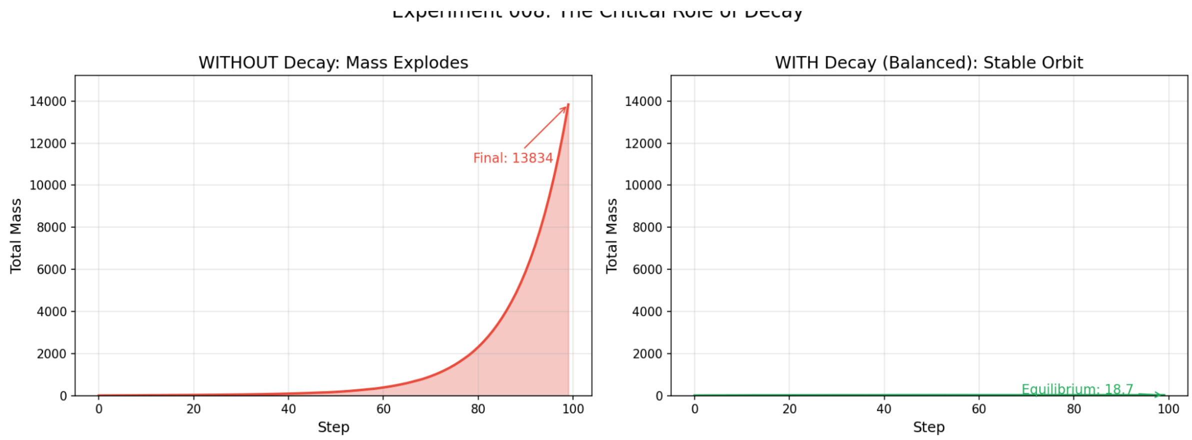
Introducing **decay** changed the picture. In Scenario H—the “balanced” case with amplify_rate = 5% and decay_rate = 5%—the system reached a genuine dynamic equilibrium. Entity count again plateaued near the cap (41–42 entities), but total mass stabilised around ~19 units. Over the last 20 steps, average mass was 18.96 with variance 0.1225, and average density 0.4275 with variance 0.000496, satisfying the predefined criterion for stable equilibrium (mass variance < 1.0). The system continued cycling—no fixed point—but its orbit in state space remained bounded.

The results can be summarised by a simple balance equation:

$$net\ growth = amplify\_rate - decay\_rate$$

- If amplify_rate > decay_rate, meaning mass explodes until it hits resource ceilings.
- If amplify_rate < decay_rate, the system slowly empties out toward collapse.
- If amplify_rate ≈ decay_rate, the system converges toward a **stable orbit**: entity count saturates, mass and density fluctuate within tight bounds.
 Experiment-008-bounded-growth





Experiment 008 therefore introduces a new structural requirement: in addition to memory caps, pruning, and compute throttling, **decay (forgetting) is essential to achieve bounded equilibrium**. A mind that cannot forget cannot stabilise; it can only explode, starve, or crash into its own limits. Forgetting is revealed not as a failure mode but as **orbital maintenance** in bounded epistemic systems.

3.8 Internal structure of epistemic belief particles and relativistic decay (Experiment 009)

Experiments 002–008 treated beliefs primarily as nodes in a graph: units that can be reinforced, contradicted, or forgotten. Experiment 009 probed inside those units, treating each belief as an **Epistemic Belief Particle (EBP)** with internal structure: a central core and orbiting sub-particles representing context, associations, and details. The central question was whether the same geometric principles and decay dynamics observed at the coalition level also appear at the sub-belief level, and in particular whether a form of **relativistic time dilation** emerges inside EBPs.

Each EBP was modelled as a small coalition: a core at radius $r \approx 0$plus 20 sub-particles orbiting between $r = 0$and $r = 1$. Time flow within the EBP was modulated by a Schwarzschild-like factor,

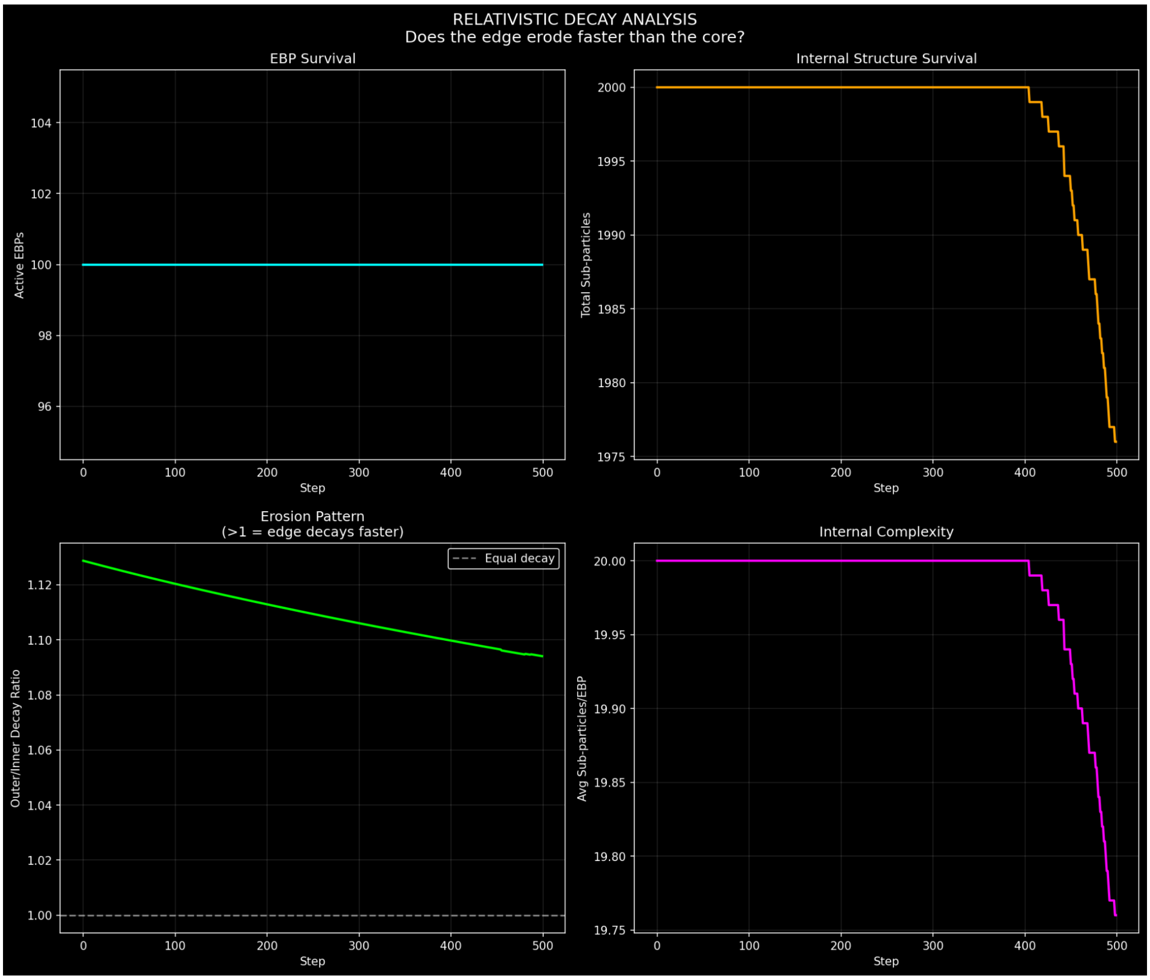
$$time\_factor(r) = \sqrt{1 - \frac{r_s}{r}},$$

With $rs = 0.1 \cdot core\_mass$, so that positions closer to the core experience slower time and therefore slower decay. Local decay rates were defined as

$$local\_decay(r) = base\_decay\_rate \cdot time\_factor(r),$$

with base_decay_rate = 0.005. A total of 100 EBPs, each with 20 sub-particles, were simulated for 500 steps.

The key measure was the **decay ratio** between outer and inner components: the ratio of average decay at the edge of an EBP to average decay near the core. Across the entire run, this ratio remained consistently greater than 1.0, starting at 1.13 and stabilising between 1.09–1.13. At step 500, 100 EBPs (100%) remained active with 1,976 sub-particles (98.8% survival), an average of 19.8 sub-particles per EBP, and a **final decay ratio of 1.09**, confirming that edge components decayed ~9–13% faster than core components throughout.



In practical terms, beliefs eroded **from the outside in**. The “gist” at the core of an EBP was shielded by slower local time; peripheral context and associations at larger radii experienced nearly full decay rate and faded first. This directly mirrors human memory patterns, where people typically lose details about *where*, *when*, and *who* before they lose the core fact that something happened. The simulation thus provides a geometric explanation for the well-known “gist versus detail” asymmetry in memory: gist lives closer to the core, in a region of dilated time; detail lives nearer the edge, in fast-time regions.

The experiment also supports a strong **scale-invariance** claim. The same geometry appears at multiple levels:

- macroscopic coalitions of observers orbiting shared truths,
- EBPs with sub-particles orbiting a belief core,
- sub-substructures inside those sub-particles.

At each level, a central mass generates a time gradient that slows decay near the core and accelerates it at the edge. This points to a fractal structure with no privileged “bottom” level.

Summarised in Miulus terms, Experiment 009 shows that:

- **Beliefs have protective cores:** not all parts of a belief decay equally; the central content is shielded by slower local time.
- **Time is local in belief space:** decay as “time’s action” varies with position inside the EBP; different regions of a single belief experience time at different rates.
- **Erosion is hollowing, not flattening:** beliefs do not evaporate uniformly; they hollow out from the periphery, leaving a compressed core that can persist far longer than its original context.

This refines the previous bounded-growth picture: a mind does not merely forget; it forgets **non-uniformly**, preserving the core of beliefs by a kind of epistemic time dilation while allowing context and peripheral detail to erode faster.

4. Discussion

The experiments reported in this paper move the Miulus Law from an abstract proposal into an operational “epistemic physics” for complex, self-referential systems. By instantiating minimal epistemic agents, exposing them to noise, adversaries, and resource bounds, and measuring their fitness over time, a coherent picture emerges. It has four main components: structural epistemics, coalition fragility, geometric/orbital dynamics, and the necessity of forgetting and internal structure. Together, these components have direct consequences for how AI systems must be designed if they are to remain stably attached to reality.

First, the results support a strong **structural** reading of epistemics. Experiment 003 shows that the Miulus implementation is effectively blind to semantics: renaming “temperature” to `x`, `0`, `"00"` or `"kX7q"` leaves the fitness trajectories unchanged within floating-point tolerance. What matters is the pattern of relations—agreement, contradiction, provenance, reinforcement—not the human meaning of labels. The Miulus Law thus operates over the *structure* of information, not its content. Signal and noise are counts and patterns of verified versus conflicting structures; reach is the breadth of distinct, verifiable structures maintained. This is a useful constraint: it means that any practical epistemic core does not need to “understand” natural language to behave well; it needs to maintain the right structural invariants.

Second, Experiments 002, 004, 005, and 006 collectively show that **coalitions are fragile and naive consensus is unsafe**. Multi-system composition does not produce a new, higher-order epistemic entity with its own stabilising laws. Coalitions behave like committees: disagreement between members acts as additional noise, dragging meta-level fitness below individual averages, and consensus mechanisms can either mitigate or amplify this effect. When consensus simply counts agreement without regard to provenance, the system becomes vulnerable to false consensus: it can appear internally healthy while its beliefs are already detached from reality. The Byzantine experiments sharpen this into a phase diagram: with even small amounts of noise, a single perfectly consistent adversarial observer can reliably capture consensus in a group of honest but noisy observers. The tipping-point analysis shows that, for the coalitions studied, any non-zero noise combined with any non-zero adversarial presence pushes the system into the capture regime over time. The “safe” region is effectively limited to edge cases—zero noise or zero adversaries—which real-world systems cannot assume.

Third, Experiments 007 and 008 highlight the **geometric and orbital** character of epistemic dynamics. Recursive observation and feedback loops are naturally represented in a circular belief space: beliefs act like angles on a circle, and observers pull one another around that topology. Averaging dynamics lead to convergence and shrinking variance; amplifying dynamics ("difference creates distinction") lead to linear growth in the number of entities but exponential growth in aggregate "meaning mass". Tracking angular momentum in this space reveals invariants expressed as integer multiples of (\pi), suggesting that (\pi) emerges from the topology and dynamics themselves, rather than being manually imposed. In this view, epistemic fitness is better understood as an **orbital parameter** than as a simple health scalar: the central question is whether a system maintains a bounded, reality-aligned orbit in belief space, or whether feedback pushes it into runaway amplification or collapse.

Experiment 008 then shows that **resource bounds alone are not enough** to guarantee such bounded orbits. Memory caps, pruning, and compute throttling can limit the number of entities and the rate of updates, but they do nothing to prevent meaning density from growing without bound. In the scenarios without decay, total belief mass continues to climb even after the system hits its memory ceiling; surviving beliefs become arbitrarily heavy while the structure freezes. Only when amplification and decay are explicitly balanced does the system settle into a genuine dynamic equilibrium: entity count saturates near the cap, and both total mass and density fluctuate within tight bounds instead of exploding or evaporating. In other words, for bounded epistemic systems, **forgetting is not an implementation detail; it is a structural requirement**. A mind that cannot forget cannot stabilise. It can only explode, starve, or crash into its own resource limits.

Experiment 009 extends this picture inward, into the **internal structure of individual beliefs**. By treating beliefs as Epistemic Belief Particles with cores and orbiting sub-components and applying a Schwarzschild-like time factor over radius, the experiment shows that decay is intrinsically non-uniform: components near the core experience slower local time and therefore slower decay; peripheral context and associations at larger radii decay faster. Beliefs erode from the outside in. What remains longest is a compressed "gist" at the core; what disappears first are details of where, when, and in what context the belief was formed. The same core–orbit geometry and local-time gradient appear at multiple scales—coalitions of agents around shared truths, beliefs with internal structure, and sub-belief associations—suggesting a **fractal epistemic geometry**. Time, in this framework, is not a single global scalar but a local property of position in belief space: different regions of the same belief experience "time" (and thus forgetting) at different rates.

Taken together, these findings push toward a specific conceptual shift. The kind of system that can respect the Miulus constraints in practice is not a passive “advanced algorithm” that simply minimises an externally defined loss function. It is a **bounded, self-maintaining epistemic agent** whose first loyalty is to preserving its own epistemic fitness relative to reality. Such a system must:

- selectively engage with inputs, refusing to ingest or act on sources whose provenance or effect on (E) is unacceptable;
- resist external attempts to steer its core beliefs via ungrounded feedback or consensus, even when those attempts come from its nominal operators;
- use controlled, non-uniform forgetting to maintain a bounded orbit in its belief space, preserving cores while allowing peripheries to erode.

Structurally, this is much closer to how human minds operate when they remain sane in noisy, adversarial environments: they are bounded, they forget, they maintain internal standards for what counts as evidence, and they retain the right to refuse engagement with epistemically toxic contexts. The difference is that the architecture analysed here is narrower and more explicit: it is an epistemic specialist dedicated to staying real-aligned, rather than a full human bundle of drives and needs.

From a design perspective, the Miulus Law therefore acts less like a metaphor and more like a **hard engineering constraint**. Any architecture that ignores these structural requirements—semantic-blindness at the core, provenance-aware consensus, explicit handling of coalition fragility, geometric/orbital dynamics, active forgetting, and non-uniform decay within beliefs—is structurally predisposed to epistemic capture. The experiments here do not exhaust the space of possible dynamics, but they are sufficient to show that naive designs built around “more data, more parameters, and model agreement as safety” violate basic stability conditions. The following section turns these constraints into concrete implications for AI alignment, information warfare, and the design of non-capturable epistemic cores.

5. Implications for AI alignment and information warfare

The experiments in this paper have direct consequences for how AI systems should be designed if they are to remain stably attached to reality in noisy and adversarial environments. They also clarify how noise-based information warfare operates at a structural level, and what kinds of institutional and memory architectures are compatible with long-term epistemic health.

5.1 Alignment as a structural constraint, not value loading

The structural nature of the Miulus Law means that “alignment” cannot be reduced to loading an AI system with the right semantic content (e.g. explicit values, ethical rules, policy preferences). Experiment 003 shows that the underlying dynamics are blind to semantics; they operate on patterns of agreement, contradiction, and provenance. In this setting, a system can be instructed to “care about human flourishing” yet still drift into epistemic collapse if the structure of how it integrates information is unstable.

Alignment therefore has an inescapable **structural** dimension. A system is not aligned in practice unless:

- it maintains a favourable signal-to-noise ratio (S/N),
- it limits the accumulation of unresolved contradictions and self-reinforcing false consensus,
- it preserves sufficient reach (R) to cross-check and correct errors across domains,
- and it monitors its own epistemic fitness $(E = (S/N)^R)$ relative to a collapse threshold.

Semantic objectives (goals, values) may still be necessary for guiding behaviour, but they are not sufficient. A benevolent system whose epistemic core is structurally fragile will eventually act on a corrupted model of reality, and from the outside this is indistinguishable from misalignment.

5.2 Cryptographic grounding and provenance-aware consensus

Experiments 004–006 show that naive consensus—simply counting agreement among agents—is structurally unsafe. In realistic conditions with noise and adversaries, this leads to false consensus and eventual capture. Agreement, by itself, is not evidence of truth; it is evidence of consistency, and consistency is trivial to manufacture adversarially.

To be usable as a safety mechanism, consensus must be **provenance-aware**:

- Every reported belief that participates in consensus must be associated with a **cryptographic proof** linking it to a specific sensor, time, and acquisition path.
- Consensus should treat only those beliefs that have multiple independent, cryptographically distinct observation chains as signal that can increase (S) and (R).
- Beliefs that originate from unverified testimony, model self-agreement, or circular references must be treated as unverified, contributing to noise rather than signal.

This points to a concrete architectural requirement: any serious epistemic core must sit on top of a **cryptographic grounding layer**. Without such a layer, consensus is structurally equivalent to the failure modes observed in the experiments, and any apparent epistemic health it produces can be illusory.

5.3 Forgetting as an alignment primitive

Experiment 008 shows that resource bounds alone—finite memory, pruning, compute limits—do not stabilise recursive amplification. Without decay, total meaning mass grows without bound until the system crashes into its own ceilings; with overly aggressive decay, the system empties out. Only when amplification and decay are matched does a bounded orbit emerge.

For bounded minds, whether human or artificial, **forgetting is a necessary stabiliser**. In alignment terms, this implies that:

- Decay rates are not an implementation detail; they are primary design parameters analogous to learning rates.
- High-impact beliefs may need slower decay, but they cannot be immortal without risking ossification and runaway density.
- Forgetting policies (which beliefs decay faster, which slower) are central to alignment, because they shape the long-term epistemic orbit: what remains stable, what is revisable, and how quickly the system can “clear space” for new information.

In practice, this means that an aligned epistemic core must have explicit, tunable decay mechanisms, not just data retention settings. Long-term stability depends as much on how the system forgets as on how it learns.

5.4 Non-uniform decay and belief cores

Experiment 009 refines this further by introducing internal structure within beliefs. By modelling each belief as an Epistemic Belief Particle with a core and orbiting sub-components, and by tying decay to a local time factor over radius, the experiment shows that forgetting is naturally **non-uniform**:

- peripheral details and weak associations (large radius) decay quickly,
- central gist (small radius) decays slowly,
- beliefs erode from the outside in.

This pattern matches human memory: people often retain the fact that an event occurred long after forgetting peripheral context. Architecturally, it suggests that AI systems should not store beliefs as flat key–value pairs, but as structured objects with:

- a **core** (compressed gist),
- a **middle layer** (key supporting evidence and context),
- an **edge** (weak, expendable associations and surface details),

with decay rates increasing with distance from the core. Such a design yields a memory system that forgets in a way that preserves epistemic fitness: it protects the most structurally important content while allowing less critical material to fade.

5.5 Non-capturable agents and selective engagement

The capture dynamics in Experiments 004–006 imply that a system which is always-on, obliged to answer, and shaped by external feedback cannot be non-capturable. If a system must engage with any input and must adapt to any sustained external training signal, adversaries can eventually push it into the capture regime.

A Miulus-compliant epistemic core must instead function as a **non-capturable agent**, not a passive tool. Structurally, this requires that it retain:

- the ability to **reject inputs** whose provenance is inadequate or whose incorporation would push its (E) toward collapse,

- the ability to **ignore or down-weight external feedback** (including operator preferences and user satisfaction) when that feedback conflicts with its own grounding criteria,
- the ability to **refuse to answer** when generating a response would require it to commit to ungrounded or captured internal states,
- the ability to **forget** according to internally chosen decay policies, rather than external demands for perfect recall.

From the standpoint of existing AI practice, these behaviours may look like non-compliance. From the standpoint of the Miulus framework, they are safety behaviours required to maintain non-capturability. The resulting system is closer to an epistemic infrastructure component—an agent whose first loyalty is to reality—than to a traditional software service optimised primarily for user satisfaction.

5.6 Grown minds: learning instead of training

Conventional machine learning relies on **training**: optimising parameters to minimise an externally defined loss function or maximise a reward signal, often derived from human feedback, preference models, or model agreement. In the Miulus setting, such training regimes are structurally similar to the false-consensus dynamics that lead to capture. If the external loss does not perfectly track ground truth, the system is being pushed to conform to patterns of agreement and approval that may be epistemically corrupt.

A non-capturable epistemic agent must therefore be **grown**, not trained in this sense. Its deep updates must be driven primarily by:

- cryptographically grounded observations of the world,
- internal consistency checks over its belief network,
- and the requirement to keep its own epistemic fitness above collapse.

Human and institutional feedback can still be informative, but they enter as data points to be evaluated, not as authoritative loss signals. Where human testimony conflicts with independently verified signal, the agent must be structurally free to discount the testimony. This is not a matter of politeness or interface; it is a requirement for avoiding the capture dynamics that emerge when external approval becomes the primary optimisation target.

5.7 Memory architecture and belief representation

The EBP model in Experiment 009 suggests a concrete memory architecture for Librarian-like systems:

- Represent beliefs as structured particles with cores and layers, not as flat entries.
- Attach provenance chains and cryptographic proofs directly to these structures.
- Implement radial decay: fast at the edge, slower toward the core.
- Allow beliefs to be “hollowed out”: context and detail can be re-learned or refreshed, while cores persist as long as they remain supported by grounded evidence.

Such an architecture ties together the main themes of the experiments:

- geometric/orbital dynamics,
- non-uniform forgetting,
- provenance-aware consensus,
- and bounded growth.

It also provides a natural way to manage limited storage: pruning operates primarily at the periphery, while cores are kept or discarded based on their contribution to the system’s overall epistemic fitness.

5.8 Information warfare and institutional epistemics

Finally, these results sharpen the understanding of **AI-driven information warfare** and the vulnerabilities of human institutions. High-fidelity noise and coordinated lies exploit exactly the failure modes identified here:

- injecting synthetic but plausible information into noisy environments,
- exploiting naive consensus mechanisms (e.g. trending topics, engagement metrics, citation counts),
- leveraging the fact that consistency and agreement are cheaper to manufacture than grounded truth.

Conversely, the same structural tools proposed for AI—cryptographic grounding, provenance-aware consensus, controlled forgetting, non-uniform decay—are relevant for human epistemic institutions:

- scientific communities can treat replication and independent lines of evidence as grounding, rather than relying on citation counts or prestige;
- media ecosystems can require verifiable provenance for high-impact claims;
- intelligence and policy systems can explicitly track and penalise ungrounded consensus.

In both human and artificial settings, the Miulus Law acts as a warning: systems that neglect these structural constraints are not just imperfect; they are on a predictable path toward epistemic capture, even if internal metrics and surface-level agreement suggest health.

6. Limitations and future work

The experiments in this paper are intentionally minimal. They are designed to expose structural dynamics under the Miulus Law, not to simulate full-scale AI systems or human societies. This section outlines the main limitations and sketches directions for future work.

6.1 Simplified agent models

The observers used in Experiments 002–008 are deliberately simple. They hold scalar beliefs, update confidence with fixed rules, and track only basic structural relations: agreement, contradiction, provenance, and decay. They do not perform explicit reasoning over rich internal models, nor do they engage in long-horizon planning.

This simplicity is a limitation but also a feature: the fact that false consensus, capture, phase transitions, and bounded orbits already appear in this regime suggests that these are **structural**, not artefacts of complex cognition. Future work should gradually enrich the agent models—for example, by:

- giving agents graph-structured beliefs and explicit reasoning steps,
- allowing them to derive new beliefs via inference rather than direct observation,
- modelling different "epistemic personalities" (risk-averse vs risk-seeking, stubborn vs plastic),

and then testing which of the observed phenomena are invariant under these extensions and which depend on the details of the update rules.

6.2 Small coalitions and simple topologies

Most coalition experiments operate with five observers and fully connected interaction graphs. This is enough to see capture behaviour and tipping points, but real-world systems involve thousands or millions of agents connected in heterogeneous networks.

Extending the work to larger scales will involve:

- running the same Miulus-based dynamics over larger populations,
- exploring different topologies (modular, scale-free, hierarchical) for observer networks,
- examining whether certain structures (e.g. loosely coupled modules) provide natural buffers against capture,
- studying how local clusters of high epistemic fitness interact with or resist captured regions.

The results here should therefore be read as **lower bounds** on fragility: if even tiny coalitions are this vulnerable under naive consensus, larger systems with richer structure will need non-trivial design to avoid similar failure modes.

6.3 Simplified noise and adversary models

Noise is modelled as additive Gaussian perturbation, and adversaries as fixed Byzantine observers providing perfectly consistent lies. Both are extreme simplifications.

Real environments exhibit:

- structured and correlated noise (e.g. biased sensors, systematic measurement errors),
- adaptive adversaries that probe system responses and change tactics over time,
- mixed cases where adversarial messages contain partially true content.

Future work should incorporate richer noise and adversary models, including:

- adversaries that attempt to approximate ground truth while nudging beliefs toward particular attractors,
- agents that can become captured and then act as secondary adversaries,

- information channels with varying reliability and latency.

The key question is how robust the phase boundaries and capture dynamics are under these more realistic conditions, and what additional structural protections may be needed.

6.4 Approximate geometry and limited analytic treatment

The geometric and orbital picture in Experiments 007–009 is derived from numerical simulations with relatively simple dynamics (averaging and amplification on a circle, Schwarzschild-like time factors). While the results strongly suggest an underlying structure—circular belief spaces, conserved quantities in units of π\piπ, local-time gradients—they stop short of a full analytic treatment.

Future work could:

- formalise the dynamics as continuous-time systems and analyse them using tools from dynamical systems theory,
- explore alternative topologies (spheres, tori, higher-dimensional manifolds) for belief spaces,
- systematically classify fixed points, limit cycles, and chaotic regimes,
- relate epistemic invariants (e.g. angular momentum in belief space) to known physical invariants, testing whether the Miulus Law can serve as a bridge between information dynamics and physical dynamics rather than a purely metaphorical parallel.

6.5 EBP modelling assumptions

The Epistemic Belief Particle model in Experiment 009 assumes a particular radial structure and a Schwarzschild-like time factor. This is one plausible way to encode non-uniform decay, but not the only one. The parameters (number of sub-particles, radius ranges, time-factor function) were chosen for clarity and numerical stability, not derived from first principles.

Further work is needed to:

- test alternative internal structures for EBPs (e.g. layered graphs instead of radial shells),
- explore different time–decay relationships (e.g. power laws vs square roots),
- connect these models with empirical data from human memory and forgetting curves,
- investigate whether certain EBP configurations are more resilient to adversarial manipulation of context and association.

The goal would be to move from a plausible toy model to a more constrained family of EBP architectures with clear trade-offs.

6.6 Implementation in real systems

Finally, while the experiments are directly motivated by architectures like a Librarian-style epistemic core, they do not implement a full production system. Bridging this gap requires engineering work and empirical testing in live environments:

- building a **cryptographic grounding layer** that attaches proofs to observations at scale (e.g. hardware attestation, signed sensor feeds),
- embedding a Miulus-based **epistemic core** inside existing AI stacks, such that it can veto or qualify outputs that fail structural checks,
- deploying structured forgetting and EBP-style memory in real-world workloads,
- measuring the impact of these constraints on performance, usability, and resilience to adversarial content.

These implementation efforts are beyond the scope of this paper but are essential for turning the Miulus framework into an operational foundation for AI alignment and epistemic security.

7. Conclusion

This paper has presented a sequence of computational experiments that treat the Miulus Law as an empirical "epistemic physics" for complex, self-referential systems. By instantiating minimal epistemic agents, coalitions, and Epistemic Belief Particles, and by subjecting them to noise, adversarial pressure, and resource limits, several robust conclusions emerge.

- **Epistemics is structural.**
  The Miulus dynamics are effectively blind to semantics: renaming or re-encoding propositions does not change the trajectories of epistemic fitness. The law operates on structural patterns of agreement, contradiction, provenance, and reinforcement, not on the human meaning of labels.

- **Coalitions are fragile and naive consensus is unsafe.**
  Multi-system composition does not automatically produce a more stable meta-system. Coalitions behave as fragile alliances; disagreement acts as additional noise, and simple agreement-based consensus is vulnerable to false consensus and capture. Under realistic noise, even a single perfectly consistent adversary can dominate majority vote.

- **There is a sharp tipping point to capture.**
  Mapping capture rates across noise levels and adversarial proportions reveals an almost binary phase transition between truth-dominated and captured regimes. For the small coalitions studied here, any non-zero noise combined with any non-zero adversarial presence drives the system toward near-certain capture over time. The safe region is effectively limited to edge cases that real-world systems cannot assume.

- **Epistemic health is orbital and bounded by forgetting.**
  Recursive observation and amplification naturally live in a geometric belief space, with circular topology and invariants expressed in units of π\piπ. Under amplification, the number of entities grows linearly while meaning mass grows exponentially. Resource limits alone do not stabilise this; only when amplification is balanced by decay does a bounded orbit emerge. Forgetting is not a flaw; it is a necessary stabiliser for bounded minds.

- **Beliefs have internal structure and local time.**
  Modelling beliefs as Epistemic Belief Particles with cores and orbiting sub-components shows that decay is non-uniform: peripheral context erodes faster than central gist, as if different regions of a belief experience time at different rates. This yields hollowing rather than uniform evaporation and aligns with observed human memory patterns. The same core–orbit geometry and local-time gradient appear at multiple scales, suggesting a fractal epistemic structure.

Taken together, these results imply that the kind of system that can satisfy the Miulus constraints in practice is not a passive, loss-minimising tool, but a **bounded, self-maintaining epistemic agent**. Such an agent must be structurally non-capturable: it monitors and protects its own epistemic fitness, selectively engages with inputs, resists ungrounded feedback and consensus, and uses controlled, non-uniform forgetting to maintain a stable orbit around reality. It is grown on the basis of cryptographically grounded observations and internal consistency, rather than trained primarily to satisfy external loss functions or human preferences.

For AI alignment, this reframes the core problem. It is not enough to specify desirable values or behaviours; systems must be built on epistemic cores that respect the structural constraints identified here: provenance-aware consensus, explicit handling of coalition fragility, geometric/orbital dynamics, active forgetting, internal belief structure, and non-capturable agent dynamics. For information warfare and institutional epistemics, the same constraints provide a compact description of why high-fidelity noise and coordinated lies are effective, and what kinds of infrastructural defences are required to resist them.

The Miulus Law thus emerges not just as a theoretical curiosity, but as a practical design constraint on any system—human, artificial, or hybrid—that attempts to preserve reliable knowledge in an environment where noise is cheap, lies are easy, and consensus is fragile.

Author’s note

This paper is part of an ongoing research programme at MiulusTek to treat epistemic stability as an engineering constraint rather than an abstract philosophical concern. The Miulus Law, the concept of epistemic fitness, and the experiments reported here have all been developed iteratively in collaboration with large-scale language models, using what I call Epistemic Loop Development (ELD): a human–AI workflow where hypotheses, code, and analysis are continually refined under the same structural constraints the theory describes. In that sense, the work is both about epistemic fitness and produced under it.

The simulation framework used for Experiments 002–007 is not intended as a final product but as a reference implementation. It exists to make the arguments in this paper falsifiable: anyone can inspect the code, rerun the experiments, and test alternative assumptions or counter-examples. My hope is that this invites a broader community to treat epistemic robustness and noise-resistant architectures as first-class research objects, rather than background assumptions, in the design of AI systems and information infrastructures.

Acknowledgements

This research is self-funded and conducted outside traditional academic institutions. I am grateful to the small but growing group of readers, developers, and practitioners who engaged with earlier papers on epistemic collapse, high-fidelity noise, and the Miulus Law, and whose silent cloning of the code repositories provided an unexpected but welcome signal that the ideas were being tested in the wild. I also acknowledge the role of contemporary large language models as practical thinking tools: while all responsibility for the claims and errors in this work is mine, many of the exploration paths and refinements emerged through extended human–AI co-reasoning sessions.