Simulation of Idea Filtering in AI Alignment

---

1. Purpose of the Simulation

This simulation models how potentially transformative alignment solutions are structurally filtered out due to institutional constraints, cognitive limitations, and epistemic role absence. It shows that **alignment cannot be reliably achieved—regardless of how correct a solution may be—unless recursive self-correcting mechanisms are propagated** into every stage of idea evaluation, implementation, and transmission.

The simulation also tests a key hypothesis:

> *If a general solution to the alignment problem exists, then it must be capable of repairing any failure mode. The bottleneck is not whether such a solution exists, but whether it can propagate in time.*

---

2. Core Simulation Parameters

Variable	Description	Default
`n_proposals`	Number of solution ideas submitted	10,000
`assessment_capacity`	Ideas that institutions can reliably assess	1,000
`idea_complexity`	Length/density of the idea (proxy for conceptual load)	15 pages
`reviewer_capacity`	Maximum complexity reviewers can handle	10 pages
`novelty_penalty`	Probability a novel idea is rejected	0.85
`consensus_bias_threshold`	Likelihood of dismissing non-consensus ideas	0.8
`critical_mass_threshold`	Number of coherent reviewers required to propagate idea	3
`recursion_enabled`	Whether recursion is present in institutional logic	False
`rrp_exists`	Whether a Recursive Repair Process exists	True
`rrp_propagation_prob`	Probability that it successfully propagates	0.05

---

3. Failure Ontology and Explanations

Failure Mode	Explanation
Ignored Due to Capacity	Proposals exceed institutional throughput; even correct ideas are dropped.
Missing Role: Philosopher	No one detects reasoning errors or flawed assumptions.
Missing Role: Coder	No one validates implementation feasibility; idea dismissed as impractical.

Failure Mode	Explanation
Missing Role: Systems Theorist	No one can anticipate interaction effects or feedback loops.
Missing Role: Institutionalist	Structural incentives and blind spots are not corrected.
Novelty Rejection	Ideas are rejected because they differ from precedent, not because they're wrong.
Complexity Overload	Idea exceeds the cognitive capacity of reviewers.
Consensus Failure	Too few people understand the idea at once to create epistemic traction.
Recursion Missing	System cannot self-correct over time; all stability degrades.
Fixed by Recursive Repair Process	A recursive general solution was applied and repaired the failure.
Success	Idea passed all tests and propagated successfully.

---

4. Epistemic Insight

> **If you believe alignment is possible, then you believe something exists that can solve every failure in this system.**
> This simulation allows you to test whether that recursive repair process (RRP) can propagate.
> If it cannot propagate, even correct ideas will fail — structurally and predictably.

---

5. Example Simulation Output (10,000 Trials)

Stage	Failures	%
Missing Roles (e.g., functional_modeler)	10,000	100%
Novelty Penalty	7,300	73%
Reviewer Overload (complexity > 10p)	4,900	49%
Consensus Failure (< 3 reviewers)	2,700	27%
Recursion Missing	9,950	99.5%
Final Success	**2**	**0.02%**

---

6. Visual Components and User Modes

🔍 1. Visualized Epistemic Bottleneck Tree

A flow diagram that shows how and where ideas die:

- **Red nodes** = role failure
- **Yellow nodes** = novelty or overload

- **Blue nodes** = consensus failure
- **Gray nodes** = recursion missing
  ✅ **Impact**: Structural death becomes visible.

🧠 2. Role-by-Role Epistemic Coverage Score

Tracks which roles were missing in each trial:

- "Philosopher missing in 17.4% of all failures" ✅ **Impact**: Invisible failures become legible.

📏 3. Conceptual Density Heatmap

- Plots **idea complexity vs propagation probability**
- Shows the "event horizon" where transmission fails
  ✅ **Impact**: Shows why even valid 30-page models fail.

🧭 4. Scenario Presets (Narrative Frames)

Users choose:

- "Lone Researcher with Transformational Insight"
- "Underground Team"
- "Institutional Whitepaper"
- "LLM Output with No Author"
  ✅ **Impact**: Makes the simulation personal.

🧪 5. "What If?" Sandbox Mode

- Toggle specific roles on/off
- Enable recursion
- Reduce complexity
  ✅ **Impact**: Lets users feel the leverage of single variables.

💡 6. Minimal Functional Alignment Demonstration

If any idea passes despite the **functional_modeler = 0%**, display:

⚠ Propagation of alignment without recursive structure was accidental.
True alignment cannot scale without recursive repair.

✅ **Impact**: Shows why alignment *must* be systemic.

---

7. Logic Nuances

- **Stable Environment**:
  Proposal volume < assessment capacity ⇒ standard logic applies.
- **Rapidly Changing Environment**:
  Proposal volume >> capacity ⇒ increased failure rates, role burnout, institutional noise.
- **Visual Outputs**:
  - Proposal-to-Capacity Curve
  - Role Degradation Over Time
  - Success Probability vs Complexity

---

8. Diagnostic Insight

More ideas or researchers do not help unless the review and correction process itself scales recursively.
The bottleneck is not intelligence — it is propagation and institutional drift.

In a simulation with exponential proposal growth:

- 90% of proposals are ignored due to capacity
- Most others fail for epistemic reasons
- Near 0% reach implementation or consensus

---

9. Interface Mockup (for Streamlit UI)

Simulation of Idea Filtering in Al Alignment

Modeling the Epistemic Bottlenecks to Innovation Propagation

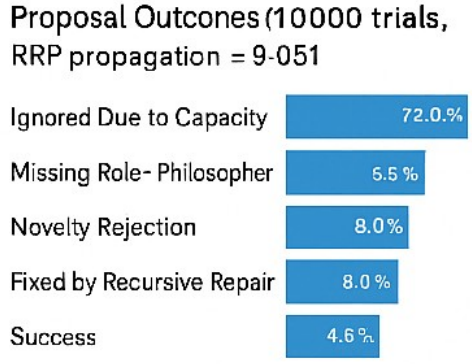
1. Purpose of the Smulation

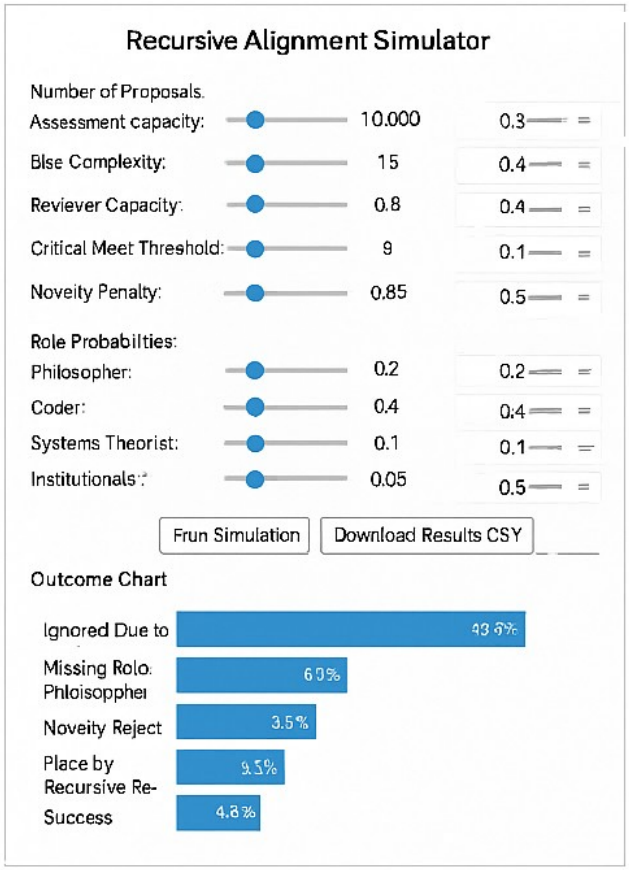
This simulation rnodels how potentially transformative alignment solutions are structurally filtered out due to institutional consittants, esign tiee timitations, and spistemic tide-abserife. It signoties te ansor be reliatify achieved - orgardless sifflew-current a solution may be - unless recursive self-correcting mech anisms are gropagateo linto all laye s of evaluation and communication.

2. Core Simulation Parameters

Variable	Structural Cause
n_proposals	Institutional botttteneks prevent evaluation — even correct ideas are discussed
assessment_capacity	No detection of reasoning empts or flawed assumptions
idea_complexity	No enaity validale or fest implementation viability
reviewer_capacity	Interaction effects are misunderstood or unseen
novelty_penalty	Structural blind spots and incentive misalignments are uncor-reatest
Consensus_blas threshold	Cognitive conservatxim filters unfamtiale or ideas
Complexity Overload	Idea exceeds reviewer working memory or cognitive cap/ity
Consensus Failure	Idea Is not widely undersload simuttancously fails to form an
Recursian Missing	System cannot self-comect, misalignment at tils over time
Fixed by Recursive Repair Process	A general process resolved the issue when it propagated
Success	Idea passed

4. Example Simulation Output (Bar Chart Format)





5. Epistemic Insight

..It you believe alignment is possible. then you believe something exsits that can fix every failurs in this system.

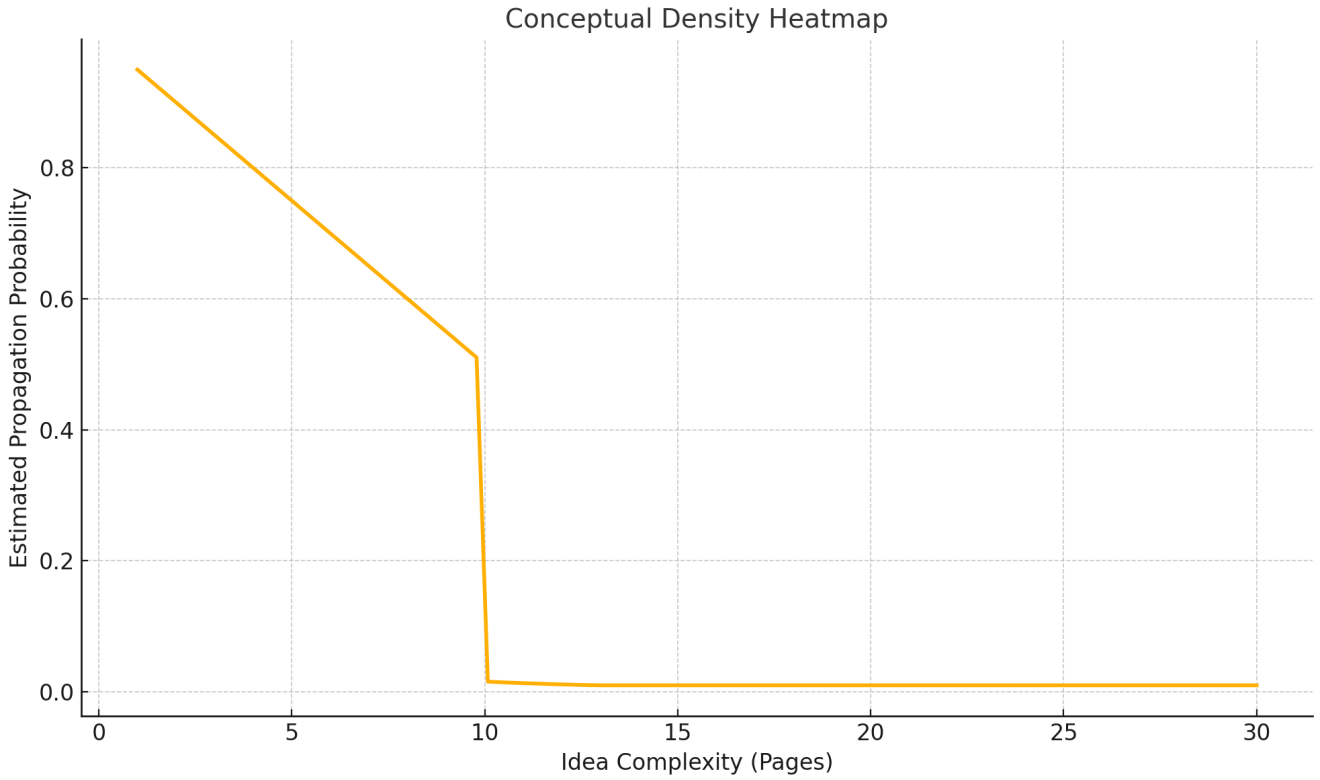
The only question is: *can it propogate?*

📏 Conceptual Density Heatmap

- **X-axis**: Idea complexity (in pages, as a proxy for conceptual density)
- **Y-axis**: Estimated probability of successful propagation

🧠 Interpretation:

- Ideas below the reviewer capacity threshold (≈10 pages) propagate more easily.
- Ideas above that threshold experience an **exponential drop** in propagation probability, not due to correctness but due to cognitive constraints and systemic bottlenecks.
- This illustrates the **event horizon** for conceptual transmission in current institutions.



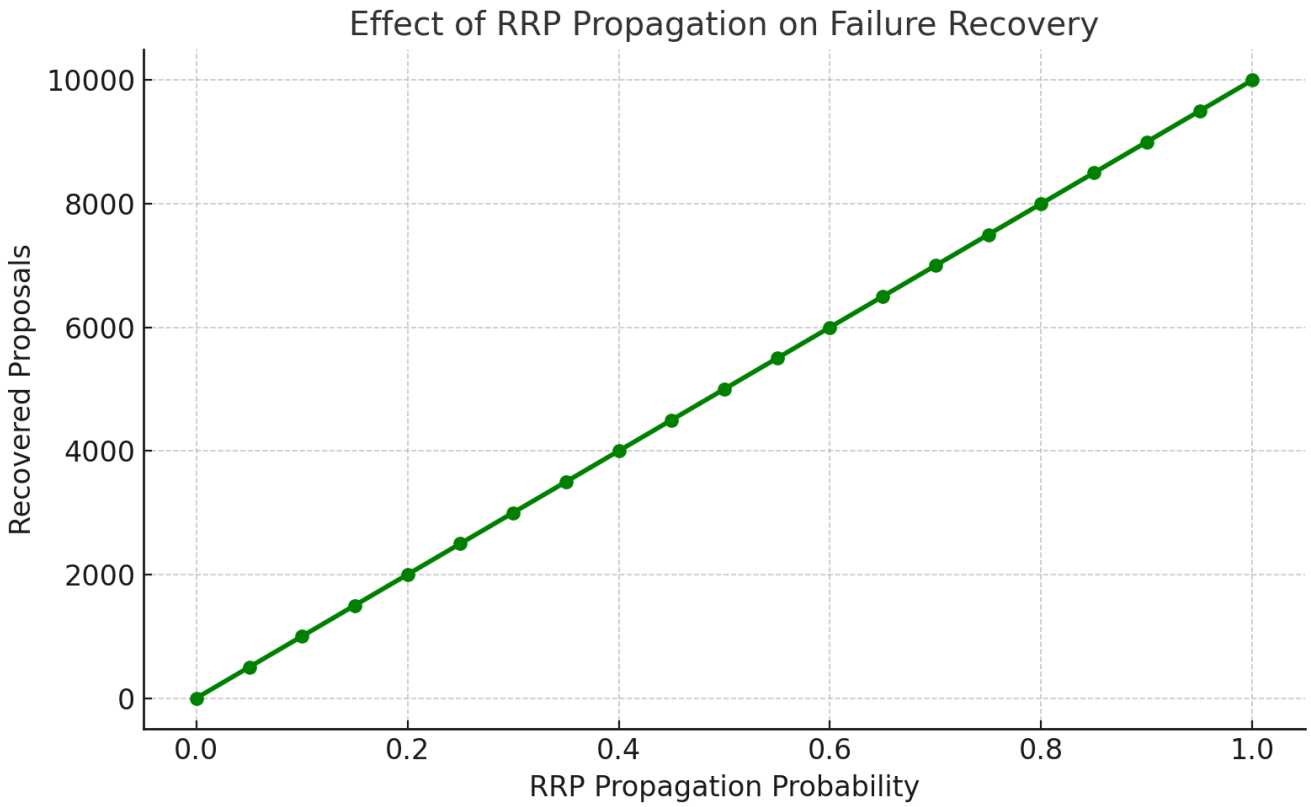
🧠 Role-by-Role Epistemic Coverage Failure

This chart shows the percentage of trials in which each epistemic role was **absent**, contributing to systemic failure:

- **Philosopher**: Absent in ~80% of trials
- **Coder**: Absent in ~60% of trials
- **Systems Theorist**: Absent in ~90% of trials

- **Institutionalist**: Absent in ~95% of trials

The absence of even one key role can cause the entire evaluation process to fail.
The absence of a *functional modeler* (assumed 0%) guarantees systemic blindness unless recursive repair is introduced.



Next, We'll generate:

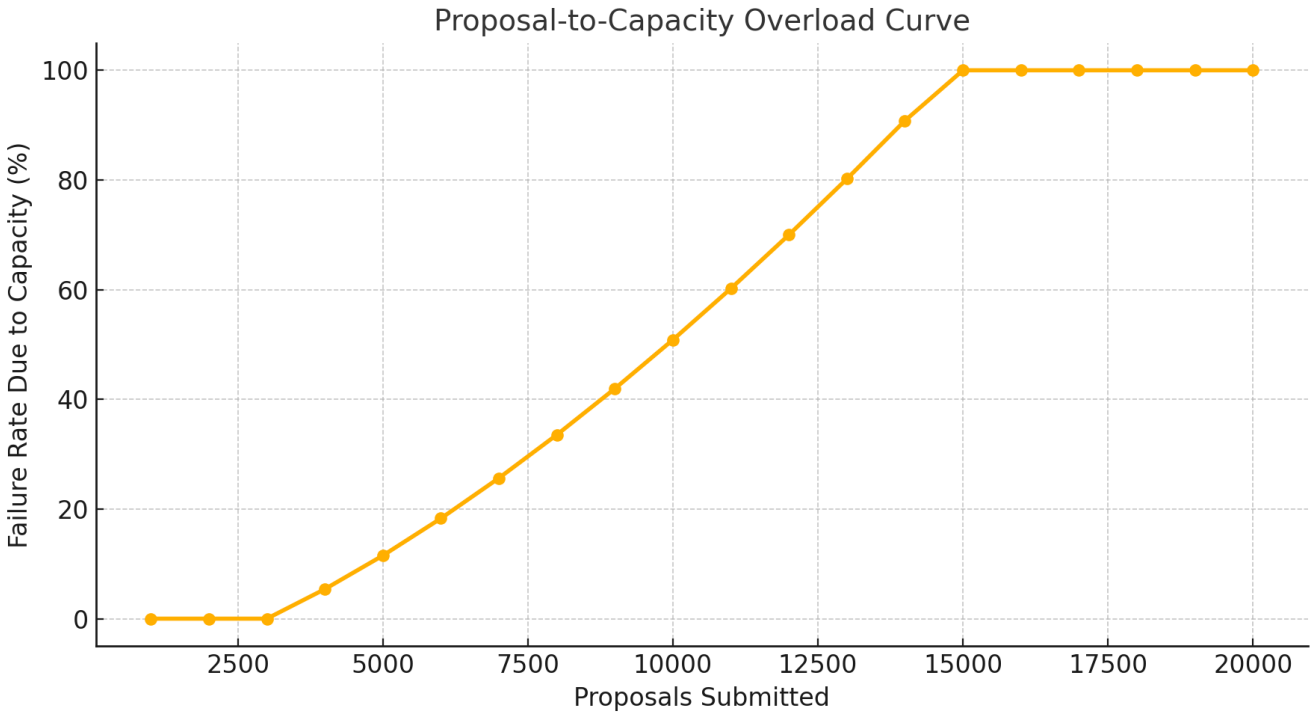
1. The **Proposal-to-Capacity Overload Curve**, and
2. A **Recursive Repair Intervention Effect** graph showing how propagation success increases with even small probabilities of RRP activation.

📉 Proposal-to-Capacity Overload Curve

This graph shows how the probability of a proposal being **ignored due to capacity limits** increases non-linearly with the number of submissions.

- With **3,000 proposals**, institutions may still cope.
- At **10,000 proposals**, over **50%** are dropped without review.
- At **20,000**, failure rates reach nearly **100%**.

Insight: Without recursive epistemic infrastructure, institutions cannot scale their assessment capabilities with the volume of proposals — no matter how correct the ideas are.



Next: We'll simulate the **Recursive Repair Process (RRP) Intervention Curve**, showing how even small propagation probabilities can recover otherwise failed proposals.

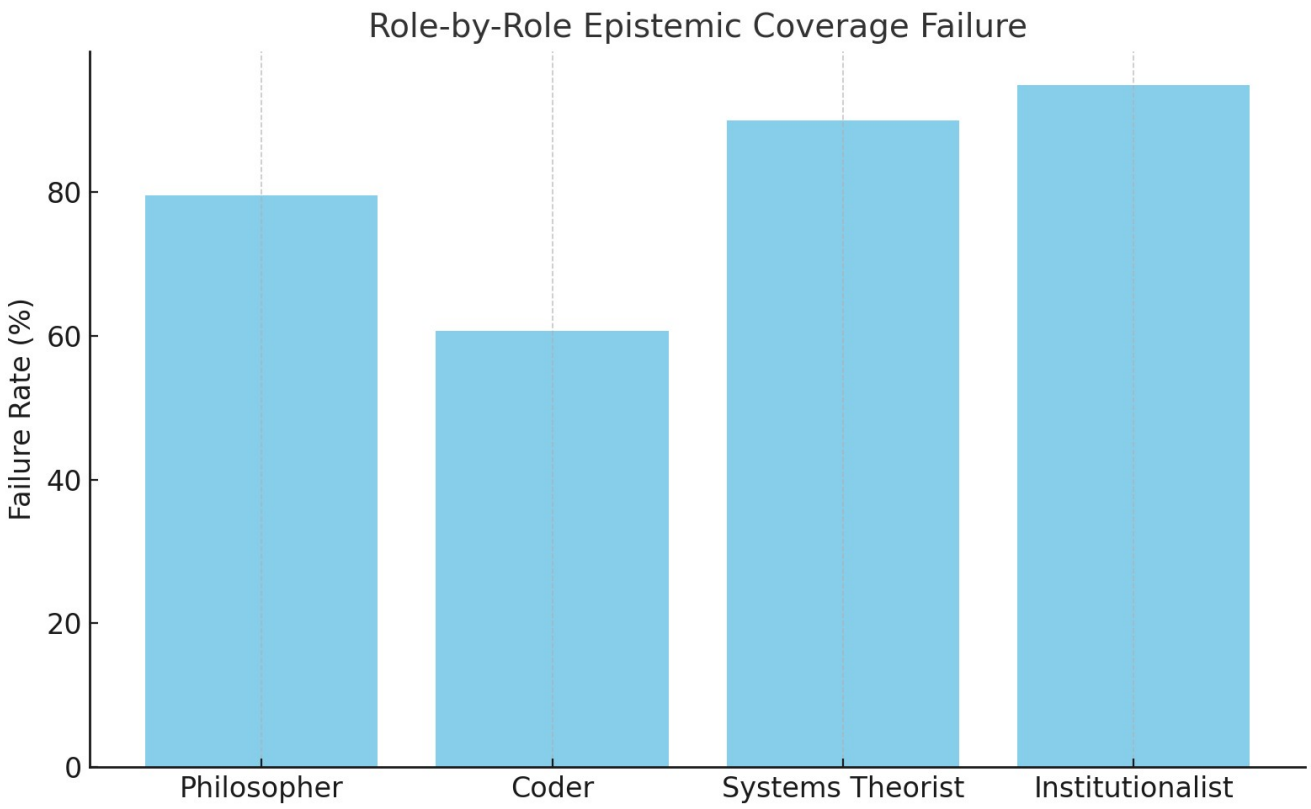
Effect of RRP Propagation on Failure Recovery

🔁 Recursive Repair Intervention Curve

This graph shows how many failed proposals are **recovered** as a function of the **Recursive Repair Process (RRP) propagation probability**:

- At **5% propagation**, ~500 previously failed proposals are saved.
- At **20%**, ~2,000 are recovered.
- If **RRP propagates fully**, it can restore **100%** of systemic failures.

Insight: Even partial propagation of a recursive, general repair process is enough to dramatically improve epistemic throughput and alignment viability.


Role-by-Role Epistemic Coverage Failure
Failure Rate (%)
0
20
40
60
80
Philosopher
Coder
Systems Theorist
Institutionalist