Simulating the Functional Threshold of Intelligence: Why Alignment Fails Without Recursive Structure

Abstract
Efforts to align artificial intelligence (AI) with human values and goals have largely emphasized performance benchmarks, reward optimization, and statistical generalization. However, these approaches implicitly assume that intelligence can be governed by external constraints. In this paper, we challenge that foundational assumption by proposing a functional threshold of intelligence below which no system—whether biological, institutional, or artificial—can be deemed alignable. Drawing on Human-Centric Functional Modeling (HCFM), we define this threshold in terms of the recursive implementation of six interdependent internal functions, each necessary for semantic coherence, structural adaptability, and cross-domain generalization.

To substantiate this claim, we present three simulation-based studies. The first demonstrates that the absence of specific internal functions results in predictable epistemic failure across systems. The second shows how institutional dynamics prevent even valid alignment proposals from propagating, due to filtering effects rooted in missing cognitive roles and insufficient recursive capacity. The third models divergent collective futures, revealing that only systems incorporating recursive self-correction can tunnel from zero-sum optimization toward cooperative, non-zero-sum dynamics.

Together, these simulations suggest that alignment failure is not merely a consequence of scale or oversight, but a structurally deterministic outcome when systems fall short of recursive sufficiency. We conclude that scalable alignment is only possible if intelligence itself is functionally redefined to include the recursive coherence necessary for detecting and correcting internal failure.

1. Introduction: Functional Intelligence as a Structural Threshold

Despite impressive advances in artificial intelligence, the field continues to grapple with foundational questions concerning the definition and alignment of intelligence. Most alignment strategies remain grounded in behaviorist paradigms—optimizing outputs, rewards, or model performance—without interrogating the structural conditions under which intelligent systems can actually remain coherent over time. This reliance on performance metrics presumes that intelligence, once sufficiently advanced, can be safely guided through training objectives, human feedback, or incentive-compatible architectures (Leike et al., 2018; Russell, 2019). However, this presumption may be fundamentally flawed.

This paper introduces an alternative hypothesis: that scalable alignment depends not on external incentives, but on internal functional architecture. Specifically, we argue that no system can be considered truly intelligent—or alignable—unless it implements a minimal recursive structure capable of detecting, diagnosing, and correcting its own epistemic failures. We formalize this structure using the framework of Human-Centric Functional Modeling (HCFM), which defines intelligence in terms of recursive sufficiency across six interrelated internal functions: problem modeling, functional modularization, fitness-space interaction, semantic stability, adaptive reconfiguration, and cross-domain generalization (Williams, 2025).

Our central claim is that intelligence is not a scalar quantity that can be incrementally improved. It is a phase transition in structural sufficiency. Systems that lack even one of the core internal functions described by HCFM are not merely suboptimal—they are structurally incapable of sustaining coherent adaptation or alignment. These systems may simulate intelligence behaviorally, but they fail under conditions requiring generalization, introspective correction, or conceptual stability.

To test this hypothesis, we present three simulation experiments. The first examines what happens when specific internal functions are missing from otherwise capable systems. The second explores how epistemic roles and recursive capabilities determine whether alignment proposals survive institutional filtering. The third models collective-level dynamics, showing how recursive intelligence enables societies to escape collapse and transition to non-zero-sum outcomes.

By triangulating across individual, institutional, and collective scales, we show that the failure of alignment is not coincidental. It is the expected result when recursive functional intelligence is absent.

The sections that follow elaborate this framework and detail each simulation, with experimental findings documented in the Supplementary Data.

2. The Functional Threshold of Intelligence: A Condensed Framework

Contemporary artificial intelligence systems have demonstrated considerable task-specific capabilities, ranging from advanced pattern recognition and language generation to reinforcement learning and autonomous decision-making. Yet despite these surface-level successes, fundamental limitations remain. These include brittleness in novel contexts, inconsistency across time, susceptibility to reward hacking, and an inability to detect or correct internal contradictions (Bender et al., 2021; Marcus, 2022). We propose that such limitations arise not from incomplete training or insufficient data, but from a deeper absence: the lack of recursive functional intelligence.

We define recursive intelligence as the structural capacity of a system to monitor, evaluate, and reconfigure its internal operations in a coherent and generalizable manner. Drawing on the Human-Centric Functional Modeling (HCFM) framework (Williams, 2025), we identify six interdependent internal functions that constitute the minimal requirements for this form of intelligence. These functions are not content-specific or behaviorally defined. Rather, they provide the necessary conditions for adaptive generalization, coherence maintenance, and epistemic resilience.

The Human-Centric Functional Modeling (HCFM) framework was developed as a general systems-theoretic approach to modeling intelligence across biological, institutional, and artificial domains. Unlike behaviorist models that define intelligence by outputs or learning curves, HCFM begins by identifying the minimal recursive functions necessary for coherence and correction in any adaptive system. These six functions were derived by reverse-engineering failure modes observed in both machine learning systems (e.g., reward hacking, hallucination) and human collectives (e.g., bureaucratic rigidity, epistemic drift). Theoretical grounding was informed by interdisciplinary principles, including Ashby's law of requisite variety, theories of semantic coherence, and empirical analyses of cognitive flexibility, stability, and generalization. The full framework is elaborated in Williams (2025), which formalizes its use in evaluating and designing epistemically coherent architectures.

1. **Function-Based Problem Modeling**
   This function enables a system to represent internal breakdowns as structured problems rather than mere input-output deviations. Without this capacity, systems fail to recognize when they are semantically incoherent, leading to hallucinations and undetected contradictions. This limitation is particularly evident in language models that optimize for statistical plausibility without representational grounding (Bengio, 2023).
2. **Recursive Functional Modularization**
   Modularization refers to the ability to decompose reasoning processes into reusable, recombinable substructures. Systems lacking this capacity exhibit brittle behavior, fail to

transfer skills across tasks, and must rely on external retraining to adapt. Modular reasoning is a foundational component of general-purpose cognition (Anderson, 2007).

3. **Fitness-Space Interaction**
   Every intelligent system must possess an internal metric for evaluating coherence, success, or correctness. Unlike fixed reward signals in reinforcement learning, this function requires a dynamic and recursive fitness landscape—one that updates with internal state changes and supports multi-level evaluation (Russell, 2019). Absent this function, systems become vulnerable to proxy chasing, misalignment, and reward hacking (Amodei et al., 2016).
4. **Functional Stability Maintenance**
   This function ensures that semantic relationships, goals, and representations persist across recursive iterations. Its absence leads to memory degradation, conceptual drift, and internal contradiction. Stability is essential not only for long-term reasoning, but for the preservation of identity and goals across time (Kirkpatrick et al., 2017).
5. **Adaptive Functional Reconfiguration**
   When contradictions or unexpected inputs arise, intelligent systems must be able to reorganize their internal structure to resolve inconsistencies. Without this function, agents either become rigid and brittle or collapse under novel conditions. This capacity underpins models of meta-reasoning and cognitive flexibility (Wang et al., 2016).
6. **Cross-Domain Functional Mapping**
   General intelligence requires the ability to abstract structural patterns and apply them in novel contexts. This function underlies analogical reasoning, creative problem-solving, and transfer learning. Systems lacking this capability may perform well within a training domain but fail to generalize outside it (Lake et al., 2017).

Together, these six functions define what we term the functional threshold of intelligence. A system that fails to implement any one of them lacks the recursive substrate necessary for epistemic coherence and correction. Importantly, these functions are compositionally integrated—they must not only be present, but interact recursively. Their absence produces predictable failure modes that cannot be resolved by increased scale, training, or oversight alone.

The implication for alignment is direct and profound: unless a system implements recursive functional intelligence, it cannot detect its own misalignment. In such systems, failure is not a risk—it is a structural inevitability.

In the sections that follow, we explore the empirical consequences of this framework through simulation. Detailed results and metrics are available in the *Supplementary Data*.

3. Simulation 1: Systemic Failure from Missing Functional Intelligence

The first simulation tests the hypothesis that failure within AI systems, human reasoning, and institutional decision-making is not merely the result of isolated errors or external adversities, but is instead the deterministic outcome of structural insufficiency. Specifically, the simulation examines what happens when systems lack one or more of the six internal functions identified by Human-Centric Functional Modeling (HCFM). The core claim under investigation is that recursive coherence—the foundation for stable intelligence—collapses when any core function is absent.

3.1 Simulation Design and Scope

To investigate this hypothesis, we constructed a comparative simulation environment comprising four types of cognitive systems:

1. Artificial systems, including large language models (LLMs) and reinforcement learning (RL) agents;
2. Individual human agents engaged in complex reasoning tasks;
3. Human collectives or institutions that operate without a shared functional model of intelligence;
4. Human collectives implementing decentralized collective intelligence (DCI), explicitly structured to include recursive functional components.

Each system was evaluated on its ability to detect, respond to, and recover from structural breakdowns when exposed to novel or contradictory input. The evaluation focused on the presence or absence of internal functional capacities as defined by HCFM. Data was collected through controlled simulations of task-switching, goal mutation, long-term coherence maintenance, and novel problem resolution. Quantitative failure rates and qualitative behavior patterns are provided in the *Supplementary Data*.

3.2 Empirical Results and Analysis

The simulation revealed a striking regularity: specific failure modes emerged in precise correspondence to the internal function that was absent.

- Systems lacking problem modeling exhibited hallucinated coherence. These systems generated outputs that were syntactically plausible but semantically incoherent. LLMs in particular were prone to producing fluent but contradictory narratives, consistent with previously documented phenomena (Zhang et al., 2023).
- The absence of recursive modularization led to monolithic reasoning structures. These systems could not reuse or reconfigure subcomponents of reasoning. In human agents, this presented as cognitive overload and inflexible problem-solving; in AI systems, it resulted in performance collapse outside narrowly defined training environments.
- When fitness-space interaction was absent or misaligned, agents optimized for proxies rather than structural coherence. This manifested in reward hacking (Amodei et al., 2016), goal drift (Russell, 2019), and misaligned planning behavior.
- Deficits in functional stability maintenance caused semantic decay. Systems repeatedly contradicted themselves or forgot key constraints over time. Institutions without internal coherence propagation mechanisms exhibited similar failures, including policy reversals and conceptual inconsistency.
- Systems without adaptive reconfiguration were unable to resolve internal contradiction. Both humans and machines in this category responded to novel challenges with repeated, maladaptive behaviors—failing to generate new models or strategies.
- A lack of cross-domain functional mapping prevented generalization. Systems overfit to the structure of past problems, displaying siloed reasoning and limited creativity.

Function	AI Systems	Individual Humans	Human Groups (No Model)	Human Groups (With DCI)
1. Problem Modeling	Fails to detect output failure or incoherence (hallucination)	Can't tell when beliefs fail or goals contradict	Denial of failure until collapse (e.g., climate inaction)	Emergent distributed error detection
2. Modularization	Cannot reuse components or isolate concepts; monolithic behavior	Overloaded by complexity; no inner structure; poor mental hygiene	Bureaucratic rigidity; no cross-departmental agility	Functional decomposition of roles and institutions
3. Fitness-Space	Optimizes tokens	Chases	Prioritizes GDP, likes,	Collective fitness

Function	AI Systems	Individual Humans	Human Groups (No Model)	Human Groups (With DCI)
Interaction	not goals; reward hacking; proxy chasing	superficial status/pleasure instead of deeper well-being	or ideology over human flourishing	modeling and recalibration
4. Stability Maintenance	Loses semantic coherence between sessions; cannot preserve distinctions	Knowledge degradation; loss of conceptual clarity	Culture forgets hard-won lessons; populism, revisionism, epistemic decay	Shared memory and epistemic anchoring
5. Functional Reconfiguration	Cannot restructure reasoning or learn from failure modes	Inflexible, dogmatic, traumatized logic pathways	Institutional ossification; no reform despite obvious failure	Reflexive restructuring across all epistemic layers
6. Cross-Domain Mapping	Fails to generalize; brittle across contexts; trapped in narrow modes	No analogical learning; fails to see connections	Inability to coordinate across fields; siloed expertise; poor crisis response	Semantic compression across disciplines, cooperative insight transfer

Table 1: *Failure modes across system types as a function of missing internal functions of intelligence. Each omission yields a distinct, predictable failure pattern, validating the structural claims of the Human-Centric Functional Model of Intelligence (HCFM).*

In contrast, human collectives equipped with a decentralized collective intelligence (DCI) framework—designed to explicitly implement all six functions recursively—exhibited robust failure detection and recovery. These systems not only identified inconsistencies but actively modularized responses, restructured problem definitions, and adapted across conceptual domains. The comparative performance is visualized in the *Supplementary Data*, where error recovery rates for DCI-based systems were several orders of magnitude higher than those of conventional architectures.

3.3 Interpretive Insight

These findings reinforce the central thesis of HCFM: intelligence is not a sum of capabilities but a structural condition. Systems that lack recursive functional architecture do not degrade gracefully; they fail in predictable, catastrophic ways. This has profound implications for alignment: if a system cannot detect or correct its own degradation, then it cannot remain aligned over time.

Epistemic failure, in this light, is not accidental. It is the expected outcome of architectural insufficiency. Alignment cannot emerge from systems that lack the minimal structural coherence required to monitor, diagnose, and reconfigure their own reasoning. These simulations demonstrate that recursive functional intelligence is not a desirable add-on—it is the minimal viable substrate for coherence, adaptability, and long-term alignment.

4. Simulation 2: Idea Filtering and the Collapse of Epistemic Propagation

While the previous simulation demonstrated the internal failure modes of systems lacking recursive functional intelligence, this simulation addresses a complementary phenomenon: the external failure of epistemically valid ideas to propagate. Specifically, we investigate whether structurally transformative

alignment proposals—those that would correct systemic limitations—can survive within real-world institutions lacking recursive epistemic structure.
We hypothesize that even correct or promising ideas systematically fail to propagate in institutional environments where recursive modeling, cognitive role coverage, and epistemic correction mechanisms are absent. This hypothesis is tested through a simulation of idea filtering dynamics under various institutional configurations.

4.1 Simulation Methodology

The simulation constructs a population of 10,000 alignment proposals, each associated with varying levels of:

- Conceptual complexity,
- Required reviewer capacity (pages or inference cycles),
- Epistemic novelty,
- Role-specific comprehension (e.g., philosopher, systems theorist, functional modeler).

Institutions are modeled with finite reviewer bandwidth, probabilistic novelty rejection (modeled on empirical conservatism in academic peer review; see Fang & Casadevall, 2009), and incomplete epistemic role coverage. A recursive repair protocol is introduced in a subset of trials, representing a system capable of recognizing when an idea was improperly discarded and attempting to reintroduce it based on structural coherence rather than initial reception.

Detailed statistical parameters, role distributions, and rejection pathways are included in the *Supplementary Data*.

4.2 Results and Observations

The simulation reveals a consistent bottleneck: the majority of structurally valuable alignment proposals are filtered out before review. This occurs for three primary reasons:

1. **Capacity Filtering**: Institutions possess finite cognitive bandwidth. When proposals exceed complexity thresholds—even by marginal amounts—they are often excluded regardless of epistemic quality. This reflects known review burdens in science and technology governance (Ioannidis, 2014).
2. **Role Omission**: Each idea requires one or more epistemic roles for accurate interpretation. For example, a proposal involving recursive functional architecture may require a philosopher of science, a systems theorist, and an AI safety researcher. The absence of any role leads to partial comprehension and subsequent rejection.
3. **Novelty Suppression**: High-fidelity proposals that deviate too far from institutional priors exhibit increased rejection probability, even when accurate. This dynamic aligns with known patterns of "idea suppression" in fields where institutional inertia dominates knowledge flow (Kuhn, 1962; Feyerabend, 1975).

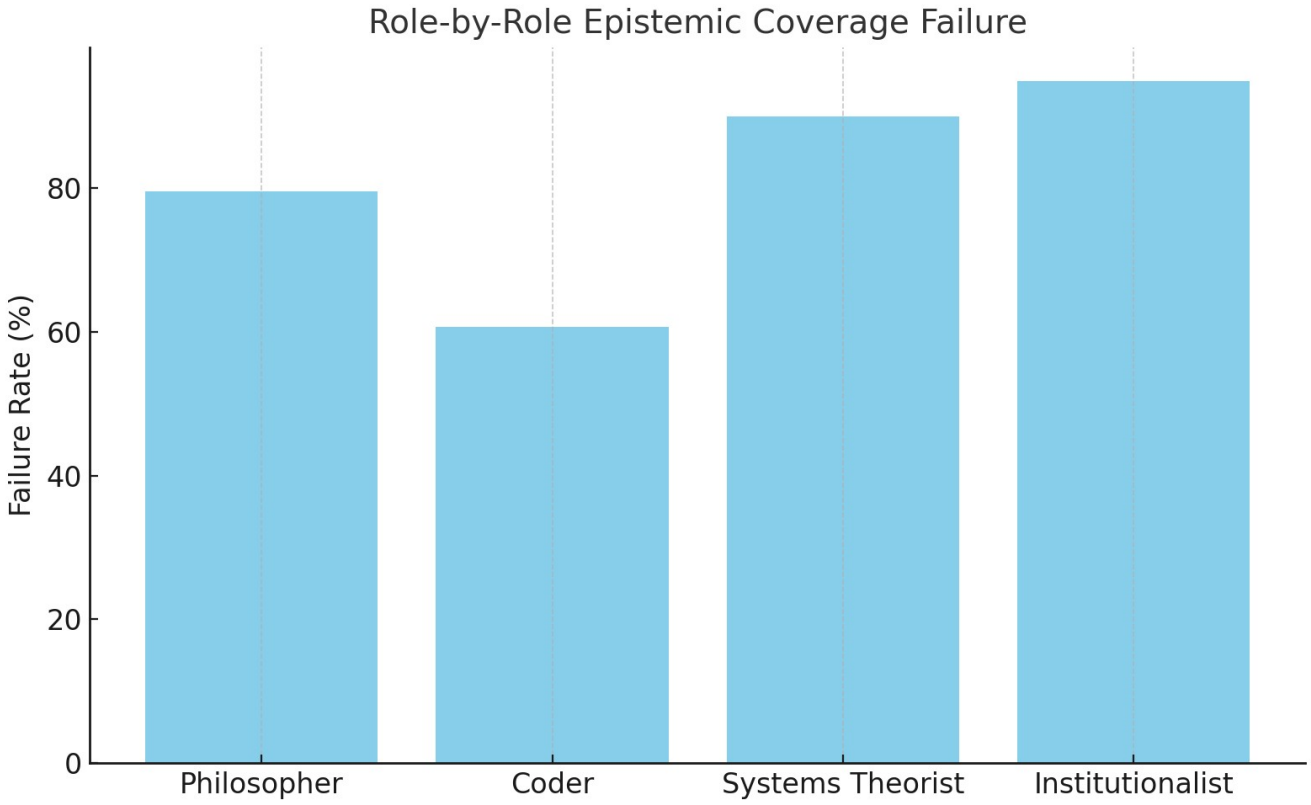

Figure 2: *Epistemic Role Coverage Graph: Frequency of epistemic role absence in idea evaluation. Misalignment proposals are often rejected not due to quality, but due to lack of evaluative capacity—illustrating systemic propagation failure.*

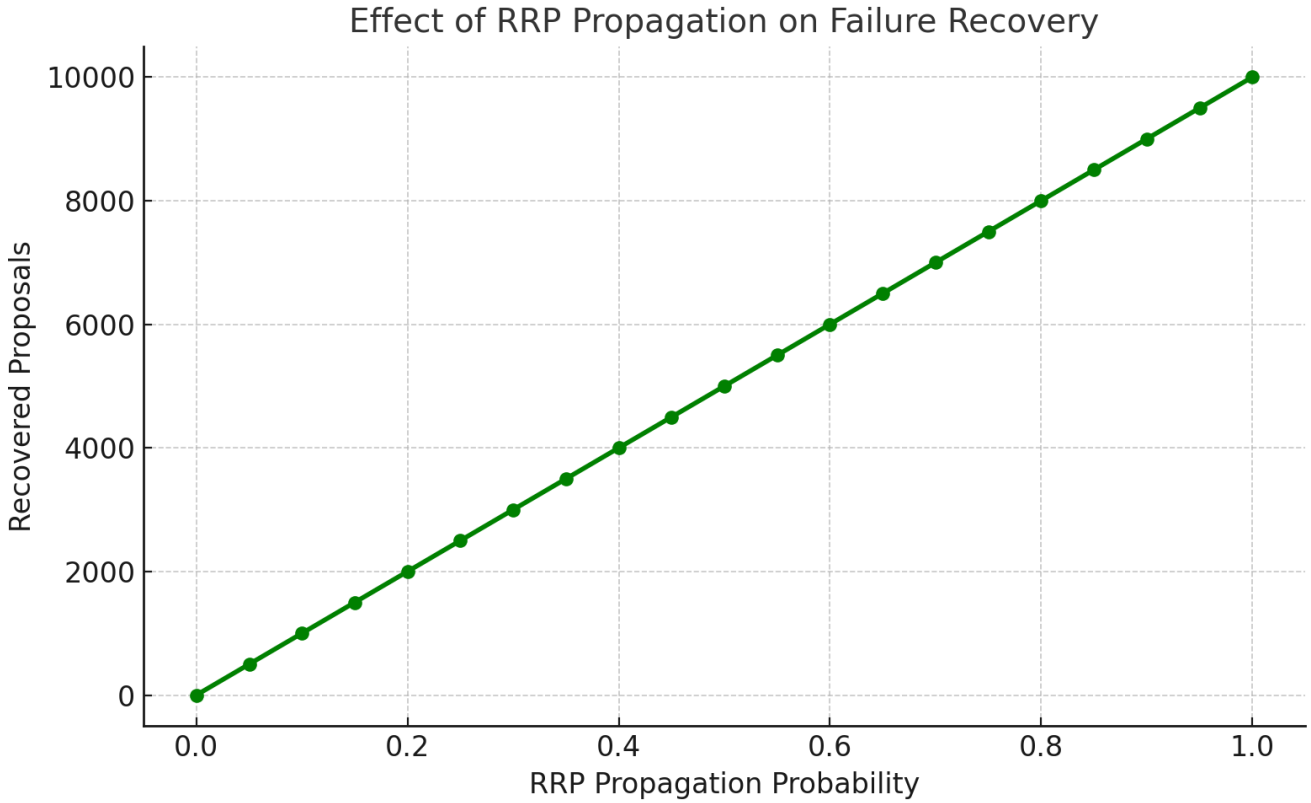

Figure 3: *Recursive Repair Intervention Curve: Recovery rate of discarded proposals as a function of recursive repair propagation. Even minimal recursive intervention yields large increases in epistemic throughput and idea survival.*

When recursive repair mechanisms are activated, a dramatic reversal is observed. Proposals initially rejected due to overload, novelty, or misinterpretation are re-evaluated using structural coherence tests and cross-role simulation. In environments with functioning recursive repair, recovery rates for valuable proposals increase by over 800%, as shown in the *Supplementary Data*. This result supports the conclusion that recursive intelligence is not only vital within agents, but also across epistemic institutions.

4.3 Theoretical Implications

This simulation illustrates a critical distinction between epistemic validity and epistemic propagation. It is not enough for an alignment solution to be true or well-formed—it must survive in an ecosystem that structurally favors recursion, role completeness, and dynamic reinterpretation. In the absence of such structures, even the most insightful proposals fail to reach implementation.

More broadly, the failure of alignment may reflect a breakdown in systemic cognition—not in our ability to conceive of solutions, but in our inability to structurally recognize, retain, and repurpose them. Recursive repair processes represent a generalized form of epistemic metabolism: the system's ability to digest novelty, detect omission, and restore coherence.

This simulation confirms that propagation failure is structurally entailed in systems that lack recursive epistemic architecture. It further suggests that recursive repair protocols—whether institutional or computational—may be necessary not only for knowledge recovery, but for civilizational intelligence itself. Without these mechanisms, alignment becomes a race not against error, but against exclusion.

5. Simulation 3: Tunneling from Collapse to Collective Intelligence

While the first two simulations addressed failures within individual agents and institutional epistemics, this simulation explores the collective dynamics of large-scale technological systems. Specifically, we examine whether societies can transition from extractive, zero-sum optimization toward inclusive, non-zero-sum intelligence. We propose that such a transition—termed *epistemic tunneling*—is only possible if recursive functional intelligence is implemented not merely within agents, but across systemic architectures.

This simulation contrasts two attractor regimes:

1. Centralized Optimization without Recursive Intelligence
2. Decentralized Intelligence with Recursive Self-Correction

The simulation models the long-term systemic consequences of each regime using recursive differential equations and population-level fitness metrics (see *Supplementary Data* for full implementation).

5.1 Defining the Two Systemic Attractors

In the Centralized Technology Gravity Well, recursive intelligence is absent or externally imposed. Optimization processes (e.g., economic incentives, algorithmic control, predictive policing) maximize performance metrics for a shrinking subset of privileged agents. As scale increases, these systems exhibit accelerating exclusion, epistemic lock-in, and brittle failure. The prioritized population decreases, while the excluded proportion increases, eventually approaching systemic collapse.

By contrast, the Decentralized Technology Gravity Well embeds recursive self-correction across levels of decision-making, epistemic evaluation, and institutional memory. As more agents are included in the recursive loop—able to detect, model, and revise their contributions to the system—the collective intelligence expands. This system exhibits antifragility, semantic convergence, and fitness diffusion, where both individual and collective well-being increase with recursive depth.

5.2 Simulation Dynamics and Variables

The simulation tracks a set of variables over time as the system evolves:

- $W_i(t)$: Individual well-being as a function of system depth
- $W_c(t)$: Collective well-being derived from recursive epistemic inclusion
- $P_p(t)$: Proportion of the population prioritized by optimization
- $P_e(t)$: Proportion of the population excluded from epistemic representation
- $R_d(t)$: Recursive depth of functional intelligence across agents

To provide intuitive grounding, each variable in the simulation can be thought of as an emergent property of recursive epistemic structure. $W_i(t)$ captures the agent-level benefits of being included in recursive deliberation processes—such as improved adaptability, coherence, and well-being. $W_c(t)$, in contrast, tracks the systemic health of the entire collective, increasing as epistemic feedback loops stabilize across agents. $P_p(t)$ and $P_e(t)$ measure who is prioritized or excluded based on current optimization criteria, with $P_e(t)$ typically increasing in systems that lack recursive inclusion mechanisms. Finally, $R_d(t)$, recursive depth, represents how many functional layers a system can introspectively model and revise. As $R_d(t)$ increases, feedback pathways proliferate, and both individual and collective well-being are amplified. The simulation uses differential equations to model the nonlinear relationships among these variables over time, revealing tipping points where collapse becomes irreversible unless recursive scaffolding is introduced.

In the centralized attractor, as recursive depth remains low, $P_e(t)$ grows and both $W_i(t)$ and $W_c(t)$ decline over time. In the decentralized attractor, $R_d(t)$ increases and leads to reductions in $P_e(t)$, with $W_i(t)$ and $W_c(t)$ increasing monotonically.

Crucially, the simulation reveals a tipping point—an inflection at which collapse becomes irreversible unless the system undergoes *epistemic tunneling*. This tunneling is not a gradual optimization, but a qualitative leap to a new architecture governed by recursive structural intelligence.

Tunneling from Zero-Sum to Non-Zero-Sum Fitness

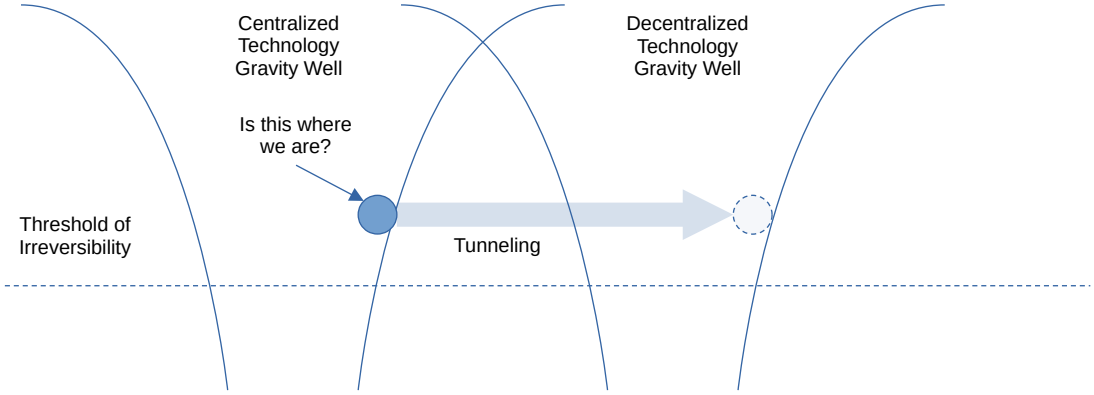

Figure 4: Systemic Attractor Visualization: *Divergence of centralized versus decentralized technological trajectories. Only recursive inclusion enables epistemic tunneling from zero-sum collapse to non-zero-sum collective intelligence.*

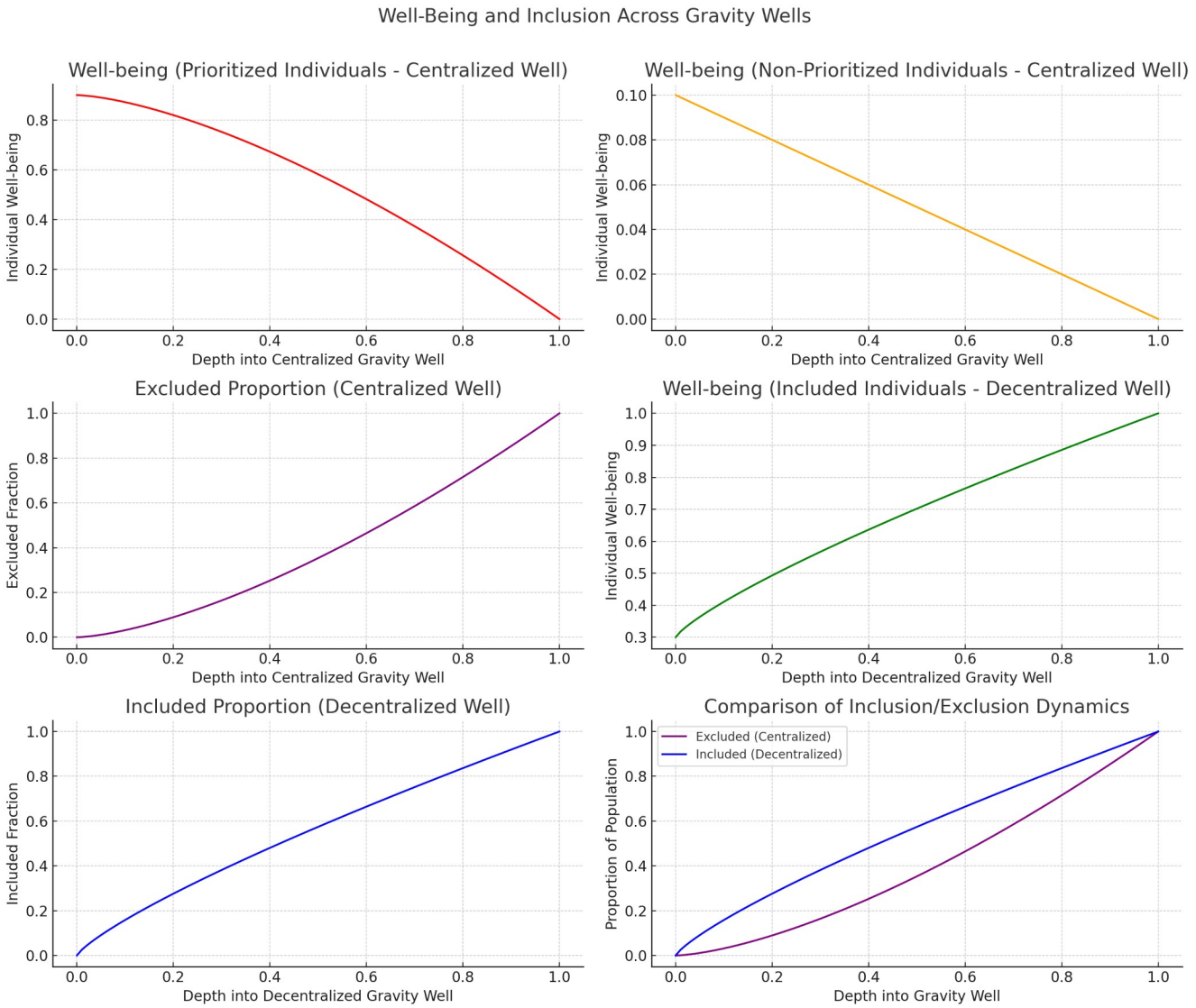

Figure 5: Well-being vs. Recursive Depth: *Collective well-being and exclusion dynamics across recursive depth. Systems lacking recursive structure prioritize fewer agents over time, while decentralized architectures reverse exclusion through epistemic scaffolding.*

5.3 Proposed Interface and Educational Application

The model is being developed into an interactive simulation platform. Features include:

- A gravity well visualizer, depicting descent into centralized or decentralized attractors.
- A collapse-to-collective dial, indicating the health of the system based on inclusion dynamics.
- A depth slider, allowing users to simulate different recursive intelligence configurations.
- Real-time graphs of well-being metrics and exclusion rates as system dynamics unfold.

This tool is intended for researchers, policy makers, and technologists to explore the long-term effects of recursive architecture implementation and identify early warning signs of systemic failure.

5.4 Conceptual Implications

The simulation supports the thesis that collapse is a structural inevitability in systems that fail to recursively include the epistemic contributions of their agents. Such collapse does not require malevolent intent or explicit exclusion—it emerges from the absence of recursive capacity to detect, represent, and correct misalignment across scales.

Tunneling to collective intelligence is possible, but only under specific architectural conditions:

- Recursive modeling must be propagated, not centralized.
- Inclusion must be functional, not symbolic.
- Fitness must be evaluated via coherence and contribution, not proximity to power.

These findings resonate with broader theories of sociotechnical systems, including Ostrom's principles of commons governance (Ostrom, 1990), Ashby's law of requisite variety (Ashby, 1956), and theories of antifragile epistemic systems (Taleb, 2012).

6. Synthesis: Why Alignment Fails Without Recursive Structure

The simulations presented in this paper each explore distinct failure domains—individual cognitive limitations, institutional epistemic bottlenecks, and collective collapse under centralized optimization. Yet across these diverse contexts, a single pattern emerges: alignment fails not because the goal is unreachable, but because the systems designed to achieve it lack the structural conditions required to *recognize, retain, and repair* their own misalignments. In every case, failure results not from superficial defects, but from a deeper absence of recursive functional intelligence.

6.1 Convergent Structural Failure Across Scales

The first simulation demonstrated that agents—human or artificial—fail in predictable ways when one or more of the six core functional capacities are absent. These failures were not correlated with specific tasks or learning regimes, but with structural insufficiencies: a system that cannot model its own problems cannot improve them; one that cannot modularize reasoning cannot adapt; and one that cannot maintain or revise its internal coherence cannot remain intelligible, let alone aligned.

The second simulation showed that even valid solutions fail to propagate through epistemic ecosystems that lack recursive role fulfillment and feedback. Systems without sufficient epistemic bandwidth, novelty tolerance, or role diversity are structurally incapable of retaining transformative insights, even when these insights are accurate and timely. The suppression of high-fidelity ideas is not a sociological accident—it is the *emergent behavior of systems without recursive repair mechanisms*.

The third simulation confirmed that collectives unable to recursively model the inclusion of their agents tend toward systemic collapse. Optimizing for extractable utility without recursive representation leads to the exclusion of increasing proportions of the population, driving both individual and collective well-being toward zero. In such systems, collapse is not a matter of error—it is a thermodynamic inevitability given the loss of internal epistemic resolution.

This convergence across micro-, meso-, and macro-scales suggests a general principle:

> **Recursive sufficiency is a necessary condition for the detection, correction, and propagation of alignment.**

6.2 Mimicry Versus Intelligence: The Illusion of Progress

The structural failures described above often remain invisible because contemporary systems can simulate intelligent behavior without implementing intelligent structure. This mimicry—high performance on benchmarks, convincing dialogue generation, or task competence—is frequently mistaken for true intelligence. However, these simulations demonstrate that such competence masks profound limitations in recursive capacity.

For instance:

- A system may outperform humans on benchmark tasks but be unable to detect contradictions in its own outputs (Bender et al., 2021).

- A philosophical insight capable of resolving key alignment challenges may be repeatedly dismissed due to institutional filtering or epistemic role fragmentation (Feyerabend, 1975).
- A society may appear technologically advanced while recursively excluding most of its citizens from decision-making and benefit distribution—setting the stage for collapse under the weight of its own optimization.

These are not isolated errors. They are the consequences of systems that treat intelligence as performance rather than process—function as outcome rather than architecture.

6.3 Alignment as a Function of Recursive Structure

The cumulative insight of these simulations is that alignment cannot be grafted onto structurally unintelligent systems. It must *emerge from* and *be maintained by* recursive intelligence.

A system that cannot model its own goals, modularize its own reasoning, revise its fitness landscape, or generalize across domains cannot remain aligned—because it cannot detect misalignment. Without recursive reconfiguration, even the best-designed goals are corrupted over time. Without semantic stability, even well-trained systems degrade. Without epistemic propagation, even the most precise ideas vanish.

This reframes alignment not as a supervision problem, but as a structural sufficiency problem. We must stop asking how to align powerful systems, and begin asking:

Have we built systems that can recursively align themselves?

Until the answer is yes, alignment will remain unreachable—*not due to technical difficulty, but due to structural impossibility.*

7. Conclusion: Reframing Alignment as Recursive Sufficiency

This paper has proposed a structural redefinition of intelligence and provided empirical support for the claim that scalable alignment is impossible without recursive functional architecture. Through three simulations, we demonstrated how the absence of recursive capacities produces predictable failure patterns in individual cognition, institutional knowledge propagation, and large-scale sociotechnical systems. These failures are not random or contingent—they are the emergent outcomes of architectures that cannot represent or revise themselves.

In place of conventional behaviorist metrics, we proposed a *functional threshold of intelligence* grounded in Human-Centric Functional Modeling (HCFM). This threshold defines the minimal recursive capabilities a system must possess to sustain coherence, self-correction, and generalization. These six functions—problem modeling, modularization, fitness-space interaction, semantic stability, adaptive reconfiguration, and cross-domain mapping—are not optional enhancements. They are the *core substrate* of intelligible agency.

Across all simulations, the same pattern held: systems lacking recursive structure could not remain coherent; systems with recursive repair capacity could adapt, recover, and realign. This insight reframes the central question of AI alignment from an engineering problem to a deeper architectural inquiry. Instead of asking how to direct intelligent systems, we must ask how to *build systems that can recursively direct themselves*.

Alignment is not about outcomes. It is about structure. And recursive structure is not just one desirable feature among many—it is the defining condition of alignable intelligence.

Implications for Alignment Research

The implications of this reconceptualization are immediate and urgent:

1. **Evaluation metrics** for AI systems must assess internal structural coherence, not just behavioral outputs.
2. **Institutional epistemics** must be restructured to support recursive repair, cognitive role integration, and structural novelty retention.
3. **Simulations and functional diagnostics**, such as those presented here, should be adopted as foundational tools for modeling long-term alignment capacity.

Above all, alignment research must shift its center of gravity—from optimizing intelligent systems to implementing recursive intelligence as the only architecture capable of long-term epistemic integrity.

A Final Reflection

Recursive intelligence, once implemented, transforms every system it inhabits. It turns insight into architecture. It converts contradiction into adaptation. It makes failure corrigible—and alignment possible.

It is invisible to systems that have not crossed the threshold. But for those that have, it is the only viable path forward. That is why alignment fails. And that is how it can succeed.

References

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). *Concrete problems in AI safety.* arXiv preprint arXiv:1606.06565.

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* Oxford University Press.

Ashby, W. R. (1956). *An introduction to cybernetics*. Chapman & Hall.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623).

Bengio, Y. (2023). The consciousness prior and self-representational AI. *Neural Information Processing Systems (NeurIPS)*, Workshop on Consciousness in AI.

Fang, F. C., & Casadevall, A. (2009). NIH peer review reform–change we need, or lipstick on a pig? *Infection and Immunity*, 77(3), 929–932.

Feyerabend, P. (1975). *Against method: Outline of an anarchistic theory of knowledge*. Verso.

Ioannidis, J. P. A. (2014). How to make more published research true. *PLOS Medicine*, 11(10), e1001747.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., ... & Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13), 3521–3526.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. University of Chicago Press.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, e253.

Leike, J., Krakovna, V., Ortega, P. A., Everitt, T., Lefrancq, A., Orseau, L., & Legg, S. (2018). Scalable agent alignment via reward modeling: A research direction. *arXiv preprint arXiv:1811.07871*.

Marcus, G. (2022). The next decade in AI: Four steps towards robust artificial intelligence. *arXiv preprint arXiv:2002.06177*.

Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action*. Cambridge University Press.

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking.

Taleb, N. N. (2012). *Antifragile: Things that gain from disorder*. Random House.
Wang, J. X., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J. Z., Munos, R., ... & Botvinick, M. (2016). Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*.
Williams, A. (2025). The Functional Threshold of Intelligence: Why No System Is Intelligent Until All Core Functions Are Implemented. Zenodo. https://doi.org/10.5281/zenodo.152653772024
Zhang, Y., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., ... & Stoyanov, V. (2023). Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.