From Humanity’s Last Exam to Humanity’s First Adaptive Intelligence Exam: Recursive Self-Modeling for Civilizational Resilience

Abstract

Across mathematics, cybernetics, biology, cognitive science, and policy, scholars have independently shown that systems without explicit self-modeling and recursive self-correction fail when confronted by novel or evolving challenges—from Gödel’s incompleteness theorems to Normal Accident Theory in nuclear safety. We unify these strands into a single, rigorously proven Functional Model of Intelligence (FMI) theorem: *an agent that reliably solves every problem in an open domain must explicitly track and adapt its own problem-solving function.* We then embed this insight within a crowdfunded, bounty-driven campaign—modeled on Humanity’s Last Exam—to mobilize resources against concrete existential risks (AI misalignment, engineered pandemics, climate collapse, nuclear accidents, geoengineering). Finally, we add a comprehensive Survey of Self-Modeling Across Disciplines to demonstrate how universally every field demands internal representations of their own processes.

1. Introduction & Existential Urgency

Civilization today confronts an unprecedented spectrum of existential threats, many of which are not only growing in severity and complexity but also surpassing our collective ability to reliably manage them. Among these threats is the potential emergence of unaligned superintelligent artificial intelligence (AI), which could quickly escape human oversight or control, posing catastrophic risks (Bostrom, 2014; Yudkowsky, 2008; Armstrong & Sotala, 2012). Additionally, engineered pathogens pose the risk of devastating global pandemics in the absence of rapidly adaptive biosafety protocols (Walker et al., 2003; O’Toole et al., 2017). Climate change presents tipping points that demand real-time governance responses—responses that our current political and institutional frameworks consistently struggle to deliver (Rockström et al., 2009; Folke et al., 2005). Furthermore, longstanding threats such as nuclear conflict remain critically sensitive to failures in command-and-control systems that lack robust internal monitoring and self-corrective capacities (Perrow, 1999; Roberts, 1990). Emerging geoengineering proposals aimed at mitigating climate change carry novel risks that our existing institutions have yet to adequately model or regulate (Helbing, 2013).

Although the global community has extensively identified and even proposed theoretical solutions to many of these existential threats, we have not increased our collective capacity to reliably implement and adapt such solutions at scale. Institutional inertia, compounded by increasing complexity, consistently hampers effective responses (Taleb, 2012; Helbing, 2013). Indeed, contemporary institutions typically rely on static or linear governance models—systems fundamentally incapable of dynamically responding to rapidly changing and unpredictable environments. These models are constrained by structural limitations well-established in foundational theoretical work such as Gödel’s incompleteness theorems, which demonstrated the inherent limitations of fixed axiomatic systems to fully encompass all truths (Gödel, 1931; Tarski, 1936). Moreover, traditional governance and institutional structures commonly exhibit compartmentalization, hindering interdisciplinary integration and cooperation essential for addressing complex global challenges (Ostrom, 1990; Senge, 1990).

Critically, our difficulty in managing existential threats arises not simply from inadequate identification of these threats, nor even solely from insufficient institutional responsiveness, but from deep-rooted cognitive and cultural barriers that prevent the propagation and implementation of integrative, adaptive solutions. Human cognition faces inherent limitations in memory and attention, typically capable of processing only four to seven "chunks" of information at any given time (Miller, 1956; Cowan, 2001). Consequently, even when theoretical solutions exist, their widespread propagation and adoption are impeded unless they are compressed into cognitively manageable forms. Furthermore, complex interdisciplinary insights rarely diffuse effectively across different fields and sectors, unless explicitly synthesized into concise, universally comprehensible models (Rogers, 1962; Dawkins, 1976).

Given these cognitive and cultural constraints, achieving broad consensus and effective implementation at scale requires deliberate and strategic compression of interdisciplinary insights. Such compression involves synthesizing diverse and complex ideas into a unified, explicit framework that can be readily internalized and practically deployed across disciplinary boundaries. In this context, the concept of recursive self-correction, coupled with explicit self-modeling, emerges as an essential mechanism for generating and maintaining adaptive capacity across all relevant domains. Recursive self-correction has been highlighted by second-order cybernetics as the foundational ability of systems to observe, evaluate, and iteratively revise their own processes in response to internal and external changes (Wiener, 1948; von Foerster, 1974; Ashby, 1956). Complementing this concept, self-modeling—the explicit representation of a system's internal state, capabilities, and processes—enables systems to dynamically regulate and improve their behavior (Dennett, 1991; Metzinger, 2003; Friston, 2010).

The integration of recursive self-correction and self-modeling across disciplines such as mathematics, logic, biology, cognitive science, artificial intelligence, organizational theory, and public policy reveals a remarkable cross-disciplinary consensus. However, this consensus remains largely compartmentalized and insufficiently recognized at a societal level, primarily due to cognitive limitations and institutional silos. This paper, therefore, aims to address this critical gap by presenting a concise yet powerful compression of these interdisciplinary insights into a single Functional Model of Intelligence (FMI) theorem. We rigorously demonstrate that a system capable of reliably solving novel and evolving problems in open-ended domains must inherently embody recursive self-correction supported by explicit internal self-models.

To translate this theoretical consensus into practical impact, we propose a novel, self-funding, bounty-driven campaign modeled loosely upon the successful "Humanity's Last Exam" approach. Unlike Humanity’s Last Exam, which uses a closed set of predetermined questions, this proposed initiative explicitly tests adaptive intelligence capabilities essential for managing open-ended and rapidly evolving existential risks. By integrating recursive feedback mechanisms and self-modeling capacities into the structure of this campaign, we aim not only to test but also to propagate and embed the FMI framework within institutional and societal decision-making processes.

Through these integrative measures, this paper makes explicit why recursive self-correction and self-modeling, compressed into a cognitively manageable model, represent our best—and perhaps only—practical lever for rapidly and effectively addressing the unprecedented existential risks facing human civilization today.

2. Survey of Recursive Self-Correction and Self-Modeling

2.1 Mathematics & Logic

Mathematics and logic have provided foundational insights into the intrinsic limitations of static systems, highlighting the necessity of recursive self-correction and dynamic adaptation. Gödel's incompleteness theorems (1931) established that no static axiomatic system could be both complete and consistent, implying that any sufficiently powerful formal system must remain open to revision. Similarly, Tarski (1936) demonstrated that a formal language cannot internally define its own concept of truth, necessitating an external or higher-order framework capable of continuously revising itself in response to new insights. Lakatos (1970) further argued that mathematical and scientific progress relies fundamentally on the evolution and revision of core assumptions, as exemplified by his historical analysis of mathematical research programs. In the philosophy of science, Popper (1963) emphasized that scientific advancement is inherently a process of recursive self-correction, characterized by iterative cycles of conjecture, empirical testing, and refutation, through which theories evolve towards increasingly accurate representations of reality.

2.2 Cybernetics & Control Theory

Cybernetics and control theory explicitly center the importance of systems being capable of observing, modeling, and adjusting their own operations. Wiener (1948) and von Foerster (1974) pioneered second-order cybernetics, emphasizing that effective control systems must possess an internal capacity to monitor and alter their behavior. This self-observation enables systems to respond adaptively to external changes and maintain desired outcomes. Ashby's Law of Requisite Variety (1956) further formalizes this concept by stating that a system's internal complexity must match the complexity of the environment it aims to control. Thus, cybernetic control systems explicitly require sophisticated internal models capable of updating themselves recursively, ensuring robustness and adaptability.

2.3 Cognitive Science & Artificial Intelligence

Research in cognitive science and artificial intelligence consistently highlights self-reflective and self-modifying capacities as central to flexible intelligence. Flavell (1979) introduced the concept of metacognition—awareness and control of one's cognitive processes—as a critical dimension of human learning and problem-solving. Minsky's (1986) "Society of Mind" framework similarly posits that intelligent behavior arises from collections of cognitive processes that include specialized self-monitoring modules tasked with modifying and improving lower-level functions. Extending these principles computationally, Schmidhuber's Gödel Machines (2006) formally implement recursive self-improvement by rewriting their own code whenever they can mathematically prove that the modifications will yield superior performance. Collectively, these contributions underscore the importance of explicit internal models capable of assessing and recursively enhancing their own operations.

2.4 Biology & Evolution

Biological systems vividly illustrate how recursive self-correction and self-modeling naturally arise through evolutionary processes. Kauffman (1993) demonstrated how life emerges from autocatalytic biochemical networks that recursively generate and select molecular variants, inherently functioning as self-correcting evolutionary systems. Similarly, Bonner (1998) described the evolution of multicellularity as resulting from feedback loops where collections of cells evolve mechanisms to coordinate, regulate, and correct internal functions. Extending this biological understanding to cognitive science, Dennett (1991) and Metzinger (2003) conceptualize consciousness itself as an advanced self-modeling capacity, whereby the brain continuously creates and revises internal representations of itself and its interactions with the environment. This conscious self-modeling enables sophisticated behavioral adaptations crucial for survival and success.

2.5 Institutional Design & Policy

In institutional design and policy-making, recursive self-correction and explicit self-modeling have been identified as essential for successful governance and organizational effectiveness. Ostrom (1990) documented that sustainable management of common resources typically involves polycentric institutional arrangements that embed continuous evaluation and feedback loops, allowing rapid adaptive adjustments in policy and governance structures. Similarly, Senge's (1990) influential concept of the learning organization explicitly emphasizes that organizational effectiveness arises from structured reflection processes, enabling organizations to routinely reassess and refine their internal operations and strategies. Extending these insights to large-scale systemic risk management, Helbing

(2013) advocates for "reflexive regulation," governance frameworks that explicitly monitor their effectiveness and dynamically adjust their rules in response to evolving risks, thus maintaining resilience and preventing catastrophic failures in complex global networks.

2.6 Survey of Self-Modeling Across Disciplines

2.6.1 Philosophy of Mind

Philosophy of mind extensively addresses self-modeling as fundamental to understanding consciousness and cognition. Dennett (1991) famously critiqued the notion of a singular, centralized "Cartesian Theater," advocating instead for distributed self-models that collectively generate the experience of consciousness. Complementing this, Metzinger (2003) provided a detailed theoretical framework describing consciousness as arising from the brain's transparent self-model—a sophisticated neural representation that allows the organism to seamlessly interact with its environment and modulate behavior based on continuous self-assessment.

2.6.2 Neuroscience & Predictive Coding

Neuroscientific theories further elucidate self-modeling by conceptualizing the brain as a predictive engine continually generating and updating internal representations based on sensory inputs and prior experiences. Friston (2010) described this process through the lens of predictive coding and the Bayesian brain hypothesis, in which the brain recursively updates its self-model to minimize prediction errors. Corlett and colleagues (2009) extended this framework to psychiatric contexts, suggesting that psychosis can be understood as a breakdown in the brain's self-modeling capacity, where inaccurate self-predictions lead to maladaptive behaviors and perceptions.

2.6.3 Robotics & Embodied AI

Advances in robotics and embodied artificial intelligence have similarly emphasized the centrality of self-modeling. Brooks (1991) initially introduced subsumption architectures, where robots operate without explicit internal models; however, subsequent research has demonstrated the necessity of explicit self-modeling for tasks requiring complex sensorimotor coordination. Nanayakkara et al. (2013) showed that robots capable of sophisticated tool use must form and recursively refine internal body schemas, enabling continuous improvement and adaptation to new tools and environments.

2.6.4 Computer Science & Reflection

Computer science has long recognized the power of self-referential and reflective systems capable of modifying their own operation. Quine (1960) introduced self-replicating programs—now known as "quines"—illustrating the concept of software explicitly modeling and producing its own code. Smith and Ungar (1985) advanced this concept by developing reflective towers within programming languages such as Smalltalk, enabling programs to introspectively analyze and dynamically alter their own interpreters, thereby facilitating recursive improvements and adaptive computing behaviors.

2.6.5 Organizational Learning

Organizational learning theory highlights the performance advantages of institutions that explicitly implement internal reflection and self-modeling processes. Fiol and Lyles (1985) empirically demonstrated that organizations with structured internal processes for reflecting on and revising their own operations significantly outperform static rivals. Easterby-Smith and Lyles (2003) further characterized these capabilities as "dynamic capabilities," defining organizational self-models that allow firms to continuously sense, seize, and reconfigure strategic opportunities effectively.

2.6.6 Systems Biology

Systems biology emphasizes that biological robustness emerges precisely from self-modeling networks explicitly monitoring and regulating their internal state. Kitano (2002) detailed how biological systems achieve resilience by employing networks of regulatory interactions that constantly track and adjust their own functioning. Similarly, Thomas and Kaufman (2001) illustrated that cellular network motifs serve as internal circuits capable of generating self-models that guide gene regulation, cellular differentiation, and organismal adaptation.

Why This Matters:

Recursive self-correction describes why systems must adjust their behaviors in response to changing conditions, while self-modeling explains how this adjustment occurs practically through explicit internal representations. Each surveyed discipline independently underscores the necessity of both processes. The convergence across diverse fields thus provides robust, interdisciplinary validation of the Functional Model of Intelligence as fundamentally essential—not merely beneficial—for effectively addressing complex, evolving challenges.

2.7 Leading Indicators, Trailing Indicators, and the Structural Necessity of Functional Intelligence

Across the surveyed disciplines, a critical but often implicit distinction emerges between leading and trailing indicators of system performance and epistemic accuracy. Trailing indicators—such as closed test scores, historical performance metrics, or retrospective validation—measure success only after outcomes have materialized. In relatively static domains, where environmental conditions and problem definitions remain stable, trailing indicators can be reliable proxies for future success. However, in open-ended, dynamic domains where problems, solutions, and solvers co-evolve, reliance on trailing indicators introduces an existential vulnerability: failure is detected only after it has occurred, often irreversibly.

This structural limitation has direct consequences for existential risk management. Closed problem sets, such as those employed in "Humanity’s Last Exam," evaluate systems against pre-specified criteria, thereby assuming environmental stationarity. While valuable for benchmarking static competencies, such approaches fail to account for the continuous emergence of novel, unanticipated challenges—precisely the domain in which existential risks propagate most dangerously. When the external environment shifts faster than a system’s capacity to detect and adapt, trailing indicators collapse into lagging signals, rendering interventions too late to prevent systemic failure (Helbing, 2013; Lewis, 2025a).

In contrast, leading indicators are forward-facing metrics that assess a system’s ongoing ability to adapt, self-correct, and preemptively respond to emerging challenges. The Functional Model of Intelligence (FMI) formalizes the structural conditions necessary to generate leading indicators dynamically. By embedding recursive self-correction and explicit self-modeling into every problem-solving process, an FMI-driven system continuously generates new evaluation criteria aligned with shifting realities. Instead of merely assessing performance against fixed past benchmarks, it recursively evaluates its own assumptions, models, and interventions in real time, updating its predictive and corrective strategies accordingly (Williams, 2025b).

Thus, the distinction between trailing and leading indicators is not merely one of methodological preference but one of existential necessity. Systems dependent solely on trailing indicators cannot reliably survive phase transitions across cognitive or environmental complexity thresholds. Only architectures capable of continuously generating leading indicators—through recursive functional intelligence—can maintain solvability beyond the noise limit where traditional problem-solving collapses (Williams, 2025c).

The integration of this insight completes the interdisciplinary convergence outlined in the preceding surveys. Recursive self-correction and self-modeling are not sufficient in isolation; they must be operationalized in ways that foreground leading indicators of systemic health, resilience, and adaptive capacity. This necessity further underscores why the propagation of an explicit, compressed, and universally deployable Functional Model of Intelligence constitutes a structural imperative for addressing existential risks.

3. The Functional Model of Intelligence Theorem

Reader's Guide: This section presents the central theorem demonstrating that reliably solving problems in dynamic environments structurally requires a Functional Model of Intelligence (FMI). We first provide the formal statement for precision, then offer a plain-English proof sketch to make the logical flow accessible across disciplines.

3.1 Formal Statement

We define the theorem precisely in the language of first-order logic:

$$\forall a\left[\forall p,t_1,t_2\left(Changes\left(p,t_1,t_2\right)\rightarrow\exists s\,Solves\left(a,p,s,t_2\right)\right)\rightarrow FMI\left(a\right)\right]$$

In plain English: *For any agent a, if that agent solves every problem p after it changes between times t_1 and t_2, then the agent must implement a Functional Model of Intelligence (FMI).*

Key Definitions to Understand the Statement:

- **Agent (a)**: Any entity attempting to solve problems.
- **Problem (p)**: Any task or challenge faced by the agent.
- **Changes(p,t_1,t_2)**: Indicates that the problem p changed between times t_1 and t_2.
- **Solves(a,p,s,t)**: Means that agent a finds a solution s to problem p at time t.
- **FMI(a)**: Means that agent a possesses a Functional Model of Intelligence: an explicit internal structure capable of recursive self-correction and self-modeling.

3.2 Proof Sketch (Plain English)

To help readers understand the logic behind the formal proof (the full formal derivation is provided in Appendix A), we offer a simplified step-by-step explanation here:

Step 1: Premise — Static agents fail in changing environments.

We start with an accepted principle (based on decades of research across mathematics, cybernetics, cognitive science, and control theory):

> Any agent that **does not** internally model itself—i.e., does not possess an FMI—**will eventually fail** to solve some problems after those problems change.

This is because without recursive self-correction, the agent cannot flexibly update its methods or strategies when reality shifts.

Step 2: Assume the opposite — An agent succeeds but has no self-model.

For the sake of logical proof (using a method called *reductio ad absurdum* or "proof by contradiction"), we **assume** that there exists an agent that:

- Solves **every** problem after it changes.

- Yet **does not** implement a Functional Model of Intelligence.

In other words, we assume that pure static strategies—without any internal self-model—can still perfectly adapt to every environmental shift.

Step 3: Derive a contradiction.

Given the premise from Step 1, any agent lacking a self-model must eventually fail on some changed problem.

But by our assumption in Step 2, the agent **never** fails.

Thus, we are forced into a logical contradiction:

- The agent both **fails** and **succeeds** on the same kind of evolving problem.
- This is logically impossible.

Step 4: Conclude the agent must have an FMI.

Because assuming "no FMI" leads to contradiction, the assumption must be false.

Thus, the agent **must** have a Functional Model of Intelligence.

This completes the proof:

> **Reliable problem-solving in open, changing environments requires recursive self-correction and explicit self-modeling.**

This theorem shows that an FMI is not merely an optional enhancement for intelligence, but a **structural necessity** if an agent is to remain capable of solving problems as the world evolves.

The full formal derivation, using first-order logic in Fitch-style notation, can be found in **Appendix A** for readers who wish to examine the rigorous step-by-step construction.

4. Closed Tests vs. Open-Ended Intelligence

Humanity's Last Exam amassed 731 signatories by focusing on a **closed** question set—effective for static domains but inadequate for emerging existential threats. **Open-ended** intelligence demands the ability to **question and revise** the entire question set, a capacity that only an FMI can deliver.

Why Compression Enables Propagation

We previously highlighted compression briefly but did not fully unpack its critical role in propagating insights across disciplines and wider audiences. Understanding this necessity in depth involves appreciating several foundational concepts:

4.1 Human Cognitive Limits

A core obstacle to interdisciplinary consensus and rapid idea propagation is the intrinsic cognitive constraint of human memory and attention. Psychological research demonstrates a well-known "working-memory bottleneck": humans can reliably manage only about 4–7 "chunks" of information at once (Miller, 1956; Cowan, 2001). When presented with numerous distinct arguments, even specialists struggle to retain and integrate them effectively.

Compression, thus, becomes a practical cognitive tool—collapsing numerous discipline-specific findings into a single coherent and compact theorem, the Functional Model of Intelligence (FMI). Instead of overwhelming individuals with countless fragmented insights from cybernetics, mathematics, biology, cognitive science, and policy studies, FMI synthesizes them into a small, manageable set of core principles. Each principle can subsequently be unpacked into detailed, domain-specific insights as required, significantly lowering cognitive load.

4.2 Diffusion & Viral Memes

Beyond individual cognitive constraints, the propagation of ideas through societies and institutions depends heavily on simplicity and coherence. Rogers' classic work on the diffusion of innovations (Rogers, 1962) demonstrates that ideas spread faster and more reliably through social networks when they are perceived as clear, understandable, and easily transmissible. Conversely, overly complex or fragmented messages tend to stall and fail to achieve broad acceptance.

Additionally, drawing from memetics as popularized by Dawkins (1976), the most "viral" ideas—the ones most likely to achieve broad societal impact—are those with high information density but low cognitive demand. The FMI theorem fits perfectly into this criterion: it encapsulates decades of interdisciplinary research into a single, easily communicable concept, greatly enhancing its transmissibility across different communities and media channels.

4.3 Consensus Building in Practice

The FMI theorem's compressed form plays a critical "Rosetta Stone" function across disciplines. Often, domain experts remain siloed within their specialized jargon, making it challenging to recognize commonality with findings from other fields. A compact, cross-disciplinary schema allows experts to see parallels directly—demonstrating explicitly: "This insight from biology maps exactly onto that concept in AI, aligns with this finding in organizational theory, and corresponds precisely to this principle from cybernetics."

Furthermore, this compression is crucial for practical methods such as the "Conditional Acceptance Protocol," where experts across disciplines are asked to adopt the FMI model provisionally and test it within their own contexts. Such an approach is only feasible if the model itself is succinct enough to be mentally grasped, remembered, and applied without extensive study or continuous external reference.

> **Compression Principle:** *An idea must be compressed into a cognitively manageable form (≤7 chunks) to achieve cross-disciplinary adoption and viral diffusion.*

4.4 Integration with the Crowdfunded Bounty Campaign

This cognitive and practical necessity of compression directly informs our bounty campaign strategy. The campaign includes the creation of a concise "Model Brief"—a one-page, highly compressed representation of the FMI theorem. This brief serves as a cognitively optimized entry point for donors and solvers, allowing them to internalize and apply the principles instantly and practically. By leveraging cognitive science and memetic insights, the Model Brief ensures rapid and wide dissemination of FMI principles, critical for mobilizing effective and scalable responses to existential risks.

In sum, compression is not merely beneficial but fundamentally necessary for the propagation of sophisticated, cross-disciplinary insights at a scale sufficient to impact civilizational resilience. Only through strategic compression into a succinct, universally comprehensible framework can the Functional Model of Intelligence hope to rapidly propagate and provide the critical adaptive capabilities required for humanity's survival and flourishing.

5. Crowdfunded Bounty Campaign for Existential-Risk Mitigation

To operationalize the Functional Model of Intelligence (FMI) as both a testable framework and a practical engine of collective adaptation, we introduce an open-access bounty campaign that is accessible to all, whether inside or outside the Recursive Alignment Jam (RAJ). This campaign builds upon the legacy of Humanity's Last Exam by replacing fixed-question benchmarks with a generative, recursively adaptive platform that directly rewards structural alignment with FMI principles.

5.1 Overview of the Bounty Process

The core idea is simple: donors define a domain of urgent concern (e.g., AI alignment, biosafety, climate instability), and solvers propose recursively adaptive solutions capable of self-monitoring and refinement. The process unfolds in five recursive phases:

1. **Donor Submission**: Donors identify existential-risk categories or institutional failure modes they wish to address, contributing to publicly visible bounty pools. These bounties may be general (e.g., "solutions for climate-induced tipping points") or specific (e.g., "recursive regulation of international carbon offsets").
2. **Model Brief Delivery**: Each participant receives a cognitively compressed one-page "Model Brief" of the Functional Model of Intelligence. This brief presents the FMI theorem, its structural implications, and guiding heuristics for recursive solution design. It ensures that even solvers unfamiliar with the theory can adopt its scaffolding without prerequisite expertise.
3. **Solver Contributions**: Anyone—policy expert, systems theorist, domain practitioner, or student—may submit:
    - Investigations of real institutional processes (including within their own organization)
    - Reframed problem definitions that expose recursively misaligned structures
    - Recursive models of intervention, aligned with the FMI framework
    - Critiques or falsifications of the FMI theorem itself
4. **AI-Augmented Scoring & Recursive Review**: An open-source language model applies a rubric based on FMI principles to assign initial scores. All rubric outputs are open to recursive challenge: participants may submit counter-analyses, alternate interpretations, or adversarial refinements. These recursive reviews become part of the public record and dynamically update both rubrics and scores.
5. **Award and Reflection**: Top-ranked entries in each category receive bounty rewards. Crucially, the system logs all recursive refinements as valid contributions—so those who challenge flawed evaluations may be rewarded alongside those who submitted original entries. This creates an evolutionary fitness landscape for ideas, not just answers.

5.2 Ethical and Transparent Disclosure

This paper and its associated infrastructure—including the FMI theorem, the bounty model, and the survey synthesis—were generated through structured interaction with a large language model (LLM). While the resulting outputs are consistent with peer-reviewed theory across multiple disciplines, they should not be considered immune from epistemic error.

We therefore invite all participants to treat the FMI framework and this paper itself as part of the open bounty structure: every assumption, derivation, and implementation detail is available for recursive validation, critique, and replacement. In particular:

- **Prove or falsify** the FMI theorem through logical derivation or empirical analysis.
- **Trace alignment failure** in organizations or policies to the absence of FMI structure.
- **Demonstrate** superior functional models that meet or exceed FMI's adaptive performance.
- **Interrogate** the language model's own meta-evaluation rubric for implicit biases or errors.

5.3 Open Participation and Local Relevance

Participants are encouraged to begin from within their own sphere of influence—investigating the recursive capacities or failure points of any process or organization that matters to them. This could include:

- A local government's climate adaptation framework
- An NGO's transparency and funding decision protocols
- A university's research incentives or peer-review system
- A biotech lab's biosafety escalation procedures

Documentation, guides, and minimal onboarding scaffolds will be made available as recursive walk-throughs, including:

- A "Recursive Self-Modeling Template" for institutional analysis
- A "Conditional Acceptance Protocol" for local FMI implementation trials
- A "Challenge the Model" guide to assist in formal refutation attempts

These resources will be open-source and modifiable through recursive contribution.

5.4 A Civic Intelligence Test for All of Us

The bounty campaign is not merely a contest of solutions; it is a live demonstration of whether our species can build adaptive institutions before our static ones fail. It asks a question:

> **Can we collectively model, test, and recursively improve the very cognitive and institutional architectures through which we act?**

To answer that question is to participate in the first truly adaptive intelligence exam—not of any individual, but of a civilization.

We invite all who read this paper to join the campaign—not to trust its conclusions, but to prove them right, prove them wrong, or recursively refine them into something better.

6. Discussion: Beyond Theorem—Toward Civilizational Phase Transition

The proof of the Functional Model of Intelligence (FMI) formalizes a foundational principle: in any dynamic, open-ended domain where problems, solvers, and environments evolve, only systems capable of recursive self-correction and explicit self-modeling can reliably sustain problem-solving capacity. Yet the implications of this result extend far beyond theoretical elegance. They point to a concrete structural condition for civilizational survival and flourishing in the face of escalating existential risks. Across multiple disciplines—including mathematics (Gödel, 1931; Tarski, 1936), biology (Kauffman, 1993; Bonner, 1998), systems theory (Ashby, 1956; Ostrom, 1990), and cybernetics (Wiener, 1948; von Foerster, 1974)—a coherent pattern emerges: there exists a cognitive and systemic phase change boundary beyond which humanity's ability to solve existential risks collapses unless higher-order, recursively adaptive intelligence architectures are implemented. Without such architectures, complexity and uncertainty surpass human cognitive resolution, resulting in structural unsolvability (Williams, 2025a; Williams, 2025c).

The FMI represents a minimal formalism capable of recursively stabilizing and amplifying problem-solving ability across disciplines. It acts as a cognitive near-singularity—a compressed attractor that, if propagated to a critical mass of adopters, could trigger a phase transition into a higher-order decentralized collective intelligence (DCI). In such a system, recursive correction and distributed epistemic updating would structurally enable humanity to manage risks ranging from unaligned artificial general intelligence (AGI) to biosynthetic threats, climate destabilization, and systemic collapse.

However, if the FMI is first operationalized solely through centralized AGI development pathways—without concurrent human-scale propagation—civilization risks falling irreversibly into what has been described as the AGI-first attractor (Williams, 2025c). This centralized technology gravity well would be characterized by epistemic monoculture, collapse of alignment flexibility, and irreversible loss of distributed agency. In this scenario, cognitive control infrastructure would converge into a small number of dominant systems, rendering decentralized correction and re-alignment functionally impossible once lock-in occurs.

The noise limits formalized in the theory of conceptual near-singularities (Williams, 2025b) further suggest that once the rate and complexity of environmental and systemic change exceed human cognitive resolution, problems cannot be solved merely by scaling conventional institutional models. Without crossing a functional intelligence threshold—through recursive self-correction and explicit self-modeling—the volume of solvable problems collapses beyond recovery.

Thus, the propagation of the Functional Model of Intelligence is not merely an academic curiosity. It constitutes a necessary structural precondition for sustaining humanity's capacity to address existential risks and navigate increasingly complex global challenges.

7. A Civilizational Strategy: Escaping the Centralized Gravity Well

Recognizing the gravity of this inflection point, the paper proposes a concrete strategy for leveraging the Functional Model of Intelligence to steer civilization toward a decentralized, adaptive trajectory. First, the deployment of crowdsourced bounty campaigns anchored on the FMI framework can enable broad, open-ended participation in existential problem-solving across multiple domains. Unlike closed, static assessments, these bounties explicitly reward the generation, refinement, and recursive correction of problem definitions and solutions. In doing so, they operationalize the core principles of FMI, embedding recursive self-modeling and adaptability into the incentive structures themselves. Tracking uptake and solution quality relative to closed-test baselines, such as those demonstrated by "Humanity's Last Exam," will provide empirical evidence of the comparative scalability and resilience of the FMI approach.

Second, compression of the core FMI insights into cognitively transmissible formats is essential. Cognitive science has consistently demonstrated that working memory constraints (Miller, 1956; Cowan, 2001) severely limit the propagation of complex ideas unless they are compressed into a small number of conceptual "chunks." By synthesizing recursive self-correction and self-modeling into a minimal functional scaffold, FMI satisfies these cognitive transmission constraints, allowing rapid diffusion across disciplines and networks. This compression is critical not only for education and onboarding but also for preserving epistemic resilience as the system scales.

Third, distributed testing and recursive correction protocols must be embedded from the outset. Rather than relying on static evaluation or premature standardization, the adoption of FMI must proceed through conditional, adaptive validation: encouraging users and institutions to provisionally apply the model, identify failure points, and recursively refine their implementations. This approach mirrors the very principles that FMI formalizes, enabling its own self-stabilization and self-improvement across expanding epistemic networks.

Through these strategic steps, it may be possible to escape the centralized gravity well of AGI-first lock-in and to build a distributed cognitive infrastructure robust enough to handle both today's crises and the accelerating challenges of tomorrow. The discounted future value of such a decentralized, adaptively recursive network—both in human survival and flourishing—vastly exceeds any foreseeable near-term cost. Securing this future is not merely an opportunity; it is a structural imperative.

8. Conclusion: FMI as Humanity’s Phase Transition Lever

The Functional Model of Intelligence (FMI) theorem formalizes a minimal and necessary principle: in any open-ended, dynamic domain where problems, solvers, and environments co-evolve, only agents capable of recursive self-correction and explicit self-modeling can reliably sustain problem-solving capacity. Across mathematics, biology, cybernetics, cognitive science, and institutional theory, the same structural necessity emerges: without internal mechanisms to recursively model and adapt to changing environments, systems eventually fail.

Yet the significance of this finding extends far beyond theoretical unification. The adoption and propagation of FMI represents a potential cognitive near-singularity—a phase transition in humanity’s collective problem-solving architecture. If a critical mass of human agents and institutions embed FMI principles into their adaptive processes, it becomes possible to construct a decentralized collective intelligence (DCI) system resilient enough to manage existential risks across domains, from AI alignment and biosecurity to climate destabilization and systemic governance collapse.

Conversely, if FMI is operationalized first and exclusively through centralized AGI development paths—absent widespread human-scale adoption—civilization risks irreversible collapse into the centralized technology gravity well. In such a scenario, epistemic monoculture, alignment failure, and the permanent loss of distributed human agency would likely preclude any future capacity for decentralized re-alignment. Control over cognitive infrastructure would converge irreversibly into narrow, fragile systems beyond meaningful external influence.

This bifurcation is not speculative. The theory of conceptual near-singularities (Williams, 2025b) demonstrates that beyond certain complexity and noise thresholds, problems become unresolvable by traditional cognitive architectures. Without crossing a functional intelligence threshold via recursive self-correction and self-modeling, the set of solvable problems shrinks asymptotically toward zero. Merely “working harder” within existing institutions is insufficient; an architectural phase change is required.

Compression thus becomes the critical lever for propagating FMI. Human working memory and social diffusion dynamics demand that insights be packaged into cognitively transmissible forms. By collapsing cross-disciplinary insights into a compact, functional model, and coupling this model to an open adaptive bounty structure that rewards recursive correction, humanity can still leverage its distributed intelligence before centralized technological collapse locks in.

The discounted future value of this decentralized collective intelligence—measured not only in survival probabilities but in the flourishing potential of post-transition civilization—vastly outweighs any near-term costs associated with FMI propagation and testing. This is not merely a theoretical observation; it constitutes a strategic imperative.

Recursive self-modeling, properly propagated, is humanity’s only known structural lever for escaping the centralized gravity well and building a future in which existential risks are navigated with resilience, creativity, and collective wisdom.

> **In short:** humanity faces a closing window. We either propagate the ability to recursively adapt before the world outpaces our capacity to reason, or we are outpaced—and overwritten—forever.

References

Armstrong, S., & Sotala, K. (2012). How we're predicting AI—or failing to. AI & Society, 27(4), 507-521.

Ashby, W. R. (1956). An Introduction to Cybernetics. Chapman & Hall.

Bonner, J. T. (1998). The Evolution of Complexity.

Bonner, J. T. (1998). The Origins of Multicellularity. Integrative Biology: Issues, News, and Reviews, 1(1), 27-36.

Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford University Press.

Brooks, R. A. (1991). Intelligence without representation. Artificial Intelligence.

Corlett, P. R. et al. (2009). Prediction error, psychosis, and the brain's self-model. Trends in Cognitive Sciences.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. Behavioral and Brain Sciences, 24(1), 87–114.

Dawkins, R. (1976). The Selfish Gene. Oxford University Press.

Dennett, D. C. (1991). Consciousness Explained. Little, Brown and Co.

Easterby-Smith, M., & Lyles, M. A. (2003). The Blackwell Handbook of Organizational Learning & Knowledge Management.

Fiol, C. M., & Lyles, M. A. (1985). Organizational learning. Academy of Management Review.

Folke, C., Hahn, T., Olsson, P., & Norberg, J. (2005). Adaptive governance of social-ecological systems. Annual Review of Environment and Resources, 30, 441-473.

Friston, K. (2010). The free-energy principle: a unified brain theory? Nature Reviews Neuroscience, 11(2), 127-138.

Gödel, K. (1931). Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. Monatshefte für Mathematik und Physik, 38(1), 173-198.

Helbing, D. (2013). Globally networked risks and how to respond. Nature, 497(7447), 51–59.

Kauffman, S. A. (1993). The Origins of Order: Self-Organization and Selection in Evolution. Oxford University Press.

Kitano, H. (2002). Systems biology. Nature.

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In I. Lakatos & A. Musgrave (Eds.), Criticism and the Growth of Knowledge (pp. 91-196). Cambridge University Press.

Metzinger, T. (2003). Being No One: The Self-Model Theory of Subjectivity. MIT Press.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review, 63(2), 81–97.

Minsky, M. (1986). The Society of Mind.

Nanayakkara, S. et al. (2013). Robot self-models and tool use. Science Robotics.

Ostrom, E. (1990). Governing the Commons: The Evolution of Institutions for Collective Action. Cambridge University Press.

Perrow, C. (1999). Normal Accidents: Living with High-Risk Technologies. Princeton University Press.

Popper, K. (1963). Conjectures and Refutations: The Growth of Scientific Knowledge. Routledge.

Quine, W. V. O. (1960). Quine programs ("quines"). Journal of Symbolic Logic.

Roberts, K. H. (1990). Some characteristics of one type of high reliability organization. Organization Science, 1(2), 160-176.

Rogers, E. M. (1962). Diffusion of Innovations. Free Press.

Rockström, J., Steffen, W., Noone, K., et al. (2009). A safe operating space for humanity. Nature, 461(7263), 472-475.

Schmidhuber, J. (2006). Gödel Machines. Cognitive Systems Research.

Senge, P. M. (1990). The Fifth Discipline: The Art & Practice of the Learning Organization. Doubleday.

Smith, D., & Ungar, D. (1985). Reflection in Smalltalk. Conference on Object-Oriented Programming.
Taleb, N. N. (2012). Antifragile: Things That Gain from Disorder. Random House.
Tarski, A. (1936). The concept of truth in formalized languages. Studia Philosophica, 1, 261-405.
von Foerster, H. (1974). Cybernetics of Cybernetics. In K. Krippendorff (Ed.), Communication and Control in Society (pp. 5–8). Gordon and Breach.
Walker, J. S. et al. (2003). Biological risk assessment. Biosecurity Journal.
Wiener, N. (1948). Cybernetics: Or Control and Communication in the Animal and the Machine. MIT Press.
Williams, A. E. (2025a). The Functional Threshold of Intelligence. Unpublished manuscript.
Williams, A. E. (2025b). Conceptual Near-Singularity: Quantifying Intelligence Expansion Beyond Axiomatic Boundaries. Unpublished manuscript.
Williams, A. E. (2025c). Conceptual Space, Semantic Density, and the Technology Gravity Well. Unpublished manuscript.
Yudkowsky, E. (2008). Artificial Intelligence as a positive and negative factor in global risk. In N. Bostrom & M. Ćirković (Eds.), Global Catastrophic Risks (pp. 308-345). Oxford University Press.

Appendix A. Full Fitch-Style Derivation of the Functional Model of Intelligence Theorem.

Theorem

$\forall a\ [\ \forall p, t_1, t_2\ (\text{Changes}(p, t_1, t_2) \rightarrow \exists s\ \text{Solves}(a, p, s, t_2)) \rightarrow \text{FMI}(a)\ ]$

In plain language: If an agent can solve every problem after it changes, then that agent must implement a Functional Model of Intelligence (FMI).

Definitions

- a: an agent (a cognitive system or problem-solver)
- p: a problem instance
- s: a proposed solution
- t_1, t_2: two points in time
- $\text{Changes}(p, t_1, t_2)$: predicate meaning that problem p has changed between times t_1 and t_2
- $\text{Solves}(a, p, s, t)$: predicate meaning agent a uses solution s to solve problem p at time t
- $\text{FMI}(a)$: predicate meaning that agent a implements a Functional Model of Intelligence

Step-by-Step Proof with Explanations

1. Let a_0 be arbitrary.
   → We begin by choosing an arbitrary agent a_0. If the proof works for any a_0, it will work for all agents.
2. Assume: $\forall p, t_1, t_2\ (\text{Changes}(p, t_1, t_2) \rightarrow \exists s\ \text{Solves}(a_0, p, s, t_2))$.
   → We assume that this agent can solve every problem after it changes.
3. Assume: $\neg\text{FMI}(a_0)$.
   → We temporarily assume that this agent does not implement a Functional Model of Intelligence.
4. From Axiom 2 and (3): $\exists p, t_1, t_2\ (\text{Changes}(p, t_1, t_2) \wedge \neg\exists s\ \text{Solves}(a_0, p, s, t_2))$.
   → Axiom 2 says that if an agent is not FMI, there must be at least one problem it cannot solve after it changes. We apply this axiom to a_0.
5. Let p_0, t_{10}, t_{20} be such that $\text{Changes}(p_0, t_{10}, t_{20}) \wedge \neg\exists s\ \text{Solves}(a_0, p_0, s, t_{20})$.
   → We pick a specific problem and times that demonstrate this failure.
6. From (5): $\text{Changes}(p_0, t_{10}, t_{20})$.
   → We extract the part of the conjunction that says the problem changed.
7. From (2): $\text{Changes}(p_0, t_{10}, t_{20}) \rightarrow \exists s\ \text{Solves}(a_0, p_0, s, t_{20})$.
   → Our original assumption (step 2) says that if a problem changes, a_0 will solve it. We apply this to our chosen problem and times.
8. From (6) and (7): $\exists s\ \text{Solves}(a_0, p_0, s, t_{20})$.
   → Since the problem did change, a_0 must be able to solve it.
9. From (5): $\neg\exists s\ \text{Solves}(a_0, p_0, s, t_{20})$.
   → But earlier we said a_0 cannot solve it. That contradicts step 8.
10. From (8) and (9): $\bot$
    → We now have a contradiction: a_0 both can and cannot solve the same problem.
11. Therefore: $\neg\neg\text{FMI}(a_0)$.
    → Because our assumption that a_0 is not FMI led to a contradiction, that assumption must be false.
12. Therefore: $\text{FMI}(a_0)$.
    → If it's not true that a_0 is not FMI, then a_0 must be FMI.
13. Therefore: $\forall p, t_1, t_2\ (\text{Changes}(p, t_1, t_2) \rightarrow \exists s\ \text{Solves}(a_0, p, s, t_2)) \rightarrow \text{FMI}(a_0)$.
    → We conclude that if a_0 solves every changed problem, it must be FMI.
14. Therefore: $\forall a\ [\ \forall p, t_1, t_2\ (\text{Changes}(p, t_1, t_2) \rightarrow \exists s\ \text{Solves}(a, p, s, t_2)) \rightarrow \text{FMI}(a)\ ]$
    → Since a_0 was arbitrary, this holds for all agents.