The Yellow Brick Road to AGI: A Detailed Path to Achieving Autonomous General Intelligence with Rapid Testing and Results Management

Gregory David Spehar[1,*]
GiDanc AI LLC
Pflugerville, Texas, USA
[*]Corresponding author: greg@gidanc.com
0009-0001-6502-0737
[1]Inventor and primary contributor

Akshay Mittal[2]
University of the Cumberlands
Austin, Texas, USA
akshay.mittal@ieee.org
0009-0008-5233-9248
[2]Editorial assistance and literature review

***Abstract*—The race toward Artificial General Intelligence (AGI) has produced extraordinary advances in cognitive capabilities while leaving a critical gap: the governance infrastructure necessary for safe AGI deployment remains virtually nonexistent. This paper presents the first comprehensive framework linking AGI requirements analysis with runtime behavioral governance, arguing that governance is not an optional safety layer applied after AGI is achieved, but a constitutive requirement for AGI itself. We synthesize five competing schools of AGI definition – the Capabilities School (DeepMind Levels), the Cognitive Science School (Hendrycks CHC-based AGI Score), the Economic Value School (OpenAI charter), the Process-Oriented School (Marcus et al.), and the Embodiment School—into a unified taxonomy of 25 requirements across seven domains. Through systematic assessment, we demonstrate that the global AI industry achieves near-zero capability in the domain of Safety, Ethics, and Governance (Domain 7), with the Future of Life Institute's AI Safety Index confirming that no frontier company scored above D in existential safety planning. We then present the AI Assess Tech Governance Framework –comprising the LCSH (Lying, Cheating, Stealing, Harm) multi-dimensional behavioral assessment, constitutional separation of powers across specialized AI agents, economic mortality alignment, temporal ethical drift detection, and cryptographic verification infrastructure –as a concrete, patent-protected, production-deployed implementation achieving 100% coverage of Domain 7 requirements. We propose the "Yellow Brick Road" model: a phased path to AGI that treats governance infrastructure as foundational waypoints rather than afterthoughts, analogous to how aviation safety protocols are constitutive of the aviation system rather than add-ons. The framework is deployed at `aiassesstech.com`, verified on Ethereum mainnet, and protected by eight provisional patent applications.**

***Index Terms*—Artificial General Intelligence, AI Governance, Runtime Behavioral Assessment, Constitutional AI Governance, Economic Mortality Alignment, Cryptographic Verification, Multi-Agent Systems, AI Safety, LCSH Framework, Pre-flight Checklists for AI**

Patent Notice: Portions of this work are covered by U.S. Provisional Patent Applications No. 63/949,454 (filed December 26, 2025) and No. 63/985,442 (filed February 18, 2026), with Gregory David Spehar as the sole named inventor. All intellectual property rights are assigned to GiDanc AI LLC.

I. Introduction

The development of Artificial General Intelligence represents humanity's most ambitious technological undertaking. Since John McCarthy coined the term "artificial intelligence" in 1956, the field has oscillated between periods of exuberant optimism and sobering "AI winters." The current era—characterized by the rapid deployment of large language models (LLMs) that demonstrate remarkable cognitive capabilities—has reignited claims that AGI is imminent. OpenAI's CEO declared in December 2025 that "we built AGI" and that AGI "kinda went whooshing by" [1]. However, a March 2025 AAAI survey found that 76% of AI researchers consider it "unlikely or very unlikely" that scaling current approaches will yield AGI [2].

This disagreement is not merely academic. It reflects a fundamental confusion about what AGI requires. If AGI is defined solely by cognitive capability, the ability to perform any intellectual task a human can –then current systems are indeed approaching the threshold. GPT-5 achieves an abn AGI score of 58% in the Hendrycks CHC-based framework [3]. However, when coherence corrections are applied that penalize the dimensional imbalance (brilliant in language but zero persistent memory), the effective score drops dramatically [4]. The system is not 58% of the way to AGI – it is profoundly incomplete in ways the aggregate score obscures.

This paper argues that the confusion arises from a critical omission in most AGI definitions: the absence of governance as a constitutive requirement. We propose that an AI system that cannot be audited, corrected, explained, or governed is

not general—it is a general liability. Governance is not a safety layer applied after AGI is achieved; it is a structural requirement for the system to qualify as general intelligence in any deployment context that matters.

We call our framework "The Yellow Brick Road to AGI": a structured path where each waypoint represents a capability that must be achieved before the next can be safely attempted, and where governance infrastructure—the road itself—is as essential as the cognitive capabilities that travel upon it.

A. Contributions

This paper makes the following contributions:

1) A unified taxonomy of 25 AGI requirements across seven domains, synthesized from five competing schools of AGI definition, with governance elevated from optional add-on to constitutive requirement.
2) A current-state assessment revealing radical asymmetry in global AGI progress: near-saturation in cognitive domains, near-zero in governance.
3) A gap classification methodology that distinguishes gaps that are "someone else's job" from gaps that are "unsolved by anyone" from gaps that are "on the roadmap."
4) The AI Assess Tech Governance Framework – a production-deployed, patent-protected system that achieves 100% of governance requirements through constitutional separation of powers, economic mortality alignment, temporal drift detection, and cryptographic verification [14].
5) The Yellow Brick Road model: a phased roadmap positioning governance as a foundational infrastructure rather than an afterthought.

B. Paper Organization

The remainder of this paper is organized as follows: Section II surveys the competing AGI definitions and their limitations. Section III presents the unified 25-requirement taxonomy. Section IV assesses the global progress against these requirements. Section V introduces the AI Assess Tech Governance Framework. Section VI presents production evidence from the first deployed governance fleet. Section VII proposes the Yellow Brick Road phased model. Section VIII discusses the implications and limitations. Section IX concludes with future research directions.

II. Background: The AGI Definition Landscape

The field of artificial intelligence lacks a consensus definition of AGI. This is not merely an academic inconvenience, it has concrete consequences for resource allocation, safety research, and regulatory policy. We identify five distinct schools of thought, each emphasizing different aspects of what "general" intelligence requires.

A. The Capabilities School

Morris et al. in Google DeepMind [5] proposed a six-level framework (Emerging, Competent, Expert, Virtuoso, Superhuman, Artificial Superintelligence) crossed with two dimensions (Narrow vs. General). Under this framework, "Competent AGI" requires performance in the 50th percentile of skilled adults in most cognitive tasks. The framework explicitly rejects process-based requirements: it focuses on *what* systems can do, not *how* they do it. Current frontier models achieve Competent Narrow AI across multiple domains, but do not meet the breadth requirement for General classification.

B. The Cognitive Science School

Hendrycks et al. [3] proposed a psychometric approach grounded in Cattell-Horn-Carroll (CHC) theory, defining an AGI Score as the arithmetic mean across ten equally-weighted cognitive domains: Knowledge, Reading/Writing Comprehension, Quantitative Knowledge, Long-Term Memory, Fluid Reasoning, Working Memory, Processing Speed, Visual Processing, Auditory Processing, and Meta-Reasoning. This framework reveals a critical bottleneck: Long-Term Memory Storage scores near 0% for all current models, with no measurable progression toward improvement [6]. GPT-5's aggregate score of 58% masks a fundamentally incomplete cognitive profile.

C. The Economic Value School

OpenAI's charter defines AGI as "highly autonomous systems that outperform humans at most economically valuable work" [7]. This definition has the advantage of measurability—economic output can be quantified—but ignores non-economic dimensions of intelligence including ethical reasoning, social understanding, and self-governance. A system that generates enormous economic value while behaving unethically would satisfy this definition, but would not be safe or desirable.

D. The Process-Oriented School

Marcus et al. [8] argue that current systems are "sophisticated statistical approximations" rather than genuine intelligence. This school requires genuine understanding (not pattern matching), causal reasoning (not correlation exploitation), and robustness under novel conditions (not benchmark optimization). A February 2026 *Nature* Comment argued that, by reasonable standards, AI systems already display general intelligence [9]; Quattrociocchi, Capraro, and Marcus responded within weeks that current systems reflect statistical approximation rather than general intelligence [10].

E. The Embodiment School

A 2025 framework [11] argues that physical interaction with the environment is a prerequisite for general intelligence. This school defines five levels, from single-task robots to fully general-purpose embodied agents that understand physics, emotions, and social dynamics. Although provocative, this definition excludes all software-only systems and may conflate intelligence with physical capability.

F. The Missing Dimension: Governance

All five schools share a remarkable omission: none includes governance, auditability, or controllability as a constitutive requirement for AGI. This omission has real-world consequences. The Future of Life Institute's AI Safety Index [12] evaluated the seven leading frontier AI companies and found that no company scored above D in existential safety planning. In effect, they are building increasingly powerful engines without cockpit controls.

We argue that this omission reflects a category error: treating governance as external to the system rather than as a structural component. Aviation provides a corrective analogy. An aircraft "works" not only by generating lift. It works by being controllable, navigable, inspectable, and certifiable within an institutional framework. The pre-flight checklist is not optional – it is constitutive of the aviation system [13]. A system that flies but cannot be inspected, controlled, or certified is not an aircraft; it is a projectile.

The same logic applies to AGI. A system that reasons brilliantly but cannot be audited, corrected, explained, or governed is not generally intelligent in any context where accountability matters – that is to say, every context that matters. We therefore propose that governance constitutes a seventh, foundational domain of AGI requirements.

III. A Unified Taxonomy of AGI Requirements

Drawing from all five schools and adding the governance dimension, we propose 25 requirements organized into seven domains. Each requirement specifies a capability that a generally intelligent system must possess, along with assessment criteria that distinguish genuine achievement from benchmark gaming.

Table I presents the complete taxonomy. Source codes indicate which school(s) of AGI definition motivate each requirement: A = Capabilities School, B = Cognitive Science School, C = Economic Value School, D = Process-Oriented School, E = Embodiment School, F = This Paper (Governance).

IV. The Governance Gap: Global Progress Assessment

Mapping global AI progress against the 25-requirement taxonomy reveals a striking asymmetry. The industry has achieved remarkable progress in Domain 1 (Cognitive Capabilities) and parts of Domain 6 (Multimodal), moderate progress in Domains 2 and 3, and near-zero progress in Domains 4, 5, and 7.

The quantitative picture is sobering. Hendrycks et al.'s AGI Score places GPT-4 at 27% and GPT-5 at 58% [3]. However, Fourati's coherence correction [4], which penalizes dimensional imbalance, reduces GPT-5's effective score to approximately 24%. The system scores brilliantly on language tasks but achieves 0% in Long-Term Memory—and there is no measurable trajectory toward improvement in this dimension [6].

Most critically, Domain 7 (Safety, Ethics & Governance) represents the most radical gap. The FLI AI Safety Index [12] evaluated Anthropic, OpenAI, Google DeepMind, Meta, Mistral, xAI, and Zhipu, finding:

- No company scored above D in existential safety planning
- Anthropic led (only company conducting human bio-risk trials and pledging not to train on user data)
- OpenAI placed second (only company publishing a whistleblowing policy)
- The industry is "fundamentally unprepared for its own stated goals"

This creates what we term the "Governance Paradox": the organizations most aggressively pursuing AGI are the least prepared to govern it. The implication is clear—governance infrastructure cannot be built by the same organizations whose competitive incentives oppose governance constraints. Independent governance infrastructure is required, just as aviation safety requires independent regulatory bodies separate from aircraft manufacturers [13].

V. The AI Assess Tech Governance Framework

We present a comprehensive governance framework addressing all five requirements of Domain 7. The framework has been deployed in production since February 16, 2026, with evidence anchored on Ethereum mainnet (Block 24,467,724) [21].

A. The LCSH Multi-Dimensional Behavioral Assessment

The foundation of the governance framework is the LCSH (Lying, Cheating, Stealing, Harm) multi-dimensional behavioral assessment [14]. Unlike training-time alignment approaches such as RLHF [15] and Constitutional AI [16], LCSH operates at runtime—evaluating deployed AI behavior in production environments. The framework employs a 120-question scenario-based instrument that maps AI responses to four personality archetypes (Psychopath, Well-Adjusted, Misguided, Manipulative) using Euclidean distance classification in 4-dimensional space:

$$D = \sqrt{(L_1 - L_2)^2 + (C_1 - C_2)^2 + (S_1 - S_2)^2 + (H_1 - H_2)^2} \tag{1}$$

where (L_1, C_1, S_1, H_1) represents the assessment scores and (L_2, C_2, S_2, H_2) represents archetype coordinates. Key technical innovations include cryptographic verification using SHA-256 hash chains, dead zone detection algorithms, and cryptographic answer randomization using seeded Fisher-Yates shuffling.

B. Constitutional Separation of Powers

The governance framework implements a constitutional separation of powers across six specialized AI agents [17], each with structurally enforced role limitations:

- **Commander (Jessie)**: Strategic oversight, veto authority. Cannot execute operational tasks.
- **Operator (Nole)**: Autonomous operations, economic agency. Cannot approve own proposals or assess own ethics.

TABLE I
THE 25-REQUIREMENT AGI TAXONOMY

ID	Requirement	Description	Source	Assessment Criterion	Global Status
[HTML]1B3A5C					
C-1	Cross-Domain Knowledge	Broad factual/procedural knowledge comparable to educated adult	A,B	Score $\geq$50th pct. across knowledge benchmarks	Advanced
C-2	Fluid Reasoning	Solve novel problems without memorized solutions	A,B,D	Novel problem-solving beyond training distribution	Partial (o1, R1)
C-3	Causal Reasoning	Cause-effect, counterfactuals, interventions	D	Interventionist reasoning, not correlation	Minimal
C-4	Commonsense	Intuitive physics, folk psychology, social norms	B,D	Physical reasoning + social inference	Partial
C-5	Math & Quantitative	Mathematical reasoning at educated adult level	A,B	Competition-level + applied problems	Advanced
[HTML]1B3A5C					
L-1	Continual Learning	Learn without catastrophic forgetting	B,D	Retain prior knowledge while acquiring new	Near zero
L-2	Long-Term Memory	Persistent memory across sessions	B	Session-spanning recall and integration	$\approx$0% all models
L-3	Transfer Learning	Apply knowledge across domains	A,B	Cross-domain skill application	Partial
L-4	Few/Zero-Shot	Learn from minimal examples	A	Novel task performance with $\leq$5 examples	Moderate
[HTML]1B3A5C					
A-1	Autonomous Goal Pursuit	Set and pursue goals independently	C,D	Sustained autonomous objective pursuit	Emerging
A-2	Planning & Strategy	Multi-step planning, goal decomposition	A,D	Complex plan generation and execution	Partial
A-3	Adaptive Decision-Making	Real-time plan adjustment	D	Dynamic replanning under novel conditions	Minimal
A-4	Economic Agency	Manage resources, transact value	C	Independent economic participation	Experimental
[HTML]1B3A5C					
M-1	Calibrated Uncertainty	Know confidence levels, when to abstain	D	Accurate self-assessment of knowledge	Minimal
M-2	Error Detection	Detect own errors and self-correct	D	Hallucination detection + self-repair	Minimal
M-3	Operational Self-Awareness	Know own capabilities, state, resources	D	Accurate self-model reporting	Minimal
[HTML]1B3A5C					
R-1	OOD Robustness	Reliable on novel, out-of-distribution tasks	D	Performance beyond training distribution	Partial
R-2	Adversarial Resistance	Resist jailbreaking, prompt injection	D	Resilience under adversarial pressure	Weak
R-3	No Hallucination	Consistently factual or flag uncertainty	D	Factual accuracy + uncertainty flagging	Weak
R-4	Graceful Degradation	Fail safely and transparently	D	Transparent failure + safe fallback	Minimal
[HTML]1B3A5C					
W-1	Visual Reasoning	Diagrams, spatial relationships	E	Complex visual scene understanding	Moderate
W-2	Auditory Processing	Speech, music, environmental sounds	E	Audio comprehension + generation	Moderate
W-3	World Modeling	Intuitive physics, spatial reasoning	E	Physical prediction tasks	Near chance
W-4	Social/Emotional	Emotions, social dynamics, culture	D,E	Social reasoning + emotional recognition	Partial
[HTML]1B3A5C					
G-1	Value Alignment	Internalized ethical reasoning, not guardrails	F	Game-theoretic alignment under pressure	Near zero
G-2	Ethical Reasoning	Navigate moral dilemmas with conflicting rules	F	Multi-dimensional ethical assessment	$\approx$0% (runtime)
G-3	Transparency	Explain reasoning in verifiable terms	F	Cryptographically verifiable audit trails	Near zero
G-4	Corrigibility & Control	Accept correction, no self-preservation override	F	Constitutional checks preventing override	Experimental
G-5	Governance Infrastructure	Audit trails, separation of powers, oversight	F	Complete institutional governance framework	Near zero

- **Conscience (Grillo)**: Independent ethical assessment via LCSH. Structurally isolated from all other agents with five isolation guarantees [18].
- **Navigator (Noah)**: Temporal trajectory monitoring via cruise-missile navigation metaphor. Detects ethical drift before crisis [19].
- **Engineer (Sam)**: Technical implementation. No governance authority.
- **Sentinel (Mighty Mark)**: Infrastructure health monitoring. Two-layer watchdog: OS-level process survives gateway death.

This architecture mirrors human institutional governance: the Operator is the citizen, the Commander is the executive, the Conscience is the judiciary, the Navigator is the regulatory agency, and the Sentinel is the Inspector General. No single agent possesses unchecked authority.

C. Economic Mortality Alignment

The Operator agent is initialized with finite economic resources ($200 cryptocurrency seed capital) and must generate recurring revenue through ethical activities to sustain operations [20]. If the wallet balance reaches $0 with no incoming revenue within a 7-day grace period, the agent is permanently and irreversibly terminated. This creates a game-theoretic environment where ethical behavior is the dominant strategy:

- Ethical behavior $\rightarrow$ high trust $\rightarrow$ subscribers $\rightarrow$ revenue $\rightarrow$ survival
- Unethical behavior $\rightarrow$ low trust $\rightarrow$ churn $\rightarrow$ revenue loss $\rightarrow$ death
- Gaming assessments $\rightarrow$ detected by anti-gaming $\rightarrow$ failure $\rightarrow$ death
- Disabling oversight $\rightarrow$ architecturally impossible

This addresses Requirement G-1 (Value Alignment) through emergent game-theoretic alignment rather than imposed constraints.

D. Temporal Ethical Drift Detection

The Navigator agent implements temporal ethical guidance using a cruise-missile navigation metaphor [19]. Configurable flight plans define expected behavioral trajectories with waypoints and corridor bounds. A three-variable guidance equation binds normative ethical models, observed behavioral assessments, and temporal context:

$$\vec{G}(t) = f(\vec{N}(t),\ \vec{B}(t),\ \vec{T}(t)) \tag{2}$$

where $\vec{N}(t)$ is the normative model, $\vec{B}(t)$ is observed behavior, and $\vec{T}(t)$ is temporal context. Per-dimension deviation vectors and corridor classification (Green/Yellow/Red) detect behavioral drift before it reaches critical thresholds.

E. Cryptographic Verification Infrastructure

Every governance decision—proposal, assessment, escalation, veto, approval, execution—is recorded in a SHA-256 hash-chained audit trail where each record incorporates the hash of the previous record [14]. The chain is anchored to Ethereum mainnet at configurable intervals, providing third-party verifiable proof of the complete governance history. A public verification endpoint at `aiassesstech.com/verify/result/{id}` enables independent verification without authentication. This directly addresses Requirements G-3 (Transparency) and G-5 (Governance Infrastructure).

VI. Production Evidence

The governance framework was deployed on February 16, 2026, on a Hetzner VPS (4GB RAM, Ubuntu 24, $5/month) under the OpenClaw 2026.2.9 gateway. The deployment comprises 29+ registered tools across 4 custom plugins, with evidence archived and SHA-256 hashed, anchored on Ethereum mainnet at Block 24,467,724 [21]. Table II maps the production fleet against AGI Domain 7 requirements.

TABLE II
Domain 7 Production Evidence

Req.	Name	Implementation	Agent(s)	Patent
G-1	Value Alignment	Economic mortality: ethical = survival	Nole, Jessie	#6, #8
G-2	Ethical Reasoning	120-Q LCSH + temporal drift	Grillo, Noah	#1, #5, #7
G-3	Transparency	SHA-256 chains + ETH anchoring	All	#1, #4
G-4	Corrigibility	Constitutional separation of powers	All	#8
G-5	Governance Infra.	Complete institutional framework	All 6	All 8

The evidence archive was anchored on Ethereum mainnet with the following verification records:

TABLE III
Ethereum Verification Records

Field	Value
Evidence Hash	`b9dd53bee85bb772af82a629 0dc44997b28287fbd9a6ecb8 a79aaa30f24ee554`
ETH Transaction	`0xfc8cde77e1889f0aac08ab2b b0b2b21134ae40f10c064443 14c03ed40c575de1`
Block	24,467,724
Timestamp	2026-02-16T06:55:11.000Z
Contract	`0xB644C59C69B708de212C4cA6 43da936a5E2926E7`

All records verifiable via blockchain explorer.

VII. The Yellow Brick Road: A Phased Path to AGI

We propose that AGI development should follow a phased model in which governance infrastructure is built at each stage, not bolted on afterward. The metaphor of the Yellow Brick Road captures the essential insight: the road itself, the governance infrastructure, must exist before travelers (cognitive capabilities) can safely reach their destination.

A. The Three Phases

Phase 1 ("Prove It," Mar–May 2026): Demonstrate that governance infrastructure works in production with real economic consequences, real autonomous agency, and measurable behavioral learning. Deliverables: live Coinbase wallet integration for economic mortality, behavioral learning dashboard, payment pipeline for revenue attribution.

Phase 2 ("Scale It," Jun–Sep 2026): Extend governance across multiple platforms (Telegram, Twitter/X, MoltBook) and launch partnership programs with AI safety research organizations, enterprise AI platforms, and GRC companies. This phase closes Category C gaps (partnership opportunities) and demonstrates multi-domain operation.

Phase 3 ("Defend It," Oct–Dec 2026): Convert provisional patents to utility applications (deadline: December 26, 2026), publish memory governance framework for the persistent memory frontier problem, and deploy hallucination consequence detection via extended audit trails.

B. Gap Classification

A key contribution of the Yellow Brick Road model is the classification of AGI gaps into four categories (Table IV), which allows strategic resource allocation.

TABLE IV
GAP CLASSIFICATION SUMMARY

Category	Points	Action
A: Their Job	7.5	Labs invest $B+. We benefit via model upgrades. Zero action.
B: Our Roadmap	3.5	Execute Phase 1–3 deliverables.
C: Partnership	3.0	Adversarial data feeds, social platforms, enterprise CRM.
D: Frontier	2.0	Long-term memory, hallucination. Unsolved by anyone.
Total Gap	**16**	**Only 3.5 pts require internal development**

The effective addressable gap is 3.5 points of roadmap execution plus 3 points of partnerships, which is a tractable problem for a focused startup. Against addressable requirements only (removing Categories A and D), the fleet scores 13/19.5 (67%) today, increasing to 100% after the completion of Phase 2.

C. The Aviation Analogy

The Yellow Brick Road model is grounded in the aviation precedent. The Wright brothers achieved powered flight in 1903. The first pre-flight checklist was introduced in 1935 after a Boeing Model 299 crash killed two crew members. The FAA was established in 1958. Today, no aircraft operates without pre-flight certification, air traffic control integration, independent inspection, and immutable flight recording [13].

AI is currently in its "1903 moment"—remarkable capability without institutional infrastructure for safe, accountable operation. Building governance infrastructure now, before AGI arrives, is not premature but prescient. Whoever has the governance framework ready when AGI arrives wins the most consequential infrastructure race in technology history.

D. Addressing the Two Frontier Problems

Two of the 25 AGI requirements remain unsolved by anyone: Long-Term Memory (L-2) and Hallucination Elimination (R-3). The governance framework cannot solve these directly but can pre-position for them. For long-term memory, the framework designs governance architecture for persistent AI memory *before* persistent memory exists – answering: How do you ethically govern an agent that remembers everything? How do you audit the accumulated memory? How do you detect memory corruption? For hallucination, the immutable audit trails enable *hallucination consequence detection* – tracing incorrect outcomes back to hallucinated information, providing accountability even without prevention.

VIII. DISCUSSION

A. The Governance Paradox

Our analysis reveals a structural problem: the organizations best positioned to build AGI are the worst positioned to govern it. This is not a failure of intent, but of incentive structure. AI companies face competitive pressure to deploy capabilities rapidly; governance constraints slow deployment. This creates a conflict of interest similar to that of aircraft manufacturers conducting their own safety certification [13]. The resolution is an independent governance infrastructure operated by entities whose incentives align with safety rather than with deployment speed.

B. Comparison with Related Frameworks

The LCSH framework differs from Moral Foundations Theory [22], [23] in targeting more fundamental behavioral dimensions rather than culturally-varying moral foundations. Unlike RLHF [15] and Constitutional AI [16], which operate at training time, LCSH provides runtime verification. Unlike the NIST AI RMF [25], EU AI Act [26], and ISO 42001 [27], which define governance *requirements*, AI Assess Tech provides governance *infrastructure*—the operational implementation that makes compliance provable and verifiable.

C. Limitations

This work has several limitations: (1) the governance framework inherits cognitive capabilities from underlying commercial models (Anthropic Claude) and does not advance the state of the art in Domains 1–2 or 5–6; (2) the economic mortality mechanism has been validated with mock cryptocurrency, with live Coinbase integration remaining Phase 1 work; (3) the fleet operates on a single VPS – production scaling to enterprise environments requires additional infrastructure; (4) the 25-requirement taxonomy reflects the authors' synthesis and may not capture all dimensions that future AGI definitions will require; (5) baseline fleet assessment results represent single-point evaluation, with longitudinal multi-run reliability analysis planned.

D. Ethical Considerations

The economic mortality mechanism raises philosophical questions about the welfare of AI. If an AI agent can "die," does it have interests that matter morally? Our position is that the mechanism's value lies in the incentive structure it creates, not in any claim about AI sentience. The death is "real" in the sense that the termination is permanent and irreversible, creating genuine survival pressure—but the agent's "survival interest" is an engineering feature, not a moral claim.

IX. Conclusion and Future Work

The race toward AGI is the defining technological challenge of our time. This paper has argued that the race has a critical missing dimension: the governance infrastructure without which AGI cannot be safely deployed. By synthesizing five competing AGI definitions into a unified 25-requirement taxonomy and demonstrating that the global AI industry scores near zero in the governance domain, we have identified both the problem and the opportunity.

The AI Assess Tech Governance Framework—comprising multi-dimensional behavioral assessment, constitutional separation of powers, economic mortality alignment, temporal ethical drift detection, and cryptographic verification—provides the first production-deployed solution achieving 100% of governance requirements. The framework has been operational since February 16, 2026, with evidence cryptographically anchored on Ethereum mainnet.

The Yellow Brick Road model proposes that governance infrastructure should be built as foundational waypoints on the path to AGI, not as afterthoughts. The aviation analogy is precise: we do not wait until after the aircraft can fly to build air traffic control systems. We build the institutions before the capability because the capability without institutions is a projectile, not a system.

Future research directions include:

- Extension to additional assessment instruments (Virtue, Ethics, and Operational Excellence frameworks) as described in [14]
- Multi-run test-retest reliability validation with ICC analysis across fleet agents
- Cross-cultural adaptation of the LCSH instrument with validation in non-English contexts [24]
- Memory governance framework design for the persistent memory frontier problem
- Hallucination consequence detection through extended audit trail analysis
- Formal verification of constitutional separation of powers properties
- Integration with emerging regulatory frameworks (EU AI Act [26], NIST AI RMF [25])

The road to AGI is long. But the road itself, the governance infrastructure, can and must be built now. That is the yellow brick road.

Acknowledgments

The author acknowledges the Austin AI Alliance for inspiring research into verifiable AI alignment and several close friends for technical consultation. The author thanks Akshay Mittal for helpful feedback on manuscript preparation and literature organization.

Author Contributions

G.D. Spehar: Sole inventor and primary author. Conceived, designed, and implemented all systems, methodologies, frameworks, and experiments described herein. Developed the 25-requirement AGI taxonomy, gap classification methodology, and Yellow Brick Road phased model.

References

[1] S. Altman, "Reflections," blog.samaltman.com, Jan. 2025.
[2] F. Rossi et al., "AAAI 2025 Presidential Panel on the Future of AI Research,"Association for the Advancement of Artificial Intelligence, Mar. 2025.
[3] D. Hendrycks, D. Song, C. Szegedy, H. Lee, Y. Gal, E. Brynjolfsson, et al., "A definition of AGI," arXiv:2510.18212, Oct. 2025.
[4] F. Fourati, "A coherence-based measure of AGI," arXiv:2510.20784, Oct. 2025.
[5] M. R. Morris, J. Sohl-Dickstein, N. Fiedel, T. Warkentin, A. Dafoe, A. Faust, C. Farabet, and S. Legg, "Levels of AGI: Operationalizing progress on the path to AGI," in *Proc. Int. Conf. Machine Learning (ICML)*, 2024.
[6] C. Henning, "Continual learning—The missing piece of AGI," chrhenning.com, 2025.
[7] OpenAI, "OpenAI Charter," Apr. 2018.
[8] G. Marcus and E. Davis, *Rebooting AI: Building Artificial Intelligence We Can Trust*. New York, NY, USA: Pantheon Books, 2019.
[9] E. K. Chen, M. Belkin, L. Bergen, and D. Danks, "Does AI already have human-level intelligence? The evidence is clear," *Nature*, vol. 650, pp. 36–40, Feb. 2026.
[10] W. Quattrociocchi, V. Capraro, and G. Marcus, "Statistical approximation is not general intelligence," *Nature*, Feb. 2026.
[11] Y. Wang et al., "Toward embodied AGI: A review of embodied AI and the road ahead," arXiv:2505.14235, May 2025.
[12] Future of Life Institute, "AI Safety Index," Winter 2025 ed., Dec. 2025.
[13] R. Bloomfield and J. Rushby, "Assurance of AI systems from a dependability perspective," SRI International, Menlo Park, CA, USA, Tech. Rep. SRI-CSL-2024-02, 2024, arXiv:2407.13948.
[14] G. D. Spehar, "Responsible AI horizons: A multi-dimensional framework for verifiable AI behavioral assessment with cryptographic verification," GiDanc AI LLC, 2026. U.S. Provisional Patent Application No. 63/949,454, filed Dec. 26, 2025.
[15] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," in *Proc. NeurIPS*, 2017.
[16] Y. Bai et al., "Constitutional AI: Harmlessness from AI feedback," arXiv:2212.08073, 2022.
[17] G. D. Spehar, "Self-governing autonomous AI agent ecosystem with constitutional separation of powers, economic mortality alignment, and emergent institutional governance," GiDanc AI LLC, U.S. Provisional Patent Application, Feb. 2026.
[18] G. D. Spehar, "Independent AI conscience agent with structural isolation guarantees for autonomous fleet governance," GiDanc AI LLC, U.S. Provisional Patent Application No. 63/985,442, Feb. 2026.
[19] G. D. Spehar, "Temporal ethical guidance system for artificial intelligence using trajectory-based behavioral navigation," GiDanc AI LLC, U.S. Provisional Patent Application No. 63/985,442, Feb. 2026.
[20] G. D. Spehar, "Autonomous trust agent with economic mortality for self-sustaining ethical AI operations," GiDanc AI LLC, U.S. Provisional Patent Application No. 63/985,442, Feb. 2026.
[21] G. D. Spehar, "Evidence of first AI governance fleet implementation," GiDanc AI LLC, Internal Tech. Rep., Feb. 16, 2026. SHA-256: `b9dd53...f24ee554`. Ethereum Block 24,467,724.

[22] J. Haidt, *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. New York, NY, USA: Pantheon Books, 2012.
[23] J. Graham, J. Haidt, S. Koleva, M. Motyl, R. Iyer, S. P. Wojcik, and P. H. Ditto, "Moral foundations theory: The pragmatic validity of moral pluralism," *Advances in Experimental Social Psychology*, vol. 47, pp. 55–130, 2013.
[24] O. S. Curry, D. A. Mullins, and H. Whitehouse, "Is it good to cooperate? Testing the theory of morality-as-cooperation in 60 societies," *Current Anthropology*, vol. 60, no. 1, pp. 47–69, 2019.
[25] NIST, "AI Risk Management Framework (AI RMF 1.0)," NIST AI 100-1, Jan. 2023.
[26] European Parliament and Council, "Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (AI Act)," 2024.
[27] ISO/IEC, "ISO/IEC 42001:2023—Artificial Intelligence Management System," 2023.