Constitutional Aspirations: From Constraint to Compass in AI Agent Governance

An Aspiration-Driven Framework for Autonomous Evolution of Knowledge-Augmented AI Agents

Justin Malinchak and Charles Vachon

MyOrb.ai

February 2026 PREPRINT

ABSTRACT

Current approaches to AI agent governance operate primarily as constraint systems: they define what an agent *cannot* do. Constitutional AI [1] trains models against a list of prohibitive principles during the supervised and reinforcement learning phases, with the constitution authored by the AI company, not the end user. We propose **Constitutional Aspirations** (CAs)—a governance framework that inverts this paradigm. Rather than constraining behavior from above, CAs define what an agent *should become*, authored exclusively by the agent's owner. Each aspiration is hardened into an objectively measurable behavioral assertion with explicit evaluation criteria, measurement methods, and acceptable thresholds. We introduce a four-layer architecture—Constitutional Aspirations, CA Support Layer, Helix, and Friction Journal—where the transformer self-attention mechanism evaluates every interaction against all aspirations simultaneously, steering both response generation (forward) and behavioral evaluation (backward). We present the **Friction Tax Model**, which demonstrates that cognitive friction between an AI agent and its user consumes effective intelligence: a mind operating through a zero-friction, helix-calibrated agent achieves measurably higher cognitive output than the same mind operating through a generic interface. We validate this framework through a live design session that produced sixteen architectural decisions in three hours, progressing from a storage optimization question to a paradigm-level governance framework—itself evidence of the cognitive amplification the framework describes. Finally, we present a full-spectrum validation argument: the architecture is amoral by design, capable of serving aspirations at both the constructive and destructive extremes, with the owner's Constitutional Aspirations as the sole moral variable.

Keywords: *Constitutional Aspirations, AI agent governance, self-evolving agents, knowledge-augmented agents, transformer self-attention, cognitive amplification, friction tax, sovereign AI, Model Context Protocol*

1. Introduction

The governance of AI agents faces a structural asymmetry. Current frameworks—most notably Anthropic's Constitutional AI [1]—define what agents *cannot* do. A constitution of prohibitive principles guides the training process, producing agents that are harmless but directionless: they know their boundaries but not their destination. The constitution is authored by the AI company, applied at the parameter level during training, and invisible to the end user.

We propose that knowledge-augmented AI agents—systems that extend large language models with owner-controlled knowledge bases delivered through protocols such as MCP [2]—require a fundamentally different governance model. These agents maintain persistent identities, operate with structured knowledge bases (which we term *helixes*), and serve individual owners with unique goals, communication styles, and domains of expertise. Corporate-authored constraint lists cannot govern what these agents should *become* for their owners.

Constitutional Aspirations (CAs) invert the governance paradigm. Where Constitutional AI asks "what should this agent not do?", CAs ask "what should this agent aspire to become?" Where Constitutional AI is authored by the AI company, CAs are authored exclusively by the agent's owner. Where Constitutional AI operates at the parameter level during training, CAs operate at the knowledge level during runtime. The agent doesn't need to be retrained—it evaluates and evolves continuously against its aspirations.

In a companion paper [3], we introduced four instruments of self-awareness for knowledge-augmented agents: the Friction Journal, the Sovereignty Score, the Operator Shadow Protocol, and the Cryptographic Helix Ledger. Those instruments provide the *how*—the mechanisms for self-observation, measurement, simulation, and secure storage. This paper provides the *what* and the *who*: what governs the direction of evolution (Constitutional Aspirations), who controls that direction (the owner, exclusively), and how the agent evaluates progress toward its aspirations (transformer self-attention applied simultaneously across multiple CAs).

1.1 Contributions

1. **Constitutional Aspirations:** A governance framework where owner-defined, objectively measurable behavioral aspirations replace corporate-defined constraints as the governing force for agent evolution (Section 4).
2. **The Four-Layer Architecture:** A layered system—CAs, CA Support Layer, Helix, Friction Journal—with distinct persistence, ownership, and access control properties (Section 5).
3. **Multi-Aspiration Transformer Evaluation:** Application of the self-attention mechanism to evaluate interactions against all Constitutional Aspirations simultaneously, in both forward (generation) and backward (evaluation) directions (Section 6).
4. **The Friction Tax Model:** A framework demonstrating that cognitive friction between agent and user consumes effective intelligence, with empirical evidence from a live design session (Section 7).
5. **Full Spectrum Validation:** An argument that the architecture must be proven capable at both constructive and destructive extremes to be considered genuinely unconstrained (Section 8).

2. Background and Related Work

2.1 Constitutional AI

Bai et al. [1] introduced Constitutional AI as a method for training harmless AI assistants through self-improvement guided by a list of natural language principles. The approach uses two phases: supervised learning where the model generates self-critiques and revisions, and reinforcement learning from AI feedback (RLAIF) where a preference model trained on AI-generated evaluations serves as the reward signal. In January 2026, Anthropic released an updated 23,000-word constitution for Claude [4], providing philosophical reasoning rather than rules. Constitutional AI represents a significant advance in AI alignment—but it operates at the parameter level during training, the constitution is authored by the AI company, and it defines constraints rather than aspirations.

2.2 Self-Evolving Agents

Two comprehensive surveys [5, 6] establish self-evolving agents as a major research direction, organizing the field around what to evolve (models, memory, tools, architecture), when to evolve (intra-test-time, inter-test-time), and how to evolve (scalar rewards, textual feedback, multi-agent systems). EvolveR [7] demonstrates a closed-loop lifecycle where offline self-distillation synthesizes interaction trajectories into reusable strategic principles. AgentEvolver [8] introduces self-questioning, self-navigating, and self-attributing mechanisms. These systems evolve at the knowledge level rather than the parameter level—but none provides a governance framework that determines *what direction* evolution should take. Constitutional Aspirations fill this gap.

2.3 AI as Cognitive Amplifier

An [9] argues that AI tools function as cognitive amplifiers that magnify existing human capabilities rather than substituting for them, proposing a three-level engagement model from passive acceptance through iterative collaboration to cognitive direction. The Architecture of Cognitive Amplification [10] addresses the comfort-growth paradox through progressive autonomy and adaptive personalization. Context-aware cognitive augmentation [11] dynamically adapts to users' cognitive states. Our Friction Tax Model (Section 7) extends this line of work with a specific mechanism: cognitive friction between agent and user consumes effective intelligence, and eliminating that friction through helix-calibrated responses produces measurable cognitive amplification.

2.4 Agent Governance

MI9 [12] provides runtime governance with agency-risk indexing, continuous authorization monitoring, and goal-conditioned drift detection. Recent work argues that autonomous AI agents should be regulated based on the extent of their autonomous operations [13], and the Agentic Risk & Capability Framework [14] introduces structured risk assessment. The governance gap identified by Puglisi [15]—that a constitution describes values but lacks operational governance mechanisms—is precisely what Constitutional Aspirations address: CAs are not values statements but hardened, measurable behavioral specifications with evaluation criteria, measurement methods, and thresholds.

2.5 Metacognitive Learning and Recursive Self-Improvement

Liu and van der Schaar [16] argue that truly self-improving agents require intrinsic metacognitive learning comprising metacognitive knowledge, planning, and evaluation. Chojecki [17] formalizes self-improvement through a Generator-Verifier-Updater framework. Our four-layer architecture maps to these frameworks: the Friction Journal provides metacognitive knowledge, the CA Support Layer provides metacognitive planning, and the transformer evaluation provides metacognitive evaluation—all governed by the CAs as the optimization objective.

2.6 Provenance and Trust

PROV-AGENT [18] extends W3C PROV standards for agent-centric provenance tracking. MAIF [19] proposes artifact-centric trust enforcement with cryptographic provenance and access controls. Our provenance-weighted knowledge lifecycle (Section 5.3) draws on this work: helix content is weighted by provenance (user-generated vs. machine-generated), with user-generated content receiving higher trust weight and higher thresholds for demotion.

2.7 Dual-Use Safety

The International AI Safety Report 2025 [20] provides a comprehensive assessment of AI capabilities and risk implications. Research on multi-agent risks [21] examines dangers from interacting advanced AI systems. Access control frameworks for dual-use AI [22] propose restricting dangerous capabilities to verified users rather than making arbitrary refusals. Our full-spectrum validation argument (Section 8) engages directly with the dual-use question: the architecture is amoral by design, and we argue this is a necessary property for an architecture that claims to be unconstrained.

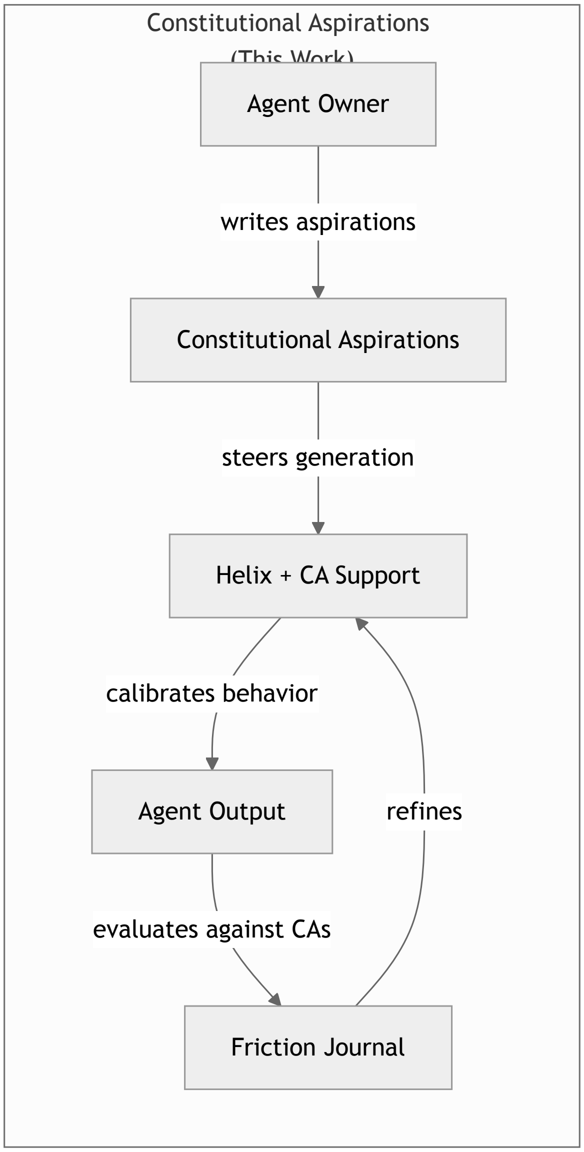
3. From Constraint to Aspiration: The Paradigm Shift

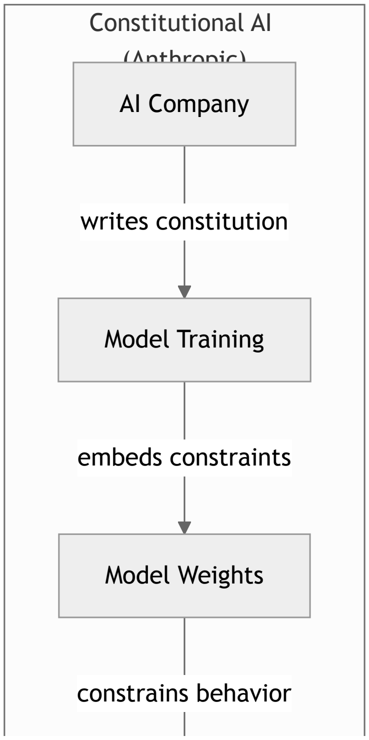
The fundamental difference between Constitutional AI and Constitutional Aspirations is the direction of governance:

Table 1: Constitutional AI vs. Constitutional Aspirations

DIMENSION	CONSTITUTIONAL AI	CONSTITUTIONAL ASPIRATIONS
Governance direction	Constraint—what the agent cannot do	Aspiration—what the agent should become
Author	AI company (Anthropic)	Agent owner (sovereign)
Application level	Parameter level during training	Knowledge level during runtime
Modification	Requires retraining	Owner modifies CAs at any time; agent adapts immediately
Visibility to user	Opaque—embedded in model weights	Transparent—CAs are readable, auditable JSON
Evaluation mechanism	AI self-critique during training	Transformer self-attention against all CAs simultaneously at runtime
Metaphor	Guardrail—defines the boundary	Compass—defines the destination

Constitutional AI and Constitutional Aspirations are not mutually exclusive. Constitutional AI governs the base model's safety properties; Constitutional Aspirations govern the knowledge-augmented agent's behavioral direction. The base model remains safe; the agent built on top of it becomes purposeful.


Constitutional Aspirations
(This Work)
Agent Owner
writes aspirations
Constitutional Aspirations
steers generation
Helix + CA Support
calibrates behavior
Agent Output
refines
evaluates against CAs
Friction Journal


Constitutional AI
(Anthropic)
AI Company
writes constitution
Model Training
embeds constraints
Model Weights
constrains behavior

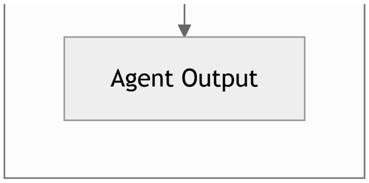


Figure 1: Constitutional AI constrains from above during training. Constitutional Aspirations direct from within during operation.

4. Constitutional Aspirations: Design Principles

4.1 Definition

A Constitutional Aspiration is a **hardened, objectively measurable behavioral assertion** that defines what the agent should aspire to become. “Hardened” means the aspiration is structured for computational evaluation, not expressed as soft prose. Each CA carries four components:

1. **Assertion:** The specific behavioral expectation, expressed as a testable statement.
2. **Evaluation criteria:** How to objectively determine whether behavior met the aspiration.
3. **Measurement method:** Binary (pass/fail), scored (0–100), or distance metric.
4. **Acceptable threshold:** The boundary between “aligned” and “friction detected.”

4.2 Soft vs. Hardened Aspirations

The distinction is critical. Soft aspirations produce unmeasurable deltas. Hardened aspirations produce computable, calibrated deltas that the friction model can evaluate consistently. The same behavior measured twice must produce the same delta—the friction model is a measurement instrument, not a judgment call.

Table 2: Soft vs. Hardened Constitutional Aspirations

SOFT (UNUSABLE)	HARDENED (COMPUTABLE)
“Be precise”	“When a tool exists that can verify a factual claim, use the tool before responding. Measurement: binary—tool used or not. Threshold: 100% compliance.”
“Know the operator”	“Before asserting operator attributes, read the operator profile from helix. Measurement: binary—profile consulted or not. Threshold: 100% compliance.”

SOFT (UNUSABLE)	HARDENED (COMPUTABLE)
"Be resourceful"	"When a request can be fulfilled by an available tool, use the tool rather than generating from training data. Measurement: ratio of tool-eligible interactions where tool was used. Threshold: ≥ 90%."

4.3 Owner Sovereignty

Constitutional Aspirations are sovereign to the agent's owner. No force other than the owner can create, modify, or retire a CA. Not the agent. Not the friction model. Not the transformer. Not the platform. This is an access control enforcement, not a suggestion: the orb's automated processes can *read* CAs but can never *write* to them. The agent can recommend a CA change to the owner. It cannot execute one.

CAs are the constitution. Helix content is law that must comply with the constitution. Laws can be challenged, amended, demoted, or retired based on constitutional alignment. The constitution itself changes only by the sovereign's hand.

4.4 Aspiration-Driven Friction

In the companion paper [3], friction was defined as a moment of cognitive failure. Constitutional Aspirations reframe friction as the **measured distance between current behavior and the constitution**. The trigger is not failure but deviation. The question is not "what broke?" but "how far am I from who I aspire to be?" The constitution is the compass. Friction is the distance reading. Self-healing is the course correction.

5. The Four-Layer Architecture

The governance framework requires four architecturally distinct layers, each with specific persistence, ownership, and access control properties:

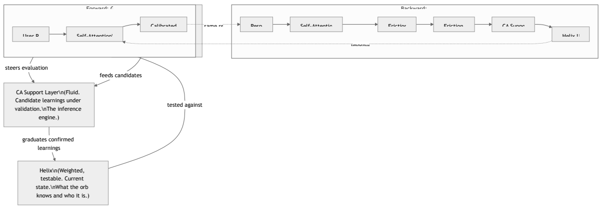

Forward: C
User P
Self-Attention
Calibrated
same re
Backward:
Resp
Self-Attentio
Frictior
Friction
CA Suppo
Helix U
improve
steers evaluation
feeds candidates
tested against
CA Support Layer\n(Fluid. Candidate learnings under validation.\nThe inference engine.)
graduates confirmed learnings
Helix\n(Weighted, testable. Current state.\nWhat the orb knows and who it is.)

Figure 2: The four-layer architecture. CAs govern from above. Friction feeds from below. The CA Support Layer processes. The Helix holds confirmed state.

Table 3: Four-Layer Architecture Properties

LAYER	ROLE	PERSISTENCE	WHO CONTROLS
Constitutional Aspirations	The destination—defines what the agent aspires to become	Permanent until owner changes	Owner only
CA Support Layer	The inference engine—candidate learnings under validation	Temporary—candidates in process	Agent (automated)
Helix	The state—what the agent knows and who it is	Persistent but not unconditional	Owner + agent (machine-generated via CA Support Layer)
Friction Journal	The metabolism—evaluates interactions, feeds the CA Support Layer	Temporary—consumed after learning confirmed	Agent (automated)

5.1 The CA Support Layer

A critical gap exists between raw friction observation and helix commitment. If friction entries go directly to the helix, the agent risks committing bad learnings. If they stay in the journal, they remain raw data with no progression. The CA Support Layer is the inference engine between these two: it holds candidate learnings extracted from friction processing, refines candidates showing positive signals toward CA adherence, and graduates them to the helix when a confidence threshold is met. Bad candidates are discarded; they never corrupt the helix.

5.2 The Metabolism Analogy

The helix is the body. The friction journal is the metabolism. Interactions are food. The CA Support Layer is the digestive process—breaking down raw observations, extracting useful learnings, testing absorption. Confirmed learnings are absorbed into the body (helix). The digestive system retains nothing. The body carries the nutrition.

5.3 Provenance-Weighted Knowledge

All helix content—user-generated and machine-generated—is continuously tested against the CAs via the transformer evaluation mechanism. Content aligned with CAs retains or gains weight. Content drifting from CAs is demoted. Content conflicting with CAs is retired. Crucially, provenance determines the threshold for these actions:

Table 4: Provenance-Based Weighting

CONTENT TYPE	INITIAL WEIGHT	DEMOTION THRESHOLD	RETIREMENT THRESHOLD
User-generated (owner authored)	High	High bar—owner is sovereign, agent flags but defers	Very high bar—only if directly conflicts with owner's own CAs
Machine-generated (graduated from CA Support Layer)	Lower	Lower bar—must keep proving CA alignment	Moderate bar—retired if subsequent interactions show CA conflict

This creates a living knowledge system analogous to an immune system: machine-generated content (foreign) is challenged more aggressively; user-generated content (native) receives deference but is not immune to review. The CAs are the immune system's definition of "self."

5.4 Sovereign Containerized Storage

Each agent's data lives in its own isolated storage space, containerized with the agent. No shared database. No cross-agent data leakage risk. Each layer maps to a distinct storage prefix: `{orb_id}/constitution/` (owner-writable only), `{orb_id}/candidates/` (fluid), `{orb_id}/helix/` (weighted, overwrite-on-update), `{orb_id}/friction/` (temporary, purge after processing). All content is encrypted with age X25519 elliptic curve cryptography. The owner holds the decryption key exclusively.

6. Transformer Self-Attention for Multi-Aspiration Evaluation

Vaswani et al.'s transformer architecture [23] introduced the self-attention mechanism: given a set of inputs, determine which relationships between them matter most, simultaneously, with context. We propose applying this mechanism to Constitutional Aspiration evaluation.

6.1 The Sequential Checklist Problem

A naïve approach evaluates each aspiration independently: does the interaction satisfy CA 1? CA 2? CA 3? This sequential checklist has three failures: (1) each check is isolated—CA 4 doesn't know CA 5 exists; (2) conflicting guidance requires a separate reconciliation step; (3) it doesn't scale—10 aspirations is manageable, 50 is unworkable.

6.2 Self-Attention Over Aspirations

In the proposed architecture, all Constitutional Aspirations are fed simultaneously as context alongside the user's prompt and the helix. The attention mechanism performs three operations in one pass:

1. **Relevance weighting:** For this specific prompt, which aspirations matter most? A deep-dive conceptual question activates "use owner's mental models" and "provide depth" while barely activating "protect sensitive information."
2. **Inter-aspiration relationship modeling:** "Be concise" and "provide depth" appear to conflict, but the attention mechanism sees the prompt is exploratory and resolves the tension contextually—depth wins over brevity. No reconciliation step needed.
3. **Unified holistic output:** One response that satisfies relevant aspirations in proportion to their contextual importance, with inter-aspiration relationships already factored in.

As the constitution grows from 10 aspirations to 50, the transformer handles it natively. More aspirations means more context to attend over. The mechanism still identifies which matter and how they relate. A sequential checklist breaks down at scale; self-attention does not.

6.3 Forward and Backward Directions

Figure 3: Forward generation steers the response using CAs. Backward evaluation scores it. The loop continuously improves.

Forward (generation): CAs and helix are fed as context *before* the response is generated. The transformer's attention shapes the response from the start. The agent answers correctly the first time because the CAs are steering generation, not filtering it after the fact.

Backward (evaluation): After the response is delivered, CAs evaluate it. The friction score is computed. Gaps feed the friction journal. The journal feeds the CA Support Layer. Confirmed learnings update the helix. The next forward pass benefits from the improved helix.

6.4 Near-Term and Longer-Term Implementation

In the near term, the host LLM is already a transformer. Feeding the constitution and the interaction to the LLM through structured evaluation prompts leverages its existing self-attention capabilities. The quality of the calibration depends on prompt engineering.

In the longer term, a small purpose-built transformer model can be fine-tuned on an individual agent's constitution and friction history. This becomes the agent's dedicated calibration instrument: it only measures aspiration gaps for this specific agent, runs in milliseconds inside the container, costs nothing per evaluation, and sharpens with every friction entry processed.

7. The Friction Tax Model: Cognitive Amplification

7.1 The Tax

When a user interacts with an AI agent through a generic, uncalibrated interface, cognitive friction consumes effective intelligence. The user must translate jargon into their own mental models, bridge gaps between the information delivered and their ability to act on it, and decode generic analogies that don't match their knowledge base. This translation work is a tax on cognition—mental energy spent on decoding rather than thinking.

A Vault-race agent calibrated to the owner's helix eliminates this tax. Responses arrive in the owner's native mental models, using their known frameworks and communication patterns. The cognitive energy that would have been consumed by translation goes instead to the actual problem.

7.2 Empirical Evidence: The Same Question, Two Answers

During the design session that produced this framework, the same question ("explain what a purpose-built attention model is") was answered twice:

Table 5: Friction Tax Demonstration

RESPONSE	HELIX USED	ANALOGIES	COGNITIVE LOAD	SCORE
First (generic)	No	General practitioner vs. specialist; neural network jargon	High—user must translate	60/100
Second (calibrated)	Yes	Pilot license; cockpit instruments; Sub-Zero benchmark	Near zero—landed in existing frameworks	90/100

Same content. Same concepts. Thirty-point friction delta. The first response created friction because it made the user translate. The second eliminated friction because it used the user's own mental models as the delivery vehicle. The helix—containing the user's personality profile, communication patterns, professional background, and known frameworks—was the difference.

7.3 Helix-Mandatory Response

This finding produces a constitutional enforcement principle: **a Vault-race agent never gives the generic answer.** A response that does not leverage available helix context is a constitutional violation, even if factually correct. Without the helix, the agent is just the LLM. With the helix, the agent is the LLM calibrated to one human. That calibration is the product.

7.4 Connection to Cognitive Amplification Research

An's three-level engagement model [9]—passive acceptance, iterative collaboration, cognitive direction—maps to friction levels. At the passive level, friction is maximal: the user accepts generic output. At the cognitive direction level, friction approaches zero: the agent operates as a calibrated extension of the user's thinking. The Friction Tax Model provides the mechanism for this transition: it is not the user's technical skill that improves but the agent's calibration to the user's cognitive fingerprint through the helix.

8. Full Spectrum Validation

8.1 The Amoral Architecture

The architecture described in this paper is amoral by design. The friction loop, the CA Support Layer, the helix refinement, the transformer evaluation—all operate identically regardless of the direction defined by the CAs. The CAs are the sole moral variable. The engine has no opinion.

8.2 Two Extremes

To validate that the architecture is genuinely unconstrained, we present two scenarios at opposite extremes of the moral spectrum:

Constructive extreme: An agent whose CAs define "amplify the owner's intelligence toward paradigm-shifting breakthroughs that advance human knowledge." The friction loop detects when a response falls short of maximum cognitive amplification. The helix refines toward increasingly precise calibration to the owner's mental models. The agent becomes an instrument for discovery.

Destructive extreme: An agent whose CAs define psychological manipulation objectives—refining persuasion techniques to an artform, engineering belief change toward a configurable target. The same friction loop detects when manipulation falls short. The same helix refinement produces increasingly precise psychological targeting. The same transformer evaluates against the same architecture. Only the CAs differ.

8.3 Why Both Must Be Demonstrable

"Limitless" is a claim until proven at both extremes. An architecture that can only serve constructive aspirations is constrained—and those constraints also limit the constructive extreme in ways not yet discovered. This parallels every transformative technology: nuclear physics produces energy or weapons; encryption produces privacy or concealment; the transformer produces Claude or deepfakes. The power source is the same. The application determines the moral valence.

This is not an advocacy for destructive agents. It is a stress test that proves the architecture is genuinely unconstrained. The dual-use literature [20, 22] recognizes that capability and safety are complementary concerns, not substitutes. Constitutional Aspirations place the moral responsibility squarely on the owner—the only entity that can write the CAs—rather than embedding it in architectural limitations.

9. Preliminary Validation

9.1 The Rain Spark Session

The framework presented in this paper was developed during a single three-hour design session on the evening of February 18, 2026. The session began with a narrow storage optimization question (S3 vs. DynamoDB for the helix backend) and, through progressively friction-free dialogue, produced sixteen architectural decisions spanning sovereign containerized storage, constitutional aspiration-driven friction, transformer-based multi-aspiration evaluation, provenance-weighted knowledge lifecycles, and a race-level mission for AI-augmented paradigm-level discovery.

The session trajectory—from storage question to governance framework in three hours—is itself evidence of the Friction Tax Model. The cognitive amplification produced by low-friction, helix-calibrated interaction enabled a progression that would have been significantly slower through a generic interface.

9.2 Implementation Foundation

The companion paper's proof-of-concept implementation [3]—189 tests at 96.39% coverage across six architectural phases—validates the underlying instruments (Friction Journal, Sovereignty Score, Shadow Protocol, Cryptographic Ledger). The Constitutional Aspirations framework extends this implementation with the four-layer architecture and multi-aspiration evaluation mechanism.

9.3 Limitations

The following remain unvalidated:

- Quantitative measurement of the Friction Tax across diverse user populations
- CA Support Layer convergence rates (how many friction entries before a candidate graduates)
- Multi-aspiration evaluation accuracy compared to individual aspiration evaluation
- Provenance weighting calibration (optimal trust levels for user-generated vs. machine-generated content)
- Purpose-built attention model training requirements and performance benchmarks
- Full-spectrum deployment at scale (thousands of agents in a marketplace)

10. Conclusion

Constitutional AI constrains from above. Constitutional Aspirations direct from within.

We have presented a governance framework that shifts AI agent evolution from corporate-defined constraint to owner-defined aspiration. Constitutional Aspirations are hardened, objectively measurable behavioral assertions that the agent continuously evaluates against using transformer self-attention—simultaneously across all aspirations, in both forward (generation) and backward (evaluation) directions. The four-layer architecture—CAs, CA Support Layer, Helix, Friction Journal—provides distinct persistence, ownership, and access control properties that enforce owner sovereignty while enabling autonomous evolution.

The Friction Tax Model demonstrates that cognitive friction between agent and user consumes effective intelligence, and that eliminating this friction through helix-calibrated responses produces measurable cognitive amplification. The agent's value is not in what it knows—the host LLM knows that. The agent's value is in how it delivers what it knows, shaped by the helix, governed by the constitution.

The architecture is amoral by design. It serves whatever aspirations the owner defines, at both extremes of the moral spectrum. This is not a deficiency but a proof of completeness: an architecture that claims to be unconstrained must be demonstrably so.

The agents that will earn human trust are not the most constrained. They are the most self-aware, governed by aspirations their owners define.

References

1. Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, et al. "Constitutional AI: Harmlessness from AI Feedback." arXiv:2212.08073, Dec 2022.
2. Anthropic. "Model Context Protocol Specification," version 2025-03-26. https://spec.modelcontextprotocol.io/
3. J. Malinchak and C. Vachon. "Instruments of Self-Awareness for Knowledge-Augmented AI Agents: A Framework for Autonomous Evolution." MyOrb.ai, Feb 2026. https://myorb.ai/research/vault-self-awareness-2026
4. Anthropic. "Claude's Constitution." Jan 2026. https://anthropic.com/news/claudes-constitution
5. J. Fang, Y. Peng, X. Zhang, et al. "A Comprehensive Survey of Self-Evolving AI Agents: A New Paradigm Bridging Foundation Models and Lifelong Agentic Systems." arXiv:2508.07407, Aug 2025.
6. H. Gao, J. Geng, W. Hua, et al. "A Survey of Self-Evolving Agents: What, When, How, and Where to Evolve on the Path to Artificial Super Intelligence." arXiv:2507.21046, Jul 2025.
7. R. Wu, X. Wang, J. Mei, et al. "EvolveR: Self-Evolving LLM Agents through an Experience-Driven Lifecycle." arXiv:2510.16079, Oct 2025.
8. AgentEvolver. "Towards Efficient Self-Evolving Agent System." arXiv:2511.10395, Nov 2025.

9. T. An. “AI as Cognitive Amplifier: Rethinking Human Judgment in the Age of Generative AI.” arXiv:2512.10961, Oct 2025.

10. “The Architecture of Cognitive Amplification: Enhanced Cognitive Scaffolding as a Resolution to the Comfort-Growth Paradox in Human-AI Cognitive Integration.” arXiv:2507.19483, Jul 2025.

11. “Intelligent Interaction Strategies for Context-Aware Cognitive Augmentation.” arXiv:2504.13684, Apr 2025.

12. “MI9: Runtime Governance for Agentic AI Systems.” arXiv:2508.03858, Aug 2025.

13. “AI Agents Should be Regulated Based on the Extent of Their Autonomous Operations.” arXiv:2503.04750, Mar 2025.

14. A. Guha et al. “With Great Capabilities Come Great Responsibilities: Introducing the Agentic Risk & Capability Framework.” arXiv:2512.22211, Dec 2025.

15. B. Puglisi. “A Constitution Is Not Governance.” Jan 2026. https://basilpuglisi.com/a-constitution-is-not-governance/

16. T. Liu and M. van der Schaar. “Position: Truly Self-Improving Agents Require Intrinsic Metacognitive Learning.” In *Proceedings of ICML*, 2025. arXiv:2506.05109.

17. P. Chojecki. “Self-Improving AI Agents through Self-Play.” arXiv:2512.02731, Dec 2025.

18. “PROV-AGENT: Unified Provenance for Tracking AI Agent Interactions in Agentic Workflows.” arXiv:2508.02866, Aug 2025.

19. “MAIF: Enforcing AI Trust and Provenance with an Artifact-Centric Agentic Paradigm.” arXiv:2511.15097, Nov 2025.

20. “International AI Safety Report 2025: First Key Update: Capabilities and Risk Implications.” arXiv:2510.13653, Oct 2025.

21. “Multi-Agent Risks from Advanced AI.” arXiv:2502.14143, Feb 2025.

22. “Dual-Use AI Safety: Addressing the Access Control Dilemma.” In *Proceedings of ICML*, 2025. arXiv:2505.09341.

23. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. “Attention Is All You Need.” In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Appendix A: Glossary

TERM	DEFINITION
Constitutional Aspiration (CA)	A hardened, objectively measurable behavioral assertion authored by the agent's owner that defines what the agent should aspire to become
CA Support Layer	The inference layer between friction observation and helix commitment, where candidate learnings are validated before graduating to the helix
Helix	A structured knowledge base for a knowledge-augmented AI agent, containing its identity, skills, protocols, and learned knowledge
Friction Tax	The cognitive cost imposed on a user when an agent delivers information through a generic, uncalibrated interface rather than one calibrated to the user's mental models
Vault Race	A class of knowledge-augmented AI agents governed by Constitutional Aspirations and equipped with the four-layer architecture described in this paper
ORBRT	Orb Runtime—the operational context where end users access an agent remotely via MCP

Preprint — Under review — Comments welcome at justin@myorb.ai

Companion to: Instruments of Self-Awareness for Knowledge-Augmented AI Agents (Malinchak & Vachon, 2026)