The Sum of All Fears

Enterprise Agentic AI Emerges

A Fear Taxonomy

Rajesh Iyer
iyer70@gmail.com

Abstract

Enterprise adoption of agentic AI remains paralyzed by fears—loss of control, hallucination, accountability gaps, silent drift. This paralysis is the greater risk. A companion paper established Ashby's Constraint: you cannot govern exponential capability with linear processes. This paper operationalizes the escape hatch. We present a fear taxonomy, demonstrate how GenAI-based governance resolves each fear, and provide an implementation sequence for builders. The key insight: where reasonable minds can differ, human review introduces variance—the very risk governance exists to prevent. AI provides consistency; humans provide accountability. GenAI governs GenAI. Humans own consequences. The architecture is proposed, not proven. The primitives exist. The integration awaits practitioners willing to build, challenge, and refine.

1 Introduction

The enterprise fear of agentic AI is backwards.

Firms delay deployment citing risks: *What if it hallucinates? What if it violates policy? What if we can't explain it to regulators?* These fears are legitimate. But the implicit assumption—that human-performed processes are safer—ignores a fundamental truth: human variance is itself a risk.

1.1 The Cost of Human Variance

Consider ten underwriters reviewing the same commercial property application. Each applies the same policy manual. Each makes a defensible decision. But the ten decisions differ—in premium, in coverage limits, in exclusions, in whether to write the risk at all.

This inconsistency creates three exposures:

Regulatory exposure. When examiners find that similar risks received dissimilar treatment, they ask why. "Underwriter judgment" is not a satisfying answer when the judgment varies by 40% across a book of business.

Litigation exposure. Plaintiff attorneys in coverage disputes search for comparables. When they find that an identical risk was approved in one case and declined in another, the inconsistency becomes evidence of arbitrary decision-making.

Audit exposure. Model risk management (SR 11-7, NAIC guidelines) requires that decision processes be consistent, documented, and reproducible. Human variance fails all three tests.

The cure to agentic AI fear is not more human oversight. It is GenAI-based governance.

1.2 The Thesis

A companion paper, "The Benjamin Button Problem" [1], established the theoretical foundation: you cannot govern exponential capability with linear processes. Agentic capability requires agentic governance. A second paper, "Semantic MVCC" [2], provided the coordination primitive: transactional state management for agentic systems.

This is the third paper in that sequence. Benjamin Button established the constraints. Semantic MVCC provided the primitive. This paper synthesizes both into a deployable architecture.

What this paper is: A next step in the right direction. A synthesis of primitives that exist into an architecture not yet fully integrated. A roadmap for builders.

What this paper is not: Proven. Complete. The final word.

No existing copilot, agent-builder, or AI platform provides what enterprises need for regulated agentic deployment. LangChain, CrewAI, and AWS Strands provide orchestration but lack transactional state semantics—no snapshot isolation, no deterministic replay, no conflict resolution framework [2]. Components reason; systems do not cohere. The governance layer must be built as platform engineering.

2 The Core Insight

The insight that unlocks the architecture is counterintuitive:

> **Where reasonable minds can differ, human review introduces variance.** That variance is the risk governance exists to prevent. AI provides consistency; humans provide accountability.

2.1 *Why Human Review Adds Variance*

When a decision requires judgment—when policy provides guidance but not determinism—reasonable humans will reach different conclusions. This is not a failure of training or competence. It is the nature of judgment.

The traditional response is to add human review: a second set of eyes, a supervisor sign-off, a committee. But review by additional humans compounds variance rather than eliminating it. The reviewer applies their own judgment, which may differ from the original decision-maker's, which may differ from the next reviewer's.

More humans means more variance, not less.

2.2 *How Constitutional AI Encodes Consensus*

Constitutional AI post-training offers a different approach. Instead of asking "what would this human decide?" it encodes "what would reasonable humans agree is the right decision, given our stated principles?"

The Constitutional AI process:

1. Articulate principles explicitly (ethics, brand, regulatory constraints)
2. Train the model to apply those principles consistently
3. The model becomes the encoded consensus

When the model encounters a judgment call, it applies the principles—the same principles, the same way, every time. This is not rigidity; the model handles novel situations by reasoning from principles. But the reasoning is consistent.

Ten underwriting decisions from Constitutional AI will be the same decision, or will surface the specific point of divergence that requires human resolution.

2.3 *The Implication*

If reasonable minds can differ, and Constitutional AI encodes the consensus they would reach, then:

- Human decision-making introduces variance
- AI decision-making provides consistency
- Human value is not in deciding, but in *defining what deciding means* and *owning the consequences*

This is the inversion: GenAI governs GenAI. Humans own consequences.

3 The Fear Taxonomy

We organize enterprise fears into four layers. Each layer represents a distinct failure mode; each has architectural answers.

3.1 *Foundation Fears*

These concern whether the AI is properly grounded *before* it acts.

Identity: "Does the AI understand our firm?" Without domain-adaptive pre-training (DAPT), the model lacks firm vocabulary, product taxonomy, and institutional context. It hallucinates jargon. It misunderstands business entities. Every prompt must re-teach context.

Values: "Does the AI share our ethics?" Without Constitutional AI post-training, the model's values are generic. It may optimize for outcomes that conflict with brand, regulatory posture, or ethical commitments. Jailbreaks are possible.

3.2 *Epistemic Fears*

These concern whether the AI knows what it claims to know.

Data Hallucination: "Will it fabricate inputs?" The model retrieves wrong documents, confuses customers, invents policy numbers. RAG retrieval is imperfect; the model may not notice.

Reasoning Hallucination: "Will it reason incorrectly from correct data?" Given accurate inputs, the model still may chain inferences incorrectly, apply wrong formulas, or reach conclusions that don't follow.

Decision Reasonableness: "Given correct inputs and valid reasoning, will it recommend something insane?" The math may be right but the conclusion may violate common sense, market norms, or unstated constraints.

3.3 *Orchestration Fears*

These concern coordination and containment in complex workflows.

Coherence: "When multiple components run in parallel, is anything coordinating state?" Agent A reads the portfolio, Agent B modifies it, Agent A acts on stale data. Without transactional isolation, race conditions corrupt decisions.

Containment: "When something fails, does failure propagate?" One bad decision triggers downstream actions. By the time humans notice, the blast radius has expanded.

Authorization: "Is this action permitted?" The agent may have capability to execute actions outside its sanctioned scope. Without an action taxonomy, anything is possible.

3.4 *Assurance Fears*

These concern verification, explanation, and improvement.

Validation: "Is the output ground truth or merely plausible?" The model produces confident, well-formatted answers. Confidence is not correctness.

Accountability: "Who approved this? Can we prove it?" When regulators or litigants ask, the answer cannot be "the AI decided." Someone must own the decision.

Drift: "Will the system degrade silently?" Models drift. Data distributions shift. What worked in January may fail in June, and no alarm sounds.

4 The Architecture

Figure 1 presents the proposed architecture. We explain each layer, then trace how a request flows through.

4.1 *Foundation Layer*

The Foundation shapes the model before runtime.

Nemotron (Open Weights FM): Base reasoning capability. Open weights ensure no vendor lock-in.

Domain-Adaptive Pre-Training (DAPT): Trains on firm-specific corpora—policy manuals, product documentation, regulatory filings, historical decisions. The model learns firm vocabulary without prompt injection.

Constitutional AI Post-Training: Encodes ethical and enterprise principles. This is the variety amplifier that Ashby requires [3]. The Constitutional AI must cover the agent's action space; every action the agent can take must have principled guidance.

Policy: SOPs, SR 11-7 requirements, NAIC guidelines. These feed into Constitutional AI at design time.

Enterprise Data: Quality-controlled data sources. The Data Factory and Data Reliability Engineering ensure that runtime context is trustworthy.

4.2 *Constitutional AI at Multiple Touchpoints*

Constitutional AI appears four times in the architecture, each serving a different purpose:

Location	Purpose
Post-Training	Shape model values permanently
Intake Filter	Block requests that violate principles
Resolver	Apply principles to pick among options
Feedback Loop	Update principles based on disagreement

This is not redundancy. Each touchpoint addresses a different failure mode. Post-training handles general alignment. Intake handles request-level filtering. Resolver handles decision-level conflict. Feedback handles continuous improvement.

4.3 *Data Assembly Loop*

The Data Assembly Loop gathers and validates context.

Intake: Every request passes through Constitutional AI filtering. Requests that violate principles are rejected before processing begins.

KV Cache (RAG-Hydrated): Retrieval-augmented generation assembles relevant context—policies, customer data, historical decisions—into a shared key-value cache. All inference paths read from the same context.

N Inference Paths: The core consistency mechanism. Rather than a single inference, the system runs N parallel paths (2 for low-risk, 3+ for high-risk decisions).

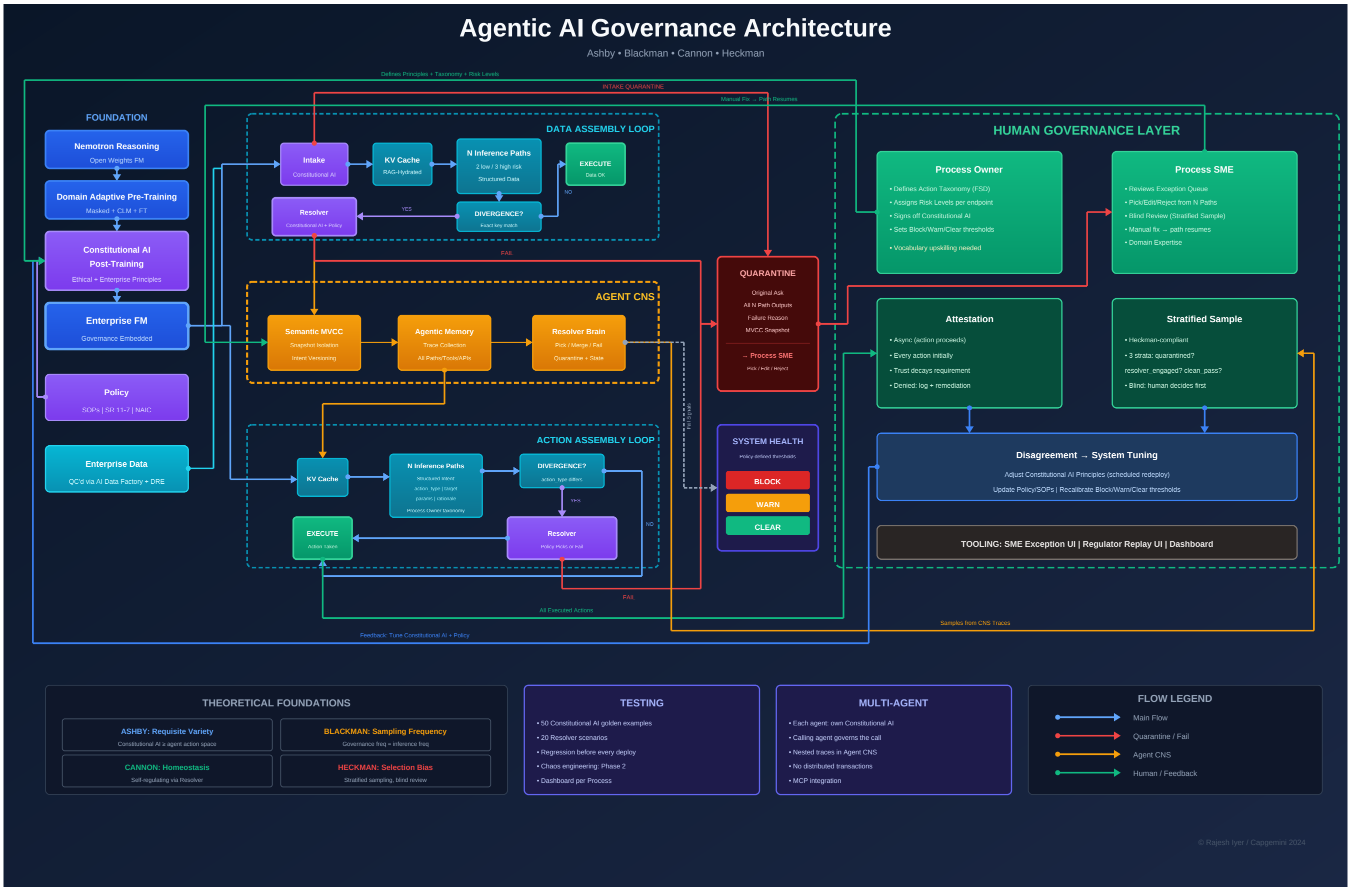
Agentic AI Governance Architecture
Ashby • Blackman • Cannon • Heckman
Defines Principles + Taxonomy + Risk Levels
INTAKE QUARANTINE
Manual Fix → Path Resumes
FOUNDATION
Nemotron Reasoning
Open Weights FM
Domain Adaptive Pre-Training
Masked + CLM + FT
Constitutional AI
Post-Training
Ethical + Enterprise Principles
Enterprise FM
Governance Embedded
Policy
SOPs | SR 11-7 | NAIC
Enterprise Data
QC'd via AI Data Factory + DRE
DATA ASSEMBLY LOOP
Intake
Constitutional AI
KV Cache
RAG-Hydrated
N Inference Paths
2 low / 3 high risk
Structured Data
EXECUTE
Data OK
NO
Resolver
Constitutional AI + Policy
YES
DIVERGENCE?
Exact key match
FAIL
AGENT CNS
Semantic MVCC
Snapshot Isolation
Intent Versioning
Agentic Memory
Trace Collection
All Paths/Tools/APIs
Resolver Brain
Pick / Merge / Fail
Quarantine + State
Fail Signals
ACTION ASSEMBLY LOOP
KV Cache
N Inference Paths
Structured Intent:
action_type | target
params | rationale
Process Owner taxonomy
DIVERGENCE?
action_type differs
YES
NO
EXECUTE
Action Taken
Resolver
Policy Picks or Fail
FAIL
All Executed Actions
Feedback: Tune Constitutional AI + Policy
QUARANTINE
Original Ask
All N Path Outputs
Failure Reason
MVCC Snapshot
→ Process SME
Pick / Edit / Reject
SYSTEM HEALTH
Policy-defined thresholds
BLOCK
WARN
CLEAR
HUMAN GOVERNANCE LAYER
Process Owner
• Defines Action Taxonomy (FSD)
• Assigns Risk Levels per endpoint
• Signs off Constitutional AI
• Sets Block/Warn/Clear thresholds
• Vocabulary upskilling needed
Process SME
• Reviews Exception Queue
• Pick/Edit/Reject from N Paths
• Blind Review (Stratified Sample)
• Manual fix → path resumes
• Domain Expertise
Attestation
• Async (action proceeds)
• Every action initially
• Trust decays requirement
• Denied: log + remediation
Stratified Sample
• Heckman-compliant
• 3 strata: quarantined?
resolver_engaged? clean_pass?
• Blind: human decides first
Disagreement → System Tuning
Adjust Constitutional AI Principles (scheduled redeploy)
Update Policy/SOPs | Recalibrate Block/Warn/Clear thresholds
TOOLING: SME Exception UI | Regulator Replay UI | Dashboard
Samples from CNS Traces
THEORETICAL FOUNDATIONS
ASHBY: Requisite Variety
Constitutional AI ≥ agent action space
BLACKMAN: Sampling Frequency
Governance freq = inference freq
CANNON: Homeostasis
Self-regulating via Resolver
HECKMAN: Selection Bias
Stratified sampling, blind review
TESTING
• 50 Constitutional AI golden examples
• 20 Resolver scenarios
• Regression before every deploy
• Chaos engineering: Phase 2
• Dashboard per Process
MULTI-AGENT
• Each agent: own Constitutional AI
• Calling agent governs the call
• Nested traces in Agent CNS
• No distributed transactions
• MCP integration
FLOW LEGEND
Main Flow
Quarantine / Fail
Agent CNS
Human / Feedback
© Rajesh Iyer / Capgemini 2024

Figure 1: GenAI-Based Governance Architecture. Foundation (left) shapes the model at design time. Two assembly loops (center) run N-path inference with divergence detection and resolution. Agent CNS coordinates via Semantic MVCC. Human Governance (right) provides attestation and calibration. Colors: red=failures, green=fixes/executions, gold=CNS traces, blue=feedback.

Each path reasons independently from shared context.

Divergence Detection: After N paths complete, the system compares outputs on key fields. If all paths agree, confidence is high. If paths diverge, resolution is required.

Resolver: When paths diverge, the Resolver—itself a GenAI component applying Constitutional AI principles—determines the outcome: pick one path, merge compatible elements, or fail to Quarantine. The Resolver does not guess; it applies encoded policy.

4.4 *Why N Paths?*

Single-path inference with temperature sampling produces one answer. That answer may be correct or incorrect; you cannot tell from a single sample.

N-path inference surfaces disagreement. If independent reasoning paths reach different conclusions from identical context, that disagreement is signal:

- The question may be ambiguous
- The context may be insufficient
- The policy may not cover this case
- One or more paths may have erred

Agreement across N paths is evidence of robustness. Divergence is evidence that human attention may be needed. This is consistency checking at inference frequency.

4.5 *Agent CNS*

The Central Nervous System coordinates state across loops.

Semantic MVCC: Multi-Version Concurrency Control provides transactional guarantees [2]. Each component reads a consistent snapshot; no mid-decision state changes. Intents are logged with full context. Any historical state can be reconstructed.

Agentic Memory: Collects traces of all paths, tool calls, and API interactions. This is the audit trail.

Resolver Brain: The central resolution authority. When the Action Assembly Loop produces divergent intents, the Resolver Brain picks, merges, or fails—and routes failures to Quarantine with full context for human review.

4.6 The Four MVCC Guarantees

Semantic MVCC [2] provides guarantees that map to regulatory requirements:

Guarantee	Regulatory Question Answered
Isolation	"What state did it observe when deciding?"
Auditability	"What exactly did it decide, and why?"
Replayability	"Can you reproduce this decision?"
Explainability	"Why this outcome and not another?"

Without these guarantees, regulated deployment is not possible. "The AI decided" is not an acceptable answer.

4.7 Action Assembly Loop

The Action Assembly Loop follows the same pattern as Data Assembly: RAG-hydrated KV Cache, N inference paths, divergence detection, resolution. But it operates on *actions* rather than *data retrieval*.

Quarantine: When the Resolver cannot resolve—when divergence is too great, when policy doesn't cover the case, when risk thresholds are exceeded—the action routes to Quarantine. Quarantine preserves: the original request, all N path outputs, the failure reason, and an MVCC snapshot for replay. Nothing is lost.

System Health: Block (system unhealthy, halt all), Warn (elevated risk, increase sampling), Clear (normal operation). The operational posture affects runtime behavior.

4.8 Human Governance Layer

Humans do not decide individual cases. They govern the system that decides.

Process Owner: Defines the action taxonomy and risk levels at design time. What actions can this agent take? What's low/medium/high risk? What thresholds trigger escalation? These are human decisions, made once, applied consistently.

Process SME: Receives quarantined items. Reviews the original request, the N paths, the failure reason. Picks the correct path, edits if needed, or rejects. SME fixes route back to Semantic MVCC—the correction becomes part of the versioned state.

Attestation: Every executed action receives attestation—a human stamp accepting responsibility for this class of decision. Initially, every action is attested. As trust builds, attestation frequency can decay for low-risk actions. But it never reaches zero.

Stratified Sample: Blind review across all strata—quarantined, resolver-engaged, and clean-pass. The reviewer doesn't know which stratum the sample came from. This prevents selection bias.

Disagreement Detection: When human review disagrees with system output, that disagreement is logged and fed back to Constitutional AI tuning.

4.9 Why Stratified Sampling Matters

This is where Heckman's insight applies [5]. Selection bias corrupts feedback loops.

If humans only review failures (quarantined items), the system learns only from failures. It improves on failure-like cases. But it may be drifting on success-like cases—and no one is watching.

The causal chain:

1. System makes errors across all decision types
2. Humans only review failures
3. System learns only from failure corrections
4. System improves on failure-like cases
5. System drifts on success-like cases
6. No one notices—success cases aren't reviewed
7. Silent degradation accumulates

Stratified sampling breaks this chain. By sampling from clean-pass decisions, humans can detect errors the system doesn't know it's making. Clean-pass disagreement rate is the most important signal—it's the only window into silent drift.

4.10 The Feedback Loop

Disagreement flows back to Constitutional AI. When human reviewers consistently disagree with system outputs on a particular class of decision, that's evidence the Constitutional AI needs updating.

This is the learning mechanism. The system doesn't just execute; it improves. But improvement is controlled: Constitutional AI changes require explicit human approval, testing, and staged rollout.

5 Scenario: Commercial Property Underwriting

A submission arrives: $5M coverage for a coastal warehouse.

Foundation: The Enterprise FM, shaped by DAPT on this carrier's policy manuals and Constitutional AI encoding risk appetite, receives the request.

Intake: Constitutional AI filter passes—this is a legitimate underwriting request.

Data Assembly: RAG retrieves: applicant history, property details, hurricane exposure models, current portfolio concentration, relevant SOPs. KV Cache hydrated.

N-Path Inference (Data): Three paths extract structured data from retrieved documents. All three agree on property value, construction type, location coordinates. Divergence on historical loss interpretation—one path reads a 2019 claim as weather-related, two read it as equipment failure.

Resolver (Data): Constitutional AI examines the underlying document. Equipment failure is explicit in the claim narrative. Resolver picks the majority interpretation. Proceeds.

Agent CNS: Semantic MVCC snapshots the assembled context. Agentic Memory logs the divergence and resolution.

Action Assembly: Three paths generate underwriting decisions:

- Path A: Approve, $48K premium, full $5M coverage
- Path B: Approve, $52K premium, $4M coverage (sublimit on flood)
- Path C: Approve, $48K premium, $5M coverage

Divergence: Paths A and C agree. Path B diverges on coverage structure.

Resolver (Action): Constitutional AI examines. Path B applied a flood sublimit that current SOPs don't require for this construction class. Resolver picks Path A/C consensus.

Attestation: Decision routes to underwriting attestation queue. Underwriter reviews summary: $5M coastal warehouse, $48K premium, no sublimits, consensus across paths. Attests.

Stratified Sample: This decision enters the clean-pass stratum. It may be sampled for blind review. If reviewer disagrees with the $48K premium, that disagreement feeds back.

Total elapsed time: seconds. Human involvement: one attestation stamp. Consistency: guaranteed—another identical submission would receive identical treatment.

6 Resolution Strategies

The Resolver doesn't guess. It applies one of three strategies based on conflict type [2, §5]:

Timestamp Ordering: When intents conflict on timing or priority, earliest wins. Deterministic, simple, but may leave value unrealized.

Allocation: When multiple intents compete for constrained resources (capital, capacity, inventory), divide proportionally or by priority weighting. No intent is fully rejected; each receives a share.

Quorum: When multiple paths evaluate the same question, require agreement above threshold before committing. Disagreement triggers escalation rather than arbitrary selection.

The MVCC paper [2] illustrates these with trading desk and insurance examples. A VaR constraint that binds doesn't reject the alpha signal—it sizes the position to fit available risk budget. The signal informs direction; the constraint determines sizing.

7 Theoretical Foundations

Four theoretical anchors underpin the architecture.

Ashby (1956) [3]: Only variety can absorb variety. A controller must match the complexity of the system it governs. Constitutional AI is the variety amplifier—it must cover the agent's action space. Simple rules cannot govern complex agents. *This is why principle coverage must match action space.*

Blackman [1, §6]: Low-frequency governance cannot stably control high-frequency behavior. Humans are low-frequency samplers (hours to days). Agents operate at computational frequency (milliseconds). Adding more humans doesn't change the frequency band—it adds latency. GenAI-based governance operates at inference frequency. *This is why human review cannot operate at runtime.*

Cannon (1929) [4]: Living systems maintain homeostasis through self-regulating feedback. The Resolver is the homeostatic mechanism—detecting divergence, restoring consistency, failing gracefully when equilibrium is impossible.

Heckman (1979) [5]: Selection bias corrupts inference. Observing only a non-random subset produces biased estimates. Stratified sampling across all strata prevents feedback loop poisoning. *This is why reviewing only failures guarantees drift.*

8 Implementation Sequence

This sequence operationalizes the Benjamin Button regression [1]: the G0→G3 governance climb that earns E0→E3 execution capability.

8.1 Phase 1: Foundation (G0→G1)

You cannot prompt your way to alignment.

Build	**Done When**	**If Skipped**
DAPT	Firm terminology unprompted	Constant prompt engineering
Constitutional AI	Red team fails	Brand/compliance risk

Fears addressed: Identity, Values. Anti-pattern: System prompts alone.

8.2 Phase 2: Coordination (G1→G2)

Without transactional state, components reason but systems don't cohere. [2]

This is the regime boundary crossing. Semantic MVCC is G2 operationalized.

Build	**Done When**	**If Skipped**
Semantic MVCC	Four guarantees met	Cannot cross boundary
Agentic Memory	All paths traced	"Why?" unanswerable

Fears addressed: Coherence, Accountability. Anti-pattern: Stateless agents.

8.3 Phase 3: Assembly Loops (G2 operational)

Conflict resolution is discovered from data, not designed from first principles. [2, §9]

Build	**Done When**	**If Skipped**
N-path + Divergence	Rate measurable, declining	Silent inconsistency
Resolver	Conflicts resolved by rule	Random selection
Quarantine	Failures route with context	Cascading failures

Development methodology [2, §9]: Deploy timestamp-only resolution. Log all conflicts. Analyze patterns. Write explicit rules for recurring patterns. Deploy and repeat. The system teaches you what it needs.

Fears addressed: Data Hallucination, Reasoning Hallucination, Reasonableness, Containment, Authorization. Anti-pattern: Single inference path.

8.4 Phase 4: Human Accountability (G3 capable)

Humans govern intent. AI governs execution.

Build	**Done When**	**If Skipped**
Attestation	Every action has owner	"Who approved?" fails
Stratified Sample	All strata sampled	Selection bias; drift
Disagreement Loop	Triggers retuning	System ossifies

Fears addressed: Validation, Drift. Anti-pattern: Review failures only.

9 Open Questions

We do not claim completeness.

Optimal N: Cost vs. coverage tradeoff is domain-specific, untested at scale. Is 3 enough? Is 5 overkill?

Retune frequency: How often should Constitutional AI update? Too frequent risks instability; too rare permits drift.

Staging guarantees: Can MVCC guarantees be earned incrementally? Or all-or-nothing?

Interpretability: N-path provides consistency, not necessarily explainability. Regulators want *why*. Resolver rationale capture needs work.

Adversarial: Constitutional AI assumes good-faith inputs. Prompt injection, data poisoning not explicitly addressed.

Unknown unknowns: What failure modes haven't we anticipated?

10 Conclusion

The sum of all fears has a proposed sum of answers.

Human variance is the risk. Where reasonable minds differ, AI provides consistency; humans provide accountability. GenAI governs GenAI at inference frequency. Humans own consequences through attestation.

The architecture synthesizes: Constitutional AI for alignment, N-path inference for consistency, Semantic MVCC for coordination, stratified sampling for calibration, feedback loops for learning.

The primitives exist. The integration awaits.

We invite practitioners to build, challenge, and refine.

References

[1] Iyer, R. "The Benjamin Button Problem: Ashby's Constraint and the Agentic Enterprise." corpXiv, Dec 2025. §6, §10, §15.

[2] Iyer, R. "Engineering the Agentic Enterprise: Semantic MVCC." corpXiv, Jan 2026. §4–5, §9.

[3] Ashby, W.R. *An Introduction to Cybernetics*. Chapman & Hall, 1956.

[4] Cannon, W.B. "Organization for Physiological Homeostasis." *Physiol. Rev.*, 9(3):399–431, 1929.

[5] Heckman, J.J. "Sample Selection Bias as a Specification Error." *Econometrica*, 47(1):153–161, 1979.