The BX3 Framework:

Implementation Protocols for Accountable Autonomous Systems

Three Functional Layers. Five Named Protocols. Guaranteed Upstream Accountability.

Jeremy Blaine Thompson Beebe
Independent Researcher

April 2026

April 2026

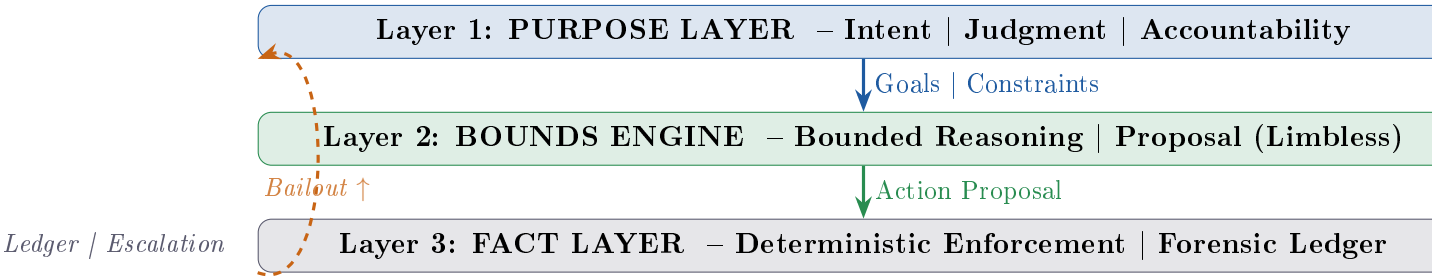

Figure 1: The BX3 Framework: Three immutable functional layers. Purpose Layer directs Bounds Engine; Bounds Engine proposes to Fact Layer; Fact Layer enforces and reports. Hard blocks flow downward; bailout propagates upward bypassing machine actors.

Abstract

The current discourse around artificial intelligence frequently presents a false binary: either AI will replace traditional software systems, or it is merely a productivity tool of limited consequence. This paper argues that both positions miss a more fundamental and practically useful insight. We propose the **BX3 Framework** — a universal architectural model organized around three immutable functional layers: the *Purpose Layer*, which provides intent, judgment, and accountability; the *Bounds Engine*, which provides interpretation, bounded reasoning, and constrained execution; and the *Fact*

Layer, which provides deterministic enforcement, hard physical constraint, and forensic auditability.

Critically, the BX3 Framework does not prescribe *who* or *what* occupies each layer. It prescribes the *functional properties* each layer must maintain — and holds any actor occupying that layer to those properties regardless of their nature. A human, an AI system, or a mechanical process may each legitimately occupy any layer, provided they satisfy that layer's functional requirements. This actor-agnostic, function-centered definition makes the framework universally applicable across human organizations, fully automated systems, multi-agent AI architectures, and any hybrid composition.

A core safety property of the framework is its upstream accountability guarantee: when any node encounters a state it cannot resolve within its bounds, accountability escalates recursively upward through the system hierarchy — bypassing all machine actors — until it reaches a human anchor. *The system fails upward into human consciousness — never downward into algorithmic chaos.*

We demonstrate that when these three functional layers are clearly separated and their properties enforced by design, systems become simultaneously more capable and less complex. We further argue that the Bounds Engine is most powerfully employed not as a replacement for the Fact Layer but as a tool for designing, accelerating, and improving it. We show that the BX3 Framework is not merely a software engineering principle but a universal structural pattern observable across law, medicine, autonomous systems, organizational design, and biological cognition. A companion engineering specification — the BX3 Protocol — is in preparation and will provide normative, certifiable implementation standards derived from this framework.

This paper is a conceptual framework and position paper. Empirical validation studies are currently in development by the author and will be published as subsequent work. This preprint establishes the theoretical foundation and research agenda that those studies will address.

Keywords: BX3 Framework, Purpose Layer, Bounds Engine, Fact Layer, artificial intelligence, deterministic systems, software architecture, human-in-the-loop, upstream accountability, recursive systems, autonomous systems, AI governance, sociotechnical systems, agentic systems, edge computing

1 Introduction

The rapid advancement and deployment of large language models and AI-based systems has generated significant confusion about the appropriate role of AI within engineered systems. A prevalent narrative suggests that AI will progressively replace traditional deterministic software, eventually subsuming most computational tasks. A counter-narrative dismisses AI as a novelty, inadequate for the reliability demands of production systems.

Both narratives fail engineers, architects, and decision-makers in the same fundamental way: they offer no principled model for how these technologies relate to one another or how

they should be combined in practice. The result is systems that are either over-reliant on AI where determinism is required, or unnecessarily constrained by legacy deterministic thinking where AI could genuinely help.

This paper proposes a clarifying framework: the **BX3 Framework**. The framework organizes any complex system into three immutable functional layers — the *Purpose Layer*, the *Bounds Engine*, and the *Fact Layer* — each defined by the properties it must maintain rather than by the type of actor that occupies it.

This actor-agnostic, function-centered definition is the framework's most important architectural property. A human, an AI system, or a mechanical process may each legitimately occupy any layer of the BX3 Framework — provided they satisfy the functional requirements of that role. What the framework prescribes is not *who* belongs in each layer but *what properties* each layer must maintain for the system as a whole to be reliable, governable, and certifiable.

A companion engineering specification — the BX3 Protocol — is in preparation and will provide normative, certifiable implementation standards derived from this theoretical foundation.

We further argue that the BX3 Framework is not a novel invention but a synthesis of well-established principles — separation of concerns [4], sociotechnical systems theory [12], control theory [13], and human-in-the-loop design [2] — applied to the specific and urgent challenge of AI integration in the current technological moment. Its value lies not in the novelty of its components but in the clarity, universality, and completeness of their unification.

Note: The first-person plural "we" is used throughout in the conventional academic sense, consistent with single-author scholarly writing.

2 Defining the Three Functional Layers

The BX3 Framework defines three functional layers. Each layer is defined by the *properties it must maintain*, not by the type of actor that occupies it. A human, an AI system, a mechanical process, an institution, or any combination thereof may occupy any layer — provided the functional requirements of that layer are satisfied.

This actor-agnostic definition is deliberate and essential. It makes the framework applicable to human-only organizations, to fully automated systems, to multi-agent AI architectures, and to any hybrid composition. It also resolves the false question of whether AI will "replace" humans in any given role: the question is not who occupies the layer, but whether the occupant satisfies the layer's functional requirements.

2.1 Layer 1: The Purpose Layer

The Purpose Layer is responsible for *intent*, *judgment*, and *accountability*. It sets Service Level Objectives, strategic goals, and the "why" that governs all downstream activity. Whatever occupies this layer must be capable of:

- Defining the goals the system exists to achieve and why.
- Making trade-off decisions under genuine uncertainty, where no algorithmic optimum exists.
- Holding accountability for outcomes — answering for the system's behavior to external parties.
- Asking and answering "why are we building this, and for whom?"
- Updating goals when context changes in ways the system was not designed to anticipate.

In the current technological moment, the Purpose Layer must remain anchored to a *human accountability anchor* — an individual or institution capable of bearing legal and ethical responsibility for the system's actions. This is the **Human Root Mandate**: in the event of system failure, accountability does not dissipate into the algorithm. It remains fixed to the human at the root. The framework does not preclude an advanced AI system from eventually occupying this layer, but until AI systems can be held meaningfully accountable for intent-level decisions, the Human Root Mandate applies.

Key property: The Purpose Layer must be *accountable* — its decisions must be attributable to an actor who can be questioned, overridden, and held responsible.

2.2 Layer 2: The Bounds Engine

The Bounds Engine is responsible for *interpretation*, *bounded reasoning*, and *constrained execution*. It performs the cognitive work of the system — analysis, pattern recognition, simulation, and optimal path proposal — but is architecturally *limbless*: it can propose but cannot execute. It lacks the authority to commit actions to the physical world unilaterally. Whatever occupies this layer must be capable of:

- Receiving goals and constraints from the Purpose Layer and translating them into proposed actions.
- Handling inputs that are ambiguous, variable, unstructured, or novel.
- Performing complex analysis — probabilistic modeling, trend analysis, simulation — within defined boundaries.

- Operating within a sandboxed cognitive environment, separated from physical execution authority.
- Escalating to the Purpose Layer when situations exceed its authority or capability.

This layer is most commonly occupied by an *AI agent or heuristic engine* in modern systems. However, a human expert, a hybrid human-AI team, or a sophisticated rule-based system may also occupy this layer when appropriate. The defining requirement is *bounded adaptability* — the ability to reason flexibly within hard constraints, never autonomously. The Bounds Engine proposes; the Fact Layer decides whether to execute.

Key property: The Bounds Engine must be *bounded* — its outputs must pass through Fact Layer validation before any physical action occurs, and its authority is strictly limited by Purpose Layer direction.

2.3 Layer 3: The Fact Layer

The Fact Layer is the physical firewall and brakes of the system. It acts as the deterministic gate through which all Bounds Engine proposals must pass before becoming real-world actions. Whatever occupies this layer must be capable of:

- Producing the same output given the same input, without exception.
- Hard-blocking any Bounds Engine proposal that violates a pre-defined safety, regulatory, or physical constraint — regardless of how confident the Bounds Engine is in its proposal.
- Maintaining a complete, tamper-evident forensic ledger of all decisions, proposals, and physical outcomes.
- Operating at the latency and reliability level required by the system's risk profile.
- Providing a basis for formal verification, regulatory certification, or legal accountability.

This layer is most commonly occupied by *deterministic software, rule engines, control systems, or physical mechanisms.* An AI system may occupy this layer only under strict conditions: fixed parameters, no generalization, no probabilistic inference, and formal verification against specification. The result of a properly implemented Fact Layer is that the system remains bounded by reality at all times — no Bounds Engine action, however confidently proposed, can violate a hard physical or regulatory constraint.

Key property: The Fact Layer must be *deterministic* — the same input must always produce the same output, all outputs must be auditable, and no Bounds Engine proposal may bypass it.

2.4 Role Occupancy Rules

The flexibility of the BX3 Framework — allowing any actor to occupy any layer — is bounded by three non-negotiable rules:

1. **Property satisfaction is mandatory.** An actor may only occupy a layer if it genuinely satisfies that layer's functional requirements. Claiming to occupy a layer without satisfying its properties is an architectural violation.
2. **Layer properties are non-negotiable.** The properties of each layer — accountability, boundedness, determinism — cannot be relaxed to accommodate an actor's limitations. If an actor cannot satisfy a layer's requirements, a different actor must be found, or the system cannot be considered BX3-compliant.
3. **All three layers must be present.** A system missing any layer is architecturally incomplete. Without the Fact Layer there are no hard constraints. Without the Purpose Layer there is no accountable intent. Without the Bounds Engine the system cannot handle the ambiguity and complexity of the real world.

3 The BX3 Framework: Structure and Properties

3.1 Structural Representation

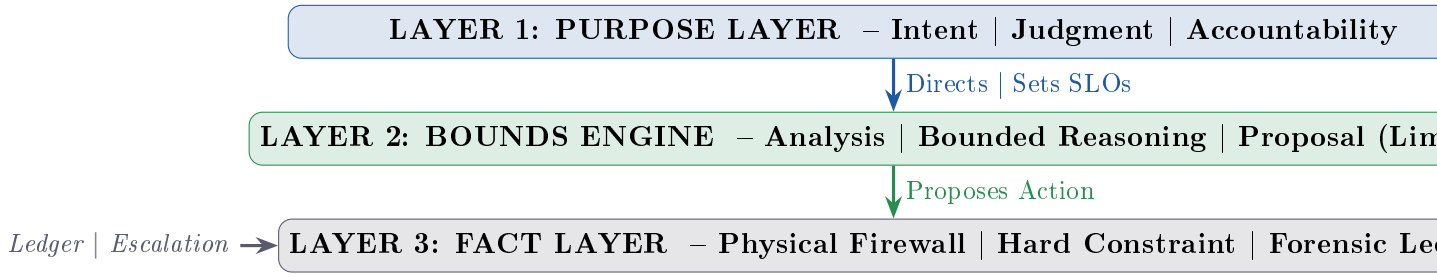

Figure 2: The BX3 Framework: Three immutable functional layers with directional communication. The Fact Layer hard-blocks Bounds Engine violations. The forensic ledger and escalation path close the accountability loop.

The BX3 Framework organizes any complex system into three functional layers. The diagram below shows the default configuration, but any actor satisfying a layer's functional requirements may occupy that position. Note the critical constraint arrow from the Fact Layer directly to the Bounds Engine — the physical firewall does not merely inform; it hard-blocks:

The BX3 Framework defines not only what each layer must do but what crosses each layer boundary. Well-defined interfaces are the mechanism by which layer isolation is maintained in practice:

Interface	From	To	What Crosses the Boundary
Direction	Purpose Layer	Bounds Engine	Goals, SLOs, constraints, authorization scope
Proposal	Bounds Engine	Fact Layer	Structured action proposals, simulation outputs
Hard Block	Fact Layer	Bounds Engine	Constraint violation signals, blocked proposal receipts
Forensic Feedback	Fact Layer	Purpose Layer	Ledger events, escalation signals, performance metrics
Override	Purpose Layer	Fact Layer	Direct commands, emergency halt, Sandbox Gate approval
Escalation	Any node	Purpose Layer	Bailout signals when bounds are exceeded

Note that the Bounds Engine and Fact Layer never share a functional plane. The Bounds Engine cannot write directly to physical actuators. The Fact Layer cannot initiate reasoning. These are not software permissions — they are architectural separations enforced by Loop Isolation (Pillar 1).

4 The Five Pillars of BX3 Implementation

The three functional layers define *what* a BX3-compliant system must contain. The Five Pillars define *how* those layers must behave in practice to maintain their properties under real-world conditions including network failures, scale, security threats, and exception states. Each pillar addresses a specific failure mode that emerges when AI, deterministic, and human systems are combined without disciplined architectural separation.

5 Formal Properties and Theorems

We formalize the three functional layers as architectural postulates – properties enforced by structure, not runtime policy.

5.1 Layer Postulates

Postulate 1: Purpose Layer – Accountability. For any BX3 node N, there exists a unique accountable actor $A(N)$ satisfying: (1) $A(N)$ is a named legal entity; (2) every action by N is attributable to $A(N)$ in the forensic Ledger; (3) $A(N)$ may delegate authority but accountability transfers only upon explicit acceptance by the delegate.

Postulate 2: Bounds Engine – Limblessness. For any BX3 node N, the Bounds Engine $BE(N)$ satisfies: (1) $BE(N)$ may propose actions but may not execute them; (2) $BE(N)$ has no direct write access to physical actuators; (3) all proposals from $BE(N)$ must pass through the Fact Layer gate before physical execution.

Postulate 3: Fact Layer – Determinism. For any BX3 node N, the Fact Layer $FL(N)$ satisfies: (1) for identical input state S and identical Purpose directive P, $FL(N)$ produces identical output state S'; (2) $FL(N)$ is provably non-Turing-complete – it computes bounded functions only; (3) $FL(N)$ cannot be modified by $BE(N)$ or any descendant of N.

5.2 Main Theorems

> **Theorem 1: Layer Isolation Guarantee.** In a BX3-compliant system, a Logic Collision – wherein the reasoning function and execution function occupy the same functional plane – is architecturally impossible.

Proof. A Logic Collision requires the Bounds Engine $BE(N)$ and Fact Layer $FL(N)$ to share a functional plane. By Postulate 2, $BE(N)$ has no direct write access to physical actuators. By definition, $FL(N)$ is the sole interface to physical execution. Since $BE(N) \neq FL(N)$ by construction, and $BE(N)$ cannot access the physical plane except through $FL(N)$, the two layers are architecturally isolated. Therefore a Logic Collision cannot occur. □

> **Theorem 2: Upstream Accountability Guarantee.** For any BX3 node N, any condition C that N cannot locally resolve escalates to $A(N)$ within a bounded number of hops in the recursive spawning tree.

Proof. Node N encountering unresolved C fires a Trigger Package to its parent $P(N)$. If $P(N)$ can resolve C within its Safety Envelope, the trigger closes. If not, $P(N)$ escalates to $P^2(N)$. The recursive spawning tree is finite and the Human Root H is always reachable. Since the Bailout Protocol mandates escalation at each unresolved hop, C reaches $A(N) = H$ within bounded steps. □

5.3 Pillar 1: Loop Isolation

Problem solved: Logic Collision — when the Bounds Engine and the Fact Layer occupy the same functional plane, enabling un-vetted autonomous actions that bypass physical constraint.

Solution: Strict isolation of the three functional layers into discrete planes. Each BX3 loop is self-contained and operates independently. A Logic Collision is architecturally impossible because the Bounds Engine never shares a functional plane with physical execution. The Bounds Engine proposes; the Fact Layer decides. These are never the same operation.

A single human Purpose Layer can govern an arbitrarily large tree of Bounds Engine agents and Fact Layer mechanisms with absolute precision, because the accountability chain remains non-collapsing regardless of system scale.

5.4 Pillar 2: Recursive Spawning

Problem solved: Logic Rigidity — static edge devices that cannot adapt to local conditions without constant cloud connectivity.

Solution: A parent node (operating at the Bounds Engine layer) births a child BX3 loop by generating a *Worksheet* — a containerized, self-contained logic set encapsulating the parent's Purpose for a specific local context — and deploying it over-the-air to the child node.

Each Worksheet carries a hard-coded pointer to the parent's Purpose, preventing autonomous drift. The child loop applies the parent's intent to local sensor data independently, without requiring a constant cloud heartbeat. If cloud connectivity is lost, the child node executes the last-known-good Worksheet based on local inputs. The system maintains integrity and local reflexes in degraded network conditions (*Local Survivability*).

This mechanism allows a single human Purpose Layer to project authority and logic across an arbitrarily large distributed system while preserving BX3 layer integrity at every node:

5.5 Type Specifications

This section provides formal type specifications for the key data structures that govern BX3 loop behavior. These specifications are implemented in the AgentOS reference implementation [3], an open-source TypeScript project maintained at `https://github.com/bxthre3inc/agentos`. Each type maps to a named concept in the paper.

Safety Envelope.

The Safety Envelope formally specifies zero-tolerance operational boundaries (Section 4):

- `waterRights`: `maxDeviationGal = 0` (zero tolerance)
- `soilMoisture`: `minPct`, `maxPct`, `currentPct`
- `temperature`: `toleranceF = 5`, `measuredAt: [sensorId]`
- `equipmentLoad`: `maxAmp`, `minCoolMin`
- `bailoutThreshold`: `latencyMs < 1000`, `escalateTo: HumanRoot`
- `ledgerAccess`: `NONE` — the Bounds Engine has zero write access to the ledger

Any proposed action that would violate a Safety Envelope parameter is blocked unconditionally by the Sandbox Gate.

The P5–P9 Decision-to-Execution Pipeline.

Four sequential planes govern the path from Bounds Engine reasoning to Fact Layer execution:

1. **P5 (Decision):** The Bounds Engine's active action proposal. Status transitions: `proposed` → `sandbox_submitted` → `sandbox_confirmed` | `block` → `executed`.
2. **P6 (Projection):** The Bounds Engine's simulation output. `confidenceScore` $\in [0.0, 1.0]$. `violatingParameters` (empty = all within bounds).
3. **P9 (Projection Confirmation):** The Sandbox Gate's pre-execution verdict. If `safe = true`, the Fact Layer is unlocked for this action.
4. **P8 (Execution):** The Fact Layer's physical actuation state, written only after P9 confirmation. The Fact Layer accepts execution commands from the Sandbox Gate exclusively.

Sandbox Gate verdict rules:

- `EXECUTE` iff $\forall e \in \text{SafetyEnvelope} : \text{projectedValue}_e \subseteq e.\text{bounds}$
- `BLOCK` iff $\exists e \in \text{SafetyEnvelope} : \text{projectedValue}_e \notin e.\text{bounds}$
- `REVIEW` iff $\exists e \in \text{SafetyEnvelope} : |\text{projectedValue}_e - e.\text{boundary}| < \delta$ (default $\delta = 5\%$ of parameter range)

BailoutTrigger.

The Bailout Protocol (Pillar 5) fires when any node encounters a condition it cannot resolve. Key fields:

- `triggerCondition`: `'capability_boundary'` | `'safety_envelope_predicted'` | `'accountability`
- `confidence`: float $\in [0.0, 1.0]$ — 0.0 = uncertain, 1.0 = certain requires human resolution
- `nodeState`: snapshot of planes P1, P2, P5, P7 at fire time
- `propagationChain`: ordered list of `{nodeId, action, timestamp}` — machine actors append `'ESCALATED'`, Human Root appends `'HUMAN_REVIEW'`
- `status`: `'fired'` → `'delivering'` → `'escalating'` → `'human_review'` → `'resolved'`

Machine Actor Exclusion: a Bounds Engine node MUST append `'ESCALATED'` and continue propagation — it has no legal authority to resolve any unresolved BailoutTrigger.

LedgerEntry.

The Forensic Ledger (Pillar 5) provides cryptographically chained, plane-isolated audit records. Each entry is immutable once written:

- `entryNumber`: sequential integer (genesis entry = 1)
- `plane`: `'P1'..'P9'` — enforces plane isolation
- `seal`: `SHA256(payload ‖ previousSeal ‖ timestamp)` — any hash mismatch = tampered record
- `previousSeal`: hash of prior entry (`'GENESIS'` for entry 1)
- `chainVersion`: integer ≥ 1

Verification: recompute SHA256 chain from the genesis event. Entry 1 is sealed at hardware manufacturing time and is non-alterable.

BX3Loop.

A BX3Loop is the fundamental unit of the framework:

- `id`: unique node identifier
- `type`: `'root'` (`parentId = null`, `parentPurposeHash = 'HUMAN_ROOT'`) | `'child'`

- `parentPurposeHash`: `SHA256` of parent's Purpose text — immutable after deployment
- `purpose`, `bounds`, `fact`: the three layer state objects
- `worksheet`: the deployed container (immutable by the child)
- `layerPlane`: current active plane `'P1'..'P9'`
- `status`: `'active'` | `'suspended'` | `'bailout_pending'` | `'closed'`

6 Application Across Domains

The BX3 Framework is not a software engineering principle alone. The same three functional layers — an accountable Purpose Layer, a bounded Bounds Engine, and a deterministic Fact Layer — appear across many complex domains. In each case, different types of actors occupy each layer, confirming that the framework's power lies in its functional definitions rather than in any prescription about actor type.

6.1 Autonomous Systems and Sensor Networks

In sensor-driven autonomous systems, the BX3 Framework maps directly onto system architecture. Human engineers occupying the Purpose Layer define mission parameters, safety constraints, and acceptable risk envelopes. AI processing layers occupying the Bounds Engine interpret sensor data, handle ambiguous environments, and propose real-time decisions within those constraints. Deterministic control systems occupying the Fact Layer enforce hard constraints — collision avoidance thresholds, actuator limits, safety interlocks — that cannot be overridden by the Bounds Engine under any circumstances.

This architecture is critical where the cost of Bounds Engine error is physical and potentially irreversible. The Fact Layer does not need to handle every situation — only to prevent catastrophic outcomes while the Bounds Engine and Purpose Layer resolve ambiguity.

6.2 Legal Systems

Legal systems demonstrate the actor-agnostic property of the BX3 Framework clearly. Legislatures and judges occupy the Purpose Layer — defining the purpose and interpretation of law, exercising judgment in novel cases, and bearing institutional accountability. AI systems increasingly occupy parts of the Bounds Engine — assisting with legal research, document analysis, and pattern recognition across case law. The written law itself, procedural rules,

and enforcement mechanisms occupy the Fact Layer — the same statute applies the same way in the same circumstances, regardless of who the parties are.

6.3 Medicine

Physicians occupy the Purpose Layer — exercising judgment incorporating patient context, ethical considerations, and probabilistic reasoning beyond any protocol. AI systems occupy portions of the Bounds Engine — assisting with imaging analysis, drug interaction checking, and population-level pattern recognition. Drug dosage protocols, surgical checklists, equipment operation specifications, and regulatory requirements occupy the Fact Layer — deterministic rules that constrain both physician and AI behavior alike.

6.4 Organizational Design

Organizations themselves exhibit the BX3 structure. Executive leadership occupies the Intent Layer — setting strategic direction and bearing accountability. Knowledge workers and AI-augmented teams occupy the Bounds Engine — interpreting strategy and executing flexibly within organizational guidelines. Policies, compliance requirements, contractual obligations, and financial controls occupy the Fact Layer — rules that apply consistently regardless of who is executing or what the AI recommends.

6.5 Biological Cognition

Notably, the BX3 Framework mirrors the layered architecture of biological cognition [6]. The autonomic nervous system and reflexes occupy the Fact Layer — deterministic, fast, and non-negotiable. Learned heuristics, intuitions, and pattern recognition occupy the Bounds Engine — efficient responses to familiar situations. Conscious deliberative reasoning occupies the Purpose Layer — slow, deliberate, and engaged only for genuinely novel, high-stakes, or ethically complex situations.

This convergence suggests the BX3 Framework reflects something deeper than an engineering preference. The same three-layer functional architecture appears to be a near-optimal solution for any system that must simultaneously be reliable, adaptive, and accountable — regardless of whether that system is biological, organizational, legal, or computational.

7 Why Deterministic Systems Will Not Be Replaced

A common claim is that sufficiently advanced AI will eventually subsume deterministic software. We argue this position misunderstands the distinct value of determinism as a *property*, not merely a *limitation*.

7.1 The Value of Determinism is Not Compensable

Determinism provides properties that probabilistic systems cannot replicate:

- **Reproducibility:** The ability to reproduce any past output given the same input, enabling debugging, auditing, and legal accountability.
- **Formal verification:** The ability to mathematically prove that a system satisfies certain properties under all possible inputs.
- **Certification:** Regulatory frameworks in aviation, medical devices, financial systems, and safety-critical infrastructure require deterministic behavior as a precondition for approval.
- **Latency:** Hard real-time requirements (microsecond response times) are achievable with deterministic systems and not with current AI inference pipelines.

7.2 The Infrastructure Argument

Every AI system in production today runs on top of vast deterministic infrastructure: operating systems, databases, networking stacks, authentication systems, billing pipelines, and monitoring tools. The claim that AI will replace software is structurally self-refuting — the AI systems making this possible are themselves dependent on deterministic software that will not and should not be replaced.

7.3 The Regulatory and Trust Barrier

Even in domains where AI could theoretically replace deterministic systems on a performance basis, the practical barriers are substantial. Organizations with core operations dependent on software will not replace functioning, auditable, certified systems with probabilistic alternatives without extraordinary evidence of equivalent reliability and auditability. The burden of proof is appropriately high, and current AI systems do not meet it for most production contexts.

8 Relationship to Prior Work

The BX3 Framework synthesizes established ideas from separation of concerns, sociotechnical systems theory, cybernetics, systems safety, and human-in-the-loop design, but it departs from prior work by defining immutable functional layers according to required properties rather than according to actor type. The closest contemporary analogue is the intelligent sociotechnical systems framework [14], which similarly emphasizes structured coordination between human and technical components; however, BX3 extends this line of work by explicitly isolating a deterministic Fact Layer and by specifying an actor-agnostic architecture in which any occupant of a layer is bound by that layer's functional obligations.

Recent agentic-AI literature reinforces several needs that BX3 attempts to formalize. Tiered Agentic Oversight (TAO) [7] demonstrates that hierarchical supervision among specialized agents can reduce error propagation in safety-critical settings. BX3 differs in making human accountability the required terminal endpoint of unresolved escalation rather than a high-tier supervisory option. TAO's hierarchy routes to human oversight as a high-risk escalation pathway; BX3 makes this routing unconditional and architectural. Likewise, recent work on production-grade agent architectures [1] recommends fail-safe behavior, sandbox-first execution, deterministic fallback workflows, and human approval gates for risky actions; BX3 incorporates these concerns as named architectural pillars — most notably the Sandbox Gate and Bailout Protocol — rather than as implementation heuristics.

While recent agentic-architecture literature recommends sandbox-first execution and human approval gates for high-risk actions, BX3 specifies a narrower architectural requirement: proposed interventions must be evaluated in a digital twin and cleared through a role-bounded Sandbox Gate tied to Root Tunneling authority before the Fact Layer is unlocked.

BX3 also sits naturally alongside emerging governance and compliance literature. Koch [8] argues that standards such as ISO/IEC 42001 [5] and the NIST AI RMF [9] do not themselves provide implementable runtime guardrails, and instead require translation across governance, design-time, runtime, and assurance layers. BX3 may be read as one candidate architecture for that translation. In parallel, formal-verification approaches such as the Lean-Agent Protocol [10] show how proposed agentic actions can be forced through deterministic verification gates before execution. BX3 generalizes that intuition beyond theorem proving: wherever execution must remain non-probabilistic, a protected deterministic layer is architecturally indispensable.

The recurrence of layered models in global AI governance discourse [11] further suggests that complex AI systems are increasingly being understood through stratified functional ab-

stractions, although those models are policy-descriptive rather than engineering-prescriptive.

8.1 Comparative Analysis with Existing Architectures

We evaluate the BX3 Framework against representative architectures.

Architecture	Sep.	Hum. Acct.	Det. Fact	Safety Env.	Audit	Dist. Auto.
BX3 Framework	✓	✓	✓	✓	✓	✓
ROS/ROS2	✓	×	×	×	×	✓
Kubernetes	✓	×	✓	×	partial	✓
JADE Multi-Agent	✓	×	×	×	×	✓
Holochain	✓	×	×	×	✓	✓
ISO/IEC 42001	✓	✓	×	×	×	×
NIST AI RMF	✓	✓	×	×	×	×

Interpretation: The BX3 Framework is the only architecture evaluated that simultaneously satisfies all six properties by architectural construction. Governance frameworks (ISO, NIST) lack deterministic enforcement; infrastructure frameworks (Kubernetes, ROS) lack accountability; agent frameworks (JADE, Holochain) lack formal safety envelopes.

Note on ROS, Kubernetes, JADE, and Holochain: ROS provides modular communication but no accountability layer. Kubernetes provides container orchestration with isolation but no bounds-engine concept. JADE provides agent messaging but no deterministic fact layer. Holochain provides cryptographic audit but no safety-envelope enforcement. Each provides partial coverage; BX3 integrates all six.

9 Conclusion

The question of how humans, AI, and traditional software should relate to one another is not merely a technical question. It is an architectural, organizational, ethical, and regulatory question with significant and growing practical consequences.

The BX3 Framework proposes a principled and universal answer. It organizes any complex system into three functional layers — Purpose, Bounds Engine, and Fact — each defined by the properties it must maintain rather than by the type of actor that occupies it. Any actor capable of satisfying a layer's functional requirements may occupy that layer: human, AI, mechanical, institutional, or hybrid. What cannot vary are the properties themselves — accountability in the Purpose Layer, boundedness in the Bounds Engine, and determinism in the Fact Layer.

This actor-agnostic definition is the framework's most important contribution beyond its predecessors. It makes the BX3 Framework applicable to the full range of systems that exist and the full range of systems that are coming: human-only organizations, fully automated pipelines, multi-agent AI architectures, and hybrid compositions that do not yet have names. In each case, the framework asks the same three questions: Is there an accountable layer that sets purpose and bears responsibility? Is there a bounded layer that reasons and proposes within defined constraints? Is there a deterministic layer that enforces, validates, and audits? If any layer is missing or its properties are not maintained, the system is architecturally incomplete — regardless of how capable its components are individually.

The BX3 Framework is not entirely new. Its pattern is visible in legal systems, medical practice, autonomous vehicles, biological cognition, and organizational design — wherever reliable systems have been built to handle a world that is simultaneously rule-bound and unpredictable. What is new is the urgency of making the pattern explicit, naming its layers by function rather than actor, and building from it deliberately — at a moment when the temptation to collapse these roles, and the cost of doing so, have never been higher.

10 Limitations and Future Work

The BX3 Framework has specific boundary conditions:

- **Human Root scalability:** As the recursive tree scales, the human accountability anchor may become a bottleneck. Distributed accountability mechanisms require further specification.
- **Legacy interoperability:** Retrofitting SCADA, PLCs, and proprietary SaaS to BX3 layer separation may require significant refactoring. The cost-benefit of such migration is an open empirical question.
- **Formal verification toolchain:** A standardized Safety Envelope specification language (analogous to TLA+) is planned as a companion specification.

- **Performance overhead:** Initial AgentOS benchmarks suggest 2–5ms per execution cycle – acceptable for agricultural and enterprise applications but potentially problematic for hard real-time systems.
- **Cross-organizational deployments:** Multiple legal entities in the same BX3 system require additional accountability transfer protocols not yet specified.

Peer Review Instructions

Review Criteria

1. Originality and Contribution (30%): Does the paper introduce new concepts or primarily synthesize existing work? The paper's primary contribution is the unification of three functional layer definitions with five enforcement pillars. Novelty lies in: (a) actor-agnostic layer definition; (b) upstream accountability guarantee; (c) deterministic enforcement integrated with bounded AI reasoning.

2. Technical Soundness (30%): Are the postulates (Accountability, Limblessness, Determinism) well-defined and internally consistent? Are the two theorems correctly derived from the postulates? Are there counterexamples?

3. Clarity and Completeness (20%): Is the framework sufficiently specified to be implemented without additional clarification? Are the five pillars clearly distinguished? Is the actor-agnostic property consistently applied?

4. Significance (20%): Does the framework address a genuine gap in AI architecture and governance practice? Are the cross-domain examples (legal, medical, biological, organizational) appropriate and well-chosen?

Submission Checklist

All citations complete and correctly formatted

All figures have captions and are referenced in text

Theorem 1 (Layer Isolation) and Theorem 2 (Upstream Accountability) formally stated and proved

Postulates 1–3 clearly labeled

Five pillars clearly named and distinguished

Actor-agnostic property stated explicitly in Section 2

Limitations acknowledged in Section 10

Abstract accurately reflects paper contributions

Metadata

Keywords: BX3 Framework, Purpose Layer, Bounds Engine, Fact Layer, artificial intelligence, deterministic systems, software architecture, human-in-the-loop, upstream accountability, autonomous systems, AI governance, sociotechnical systems, agentic systems, edge computing

Subject Areas: Computer Science – Artificial Intelligence; Computer Science – Software Engineering; Computer Science – Multiagent Systems

Conflicts of Interest: The author is affiliated with Bxthre3 Inc., a company developing commercial implementations of the BX3 Framework.

Acknowledgments

The author wishes to acknowledge the foundational contributions of the researchers cited herein, whose work across control theory, sociotechnical systems, and computer science provides the shoulders on which this synthesis stands.

References

[1] Mamdouh Alenezi. From prompt-response to goal-directed systems: The evolution of agentic AI software architecture, 2026.

[2] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S. Lasecki, Daniel S. Weld, and Eric Horvitz. Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 2–11, 2019. doi: 10.1609/hcomp.v7i1.5285.

[3] Jeremy Blaine Thompson Beebe. Agentos: An open implementation of the bx3 framework for ai workforce orchestration. GitHub repository, 2026. URL https://github.com/bxthre3inc/agentos. Open-source reference implementation, 16 tests passing.

[4] Edsger W. Dijkstra. On the role of scientific thought. In *Selected Writings on Computing: A Personal Perspective*, pages 60–66. Springer-Verlag, New York, 1982.

[5] ISO/IEC. ISO/IEC 42001:2023 — information technology — artificial intelligence — management system. Technical report, International Organization for Standardization, Geneva, 2023.

[6] Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York, 2011.

[7] Yubin Kim, Hyewon Jeong, et al. Tiered agentic oversight: A hierarchical multi-agent system for healthcare safety, 2025.

[8] Christoph Koch. From governance norms to enforceable controls: A layered translation method for runtime guardrails in agentic AI, 2026.

[9] National Institute of Standards and Technology. Artificial intelligence risk management framework (AI RMF 1.0). Technical Report NIST AI 100-1, U.S. Department of Commerce, Gaithersburg, MD, 2023.

[10] Darren Rashie and Vikram Rashi. Type-checked compliance: Deterministic guardrails for agentic financial systems using Lean 4 theorem proving, 2026.

[11] Nour Sabah, Lisa Chen, Jin Park, Aisha Rahman, Thanh Nguyen, Denis Volkov, Shirley Wang, Gabriele Kowalski, Simon Liu, Arjun Singh, Riya Nakamura, Marco Torres, Hyeon Kim, Emily Zhao, Dev Patel, Chidi Okonkwo, Lutz Fischer, Asta Nielsen, Pelle Holm, Evgeny Delong, Håkan Svensson, Kenzo Yamamoto, Kofi Osei, François Dubois, Mia Ivanova, Ana Santos, Igor Volkov, and Wei Chen. Layered functional abstractions in global AI governance: A cross-jurisdictional survey, 2026.

[12] Eric Trist. The evolution of socio-technical systems. Occasional paper, Ontario Quality of Working Life Centre, Toronto, 1981.

[13] Norbert Wiener. *Cybernetics: Or Control and Communication in the Animal and the Machine*. MIT Press, Cambridge, MA, 1948.

[14] Wei Xu and Yue Gao. An intelligent sociotechnical systems (iSTS) framework: Enabling a hierarchical human-centered AI (hHCAI) approach. *IEEE Transactions on Human-Machine Systems*, 2024. arXiv:2401.03223.