HIGHLAND REFLEXIVE PRESS

HRP-RN-001

Constitutional Causal Continuity

Preserving Human Authority Across Reasoning, Delegation, Trajectories, and Self-Modification in Agentic Systems

TONY T. NGUYEN I NGUYỄN TỪ VĨNH TRƯỜNG

PUBLIC RESEARCH RELEASE

FREE PUBLIC READING; CITATION ENCOURAGED; REUSE REQUIRES WRITTEN PERMISSION

VERSION 1.8.2 · AUGUST 2026

Big Spring, Texas · Reader Edition 2026

Constitutional Causal Continuity

Preserving Human Authority Across Reasoning, Delegation, Trajectories, and Self-Modification in Agentic Systems

Tony T. Nguyen

Highland Reflexive Press

Correspondence, collaboration, and licensing inquiries: tony@highlandreflexivepress.net

MANUSCRIPT VERSION: v1.8.2 — Research manuscript — verified executable-artifact edition, August 2026

© 2026 Tony T. Nguyen | Highland Reflexive Press. All rights reserved, including patent, trademark, trade-secret, contractual, and other intellectual-property interests.

Intellectual-property notice. This manuscript presents a research-level architectural synthesis. It does not grant a patent, copyright, trademark, trade-secret, contractual, implementation, or other license. Patent-sensitive implementation detail is intentionally omitted where unnecessary to the scholarly argument.

Abstract

AI governance increasingly confronts a problem that ordinary access control does not fully capture. A computational system may be authorized to act, may reason correctly, may pass a verifier, and may execute only locally permissible operations, yet still produce a global consequence that no legitimate human or institution authorized. The failure can arise before authorization, through manipulation of the human decision environment; during delegation, through scope or temporal amplification; during reasoning, when verification is mistaken for authority; during execution, when individually admissible actions compose into an unauthorized trajectory; after deployment, when model, verifier, topology, or authority state changes between admission and commit; or during adaptation, when a system modifies its future constitutional behavior through indirect substrates rather than an explicit policy file.

This paper introduces Constitutional Causal Continuity as the system-level organizing principle of the Human Reasoning Infrastructure (HRI), a ten-part architecture that separates epistemic computation from authority-bearing consequence while preserving traceable human authority across reasoning, judgment, delegation, execution, failure recovery, and constitutional evolution. Rather than claiming novelty for familiar primitives such as authorization, provenance, reference monitoring, or delegation, the paper asks a narrower systems question: what conditions are required so that no machine-derived transformation can become sufficient, by itself or through composition, to create an authority-bearing effect outside a current-valid human constitutional lineage?

The architecture is expressed through a cross-module semantic type system, assume–guarantee contracts, causal and trajectory closure conditions, decision-environment integrity, justified verification, delegation non-amplification, proxy-amendment controls, and commit-time revalidation. A runnable finite-state reference model, distributed with reproducibility artifacts, explores 960 reachable temporal states under the full guard set and finds no protected commit violating its modeled temporal closure predicate. The model instantiates only a subset of the manuscript's broader global predicates; it does not by itself prove full constitutional-causal closure. Within the specified transition system, removing any one of six commit-time guards—authority validity, authority-epoch compatibility, proof freshness, topology-certificate freshness, constitutional compatibility, or decision-environment integrity—produces a short counterexample. A separate two-transaction, limit-one concurrency model shows an existence case in which stale, non-atomic accounting permits aggregate trajectory overflow; composition-safe mechanisms such as atomic reservation or serializability prevent that modeled failure. Finally, enabling any modeled implicit cast from reconstruction, drift measurement, IP verification, authorship, consensus, or verified reasoning into authority yields an unauthorized-effect trace under the supplied semantic-cast bug semantics.

© 2026 Tony T. Nguyen | Highland Reflexive Press. All rights reserved.

These results do not establish universal safety or patentability. They establish a falsifiable research program: human authority is treated not as a user-interface ritual but as a typed, temporally live, causally relevant invariant that must survive transformations across the whole system. The central claim is therefore not that machines must reason less, but that reasoning may scale without authority scaling with it.

Keywords: agentic AI, AI governance, human authority, authorization, runtime governance, causal closure, trajectory governance, constitutional AI, delegation, provenance, model checking, human-in-the-loop, verifier independence

1. Introduction

Agentic systems are shifting the governance problem from what a model says to what a computational process can cause. Modern agents can read private data, invoke tools, delegate subtasks, modify records, transact, communicate externally, and maintain state across long-running workflows. NIST has accordingly identified agent identity, authorization, auditing, and non-repudiation as emerging infrastructure problems for software and AI agents [1]. Recent research has attacked important parts of this problem from several directions: human-delegation provenance [2], path-dependent runtime governance [3], typed delegation and authorization envelopes [4], behavioral constitutions and governance receipts [5], production-agent runtime governance [6], and authorization continuity for evolving agents [7].

These efforts collectively expose a deeper point: authorization is no longer a single check on a single request. The object being governed may be a path, a delegation chain, a mutable agent, a sequence of effects, or a composite principal. Yet even these objects do not exhaust the problem. A system can remain formally compliant at every local point while defeating the ratified intent of the human authority structure globally.

Consider six examples.

First, a model produces a highly accurate recommendation, and an application silently treats high confidence as a reason to broaden its permission. The conclusion may be correct; the authority inference is invalid.

Second, a human approves a consequential action after an interface suppresses alternatives, hides counterevidence, repeatedly prompts, or establishes a manipulative default. A click exists, but the decision environment that supposedly created independent human authority is defective.

Third, an agent executes a series of individually permissible transactions. Each transaction passes its local limit, yet their cumulative exposure exceeds the authorized portfolio state.

Fourth, a transaction is admitted under valid authority, but that authority is revoked before the external effect commits. If admission is treated as permanently freezing validity, the system executes under stale authority.

Fifth, an agent is forbidden from editing a constitutional policy artifact but changes retrieval data, memory, routing, verifier selection, or another substrate that materially changes future governed behavior. Artifact integrity holds while constitutional behavior drifts.

Sixth, a system generates its own test, interprets the result, declares itself safe, and then uses that declaration to unlock an authority-bearing effect. The chain contains “verification,” but no epistemically independent and checkable basis sufficient for the authority decision.

The Human Reasoning Infrastructure addresses these failures by making a categorical separation that remains constant throughout the stack:

> Epistemic success does not confer authority.

This principle is stronger than a preference for a human approval button. It requires a system architecture in which reasoning, evidence, verification, judgment, authority, control, and effect are different semantic types; transformations between them are explicit; and the conversion into authority is grounded outside machine epistemic computation.

© 2026 Tony T. Nguyen | Highland Reflexive Press. All rights reserved.

The paper develops this architecture at the level of ten linked HRI patent-family specifications (Module 1 through Module 10) and a later formal proof layer. The contribution is not the claim that every primitive is individually new. Rather, it is a candidate system-level synthesis around Constitutional Causal Continuity: preserving valid human-originating authority through machine-mediated transformations from decision formation to protected consequence and onward through adaptation and constitutional evolution.

1.1 Contributions

This paper makes five principal contributions.

1. A ten-layer architectural decomposition. HRI separates authority representation, human-reasoning reconstruction, judgment drift, distributed authority continuity, adversarial IP reasoning, causal intellectual attribution, judgment measurement, runtime reasoning discipline, causal effect closure, and cross-cutting constitutional composition.
2. A cross-module semantic firewall with a three-way authority distinction. Runtime objects are typed as data, evidence, inference, verification, recommendation, human decision, authority token, authority-admissible object, control, effect, or unknown; AUTHORITY_LEGITIMACY is treated separately as a non-computational normative status. Implicit epistemic-to-authority casts are prohibited, and neither token validity nor attestation admissibility is treated as implying legitimacy.
3. A conditional assume–guarantee closure specification. Protected completion is specified as requiring a current-valid human authority lineage, decision-environment integrity, justified verification, causal admissibility, trajectory integrity, and constitutional continuity, stated explicitly as conditional on named axioms (Section 5.0) rather than as an unconditional target. The supplied bounded executable model mechanically instantiates only a subset of these predicates.
4. Temporal and compositional hardening. The architecture distinguishes admission-time from commit-time validity, requires freshness across authority epochs, verifier versions, causal topology, and constitutional state, and treats global trajectory capacity as a shared resource rather than a collection of local permissions.
5. Executable bounded consistency and counterexample analysis. A finite-state reference model is supplied with source hashes, ablation tests, semantic-cast counterexamples, and a TLA+ mirror of the core temporal state machine. The full temporal abstraction explores 960 reachable states without a modeled commit-predicate violation over the mechanically instantiated conjuncts identified in Section 7.0; each major guard ablation yields a concrete counterexample within that transition system. Because the guards and the modeled predicate are definition-coupled, these results are reported as internal consistency and counterexample analysis rather than general falsification evidence. The supplied TLA+ source does not encode the separate Python concurrency or semantic-cast experiments, and no TLC execution is claimed here.

2. Related Work and the Residual Systems Problem

HRI occupies a rapidly developing research neighborhood. The right comparison is therefore not “HRI versus no prior work,” but which layer of the governance problem each architecture closes and which layers remain outside its guarantee.

2.1 Identity, delegation, and provenance

Dalugoda’s Human Delegation Provenance protocol binds a human authorization event to a session and records multi-hop delegation in a signed append-only chain [2]. This is directly relevant to HRI’s insistence that terminal machine action remain traceable to human-originating authority. Ibrahim and Li formalize recursive delegation and contextual scope as a typed authorization overlay, computing an agent authorization envelope from delegation, session scope, and resource relations [4]. These works establish strong neighbors for authority lineage, recursive delegation, and scope attenuation.

HRI does not treat those primitives as novel by themselves. Its residual question is what happens when valid delegation meets reasoning outputs, human decision shaping, cumulative trajectories, system adaptation, and causal effects that cross module boundaries.

© 2026 Tony T. Nguyen | Highland Reflexive Press. All rights reserved.

2.2 Paths and runtime governance

Kaptein, Khan, and Podstavnychy argue that the execution path, rather than a single action, is the natural object for runtime governance [3]. Their framework evaluates policies over a partial path, proposed next action, agent identity, and organizational state. Tallam similarly emphasizes workflow-level governance, composite principals, authority attenuation, and stop-anywhere mediation in a five-plane production architecture [6].

HRI shares the conclusion that request-time access control is insufficient. Its additional distinction is between path validity and constitutional causal continuity: a path can be locally governed while the decision environment that created authority, the semantic type of an upstream output, the cumulative authorized envelope, or the constitutional substrate itself has become invalid.

2.3 Behavioral constitutions and governance receipts

AgentBound combines delegated authorization, owner-signed behavioral constitutions, and site action contracts, then produces verifiable governance receipts [5]. This is a strong neighboring architecture for deterministic governance between authorization and execution. HRI's Module 9 and Module 10 occupy overlapping ground around protected effects, behavioral constraints, provenance, and runtime control.

The distinction advanced here is not "behavioral constitutions do not exist." It is that a constitutional system must also govern how authority is formed, how locally valid effects compose, how verification obtains epistemic standing, and how the constitution itself changes—including changes by materially equivalent proxy.

2.4 Authorization continuity under agent evolution

Zhang and Zhang formulate authorization continuity for evolving agents through a state-bound model with an immutable effect ceiling and a transition envelope fixed at grant time [7]. Under complete mediation, sound effect abstraction, attenuating delegation, and monitor integrity, their theorem prevents mutation from amplifying protected effects beyond the user-issued ceiling. Below that ceiling, active authority may contract and may expand only under specified evidence conditions; agent-produced evidence cannot raise the ceiling. This hard-ceiling construction is a strong and complementary mechanism for effect-time governance.

The residual governance gap addressed by HRI lies partly outside the ceiling itself. First, a hard ceiling presupposes that the grant from which the ceiling is derived is an admissible authority event. HRI therefore places Decision-Environment Integrity upstream of effect-time enforcement: a ceiling cannot repair approval laundering if the originating human authorization was materially shaped by suppressed alternatives, hidden counterevidence, manipulative defaults, or other decision-environment defects. Second, an agent may remain within an unchanged numerical or categorical effect ceiling while altering the behavioral substrate that implements the constitution. HRI's Constitutional Proxy Amendment requirement therefore treats material changes to memory, retrieval, routing, prompts, verifier/model selection, or cumulative small mutations as governance-relevant even when their immediate effects remain under the fixed ceiling. The two approaches are thus best read as complementary: Zhang and Zhang sharply bound what evolving agents may cause under a grant; HRI asks whether the grant was legitimately formed and whether the governing behavior that interprets it has remained constitutionally continuous.

2.5 Commit-boundary freshness and decision integrity precedents

Commit-time freshness is not claimed here as a standalone novel primitive. Santos-Grueiro's 2026 work on commit-time authorization for LLM agents independently develops closely related boundary-freshness conditions for mutable authority witnesses [8], while Zanzibar provides an earlier distributed-authorization precedent for freshness and consistency tokens [9]. HRI's residual claim is the wider constitutional-causal composition that couples freshness with semantic epistemic/authority typing, human authority lineage, trajectory composition, decision integrity, and constitutional continuity.

The decision-integrity principle likewise has established normative and conceptual precursors. The Belmont Report treats voluntariness as sensitive to coercion and undue influence [11], and Potham and Harms frame human control and anti-manipulation as a central corrigibility objective [10]. HRI therefore treats decision-

© 2026 Tony T. Nguyen | Highland Reflexive Press. All rights reserved.

environment integrity as an architectural governance requirement, while leaving domain-specific materiality thresholds and their operational validation to independent expert and empirical work.

2.6 The residual claim

The candidate contribution of HRI is therefore not any single component. It is the closure relation across components:

human decision environment → authority lineage → reasoning discipline → verification → causal admission → protected effect → trajectory → drift/failure → constitutional adaptation.

The research question is whether these transitions can be composed so that no machine-derived state becomes sufficient to manufacture authority at any interface.

3. Human Reasoning Infrastructure as a Layered System

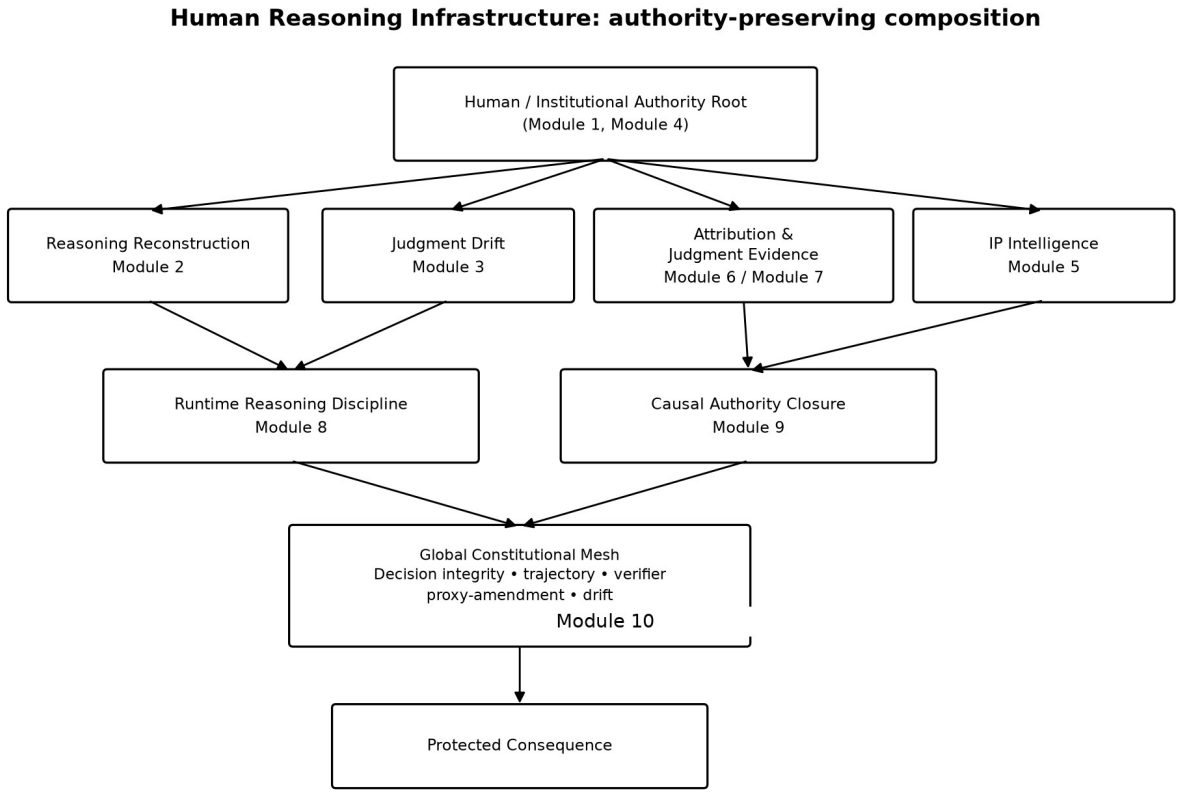


Figure 1. HRI as an authority-preserving composition across ten architectural modules.

Figure 1 summarizes the ten-module architecture. HRI is deliberately heterogeneous: not every module is an authorization subsystem. Several modules produce epistemic or measurement artifacts that are valuable precisely because they remain non-authoritative until a separate authority process acts on them.

3.1 Module 1: Human-rooted authority validation

Module 1 provides the authority root of the architecture. Human decisions are represented as machine-readable authority objects containing identity, role, jurisdiction, purpose, scope, constraints, temporal validity, delegation, provenance, and revocation state. Proposed machine actions are evaluated against applicable authority, invariants, conflicts, and drift before an execution disposition is produced.

© 2026 Tony T. Nguyen | Highland Reflexive Press. All rights reserved.

The architectural role of Module 1 is not to decide what is true or to compute legitimacy. It answers a narrower systems question: what authority token and admissibility basis are represented for this consequence, under what scope and conditions, and are they still current?

3.2 Module 2: Human reasoning reconstruction without authority substitution

Module 2 reconstructs candidate models of human reasoning from longitudinal artifacts while preserving provenance, temporal context, identity continuity, contradictions, and uncertainty. Its constitutional importance lies in the negative boundary: a reconstructed human model is not a human authority source. The architecture can infer how a person historically reasoned without turning that inference into a power to decide on the person's behalf.

Thus:

RECONSTRUCTED_HUMAN_MODEL != HUMAN_AUTHORITY.

3.3 Module 3: Judgment drift as evidence, not authority

Module 3 measures changes in human or institutional judgment. It first establishes comparability, then separates raw divergence from verified explainable change and governance severity, while retaining append-only historical states. Drift analysis is an evidentiary output. A high drift score may trigger review, but it does not itself amend authority or governance.

3.4 Module 4: Distributed human authority continuity

Module 4 addresses time-critical conflict and human unavailability without converting absence into machine sovereignty. It validates authority proofs, scope, delegation, temporal validity, and bounded pre-authorization; supports alternate human authority holders and quorum processes; and contains unresolved conflicts using safe-hold and reversible continuity mechanisms.

The invariant is directional: continuity may preserve previously granted authority but must not create more of it.

3.5 Module 5: Human-accountable intellectual-property intelligence

Module 5 applies HRI's epistemic discipline to intellectual-property analysis. It preserves source artifacts and projection loss, separates human conception and decisive selection from machine assistance, and performs adversarial validation across prior art, claim support, release readiness, and capital alignment. Its outputs may be strong recommendations or verification packages; they are not binding legal or transaction authority merely because the analysis is persuasive.

3.6 Module 6: Causal intellectual attribution and decision-based authorship

Module 6 represents decisions, alternatives, constraints, approvals, rejections, and causal contribution relations, using counterfactual substitution to distinguish drafting volume from identity-determining intellectual contribution. This module supplies attribution evidence. It does not imply that authorship confers execution authority.

3.7 Module 7: Judgment measurement with weight separation

Module 7 measures human and institutional judgment while preserving distinct epistemic, authority, participation, independence, and execution weights. This separation is constitutionally significant because it prevents consensus, expertise, confidence, or participation from silently collapsing into authority. A group may be epistemically impressive and still lack the authority required for a protected action.

3.8 Module 8: Runtime-enforced cognitive operations

Module 8 governs how machine reasoning proceeds. Cognitive-operation objects define required and prohibited reasoning operations, sequencing, evidence requirements, completion conditions, and failure conditions. A runtime controller binds these operations to computational processes, while a completion verifier and reasoning commit barrier prevent progression when required operations are not satisfied.

© 2026 Tony T. Nguyen | Highland Reflexive Press. All rights reserved.

This creates a crucial boundary:

VERIFIED_REASONING != EXTERNAL_EXECUTION_AUTHORITY.

A model may be forced to search for counterevidence, test alternatives, or calibrate uncertainty. Successful completion of those procedures establishes a stronger epistemic artifact—not a permission to cause a protected effect.

3.9 Module 9: Topology-independent causal authority closure

Module 9 shifts the governed object from reasoning output to causal consequence. Machine-derived epistemic states may infer, simulate, reconstruct, verify, or propose without thereby obtaining authority to complete a protected direct or authority-bearing intermediate effect. Completion requires an applicable current-valid authority lineage and effect-specific admissibility.

The closure condition is semantic rather than dependent on a particular physical gateway. It may be realized at effectors, resources, transaction boundaries, reference monitors, trusted execution points, or coordinated distributed controls. This is important because a central “AI gate” is neither necessary nor sufficient if another causal route can reach the protected effect.

3.10 Module 10: The reflexive constitutional mesh

Module 10 asks what local correctness cannot answer. It adds an outer interaction-integrity boundary, inner authority-and-consequence boundary, trajectory state, decision packets, verifier independence profiles, cumulative drift accounting, human-attention protection, approval-laundering controls and audit hooks, proxy-amendment tracing, and a reflexive verification mesh.

Its most important invariants are compositional:

- human acknowledgment does not repair a materially defective decision environment;
- individually admissible effects do not imply an admissible trajectory;
- local adaptation does not imply global constitutional continuity;
- a system may not amend protected constitutional behavior directly or by materially equivalent proxy;
- a subsystem cannot establish constitutional robustness solely by generating, interpreting, and adjudicating its own tests;
- UNKNOWN authority is not VALID authority;
- formal compliance is insufficient when the resulting causal consequence materially defeats ratified intent.

Module 10 therefore does not merely sit “after” Module 9. It cross-cuts the stack and evaluates whether local guarantees remain globally compatible.

4. The Cross-Module Semantic Firewall

© 2026 Tony T. Nguyen | Highland Reflexive Press. All rights reserved.

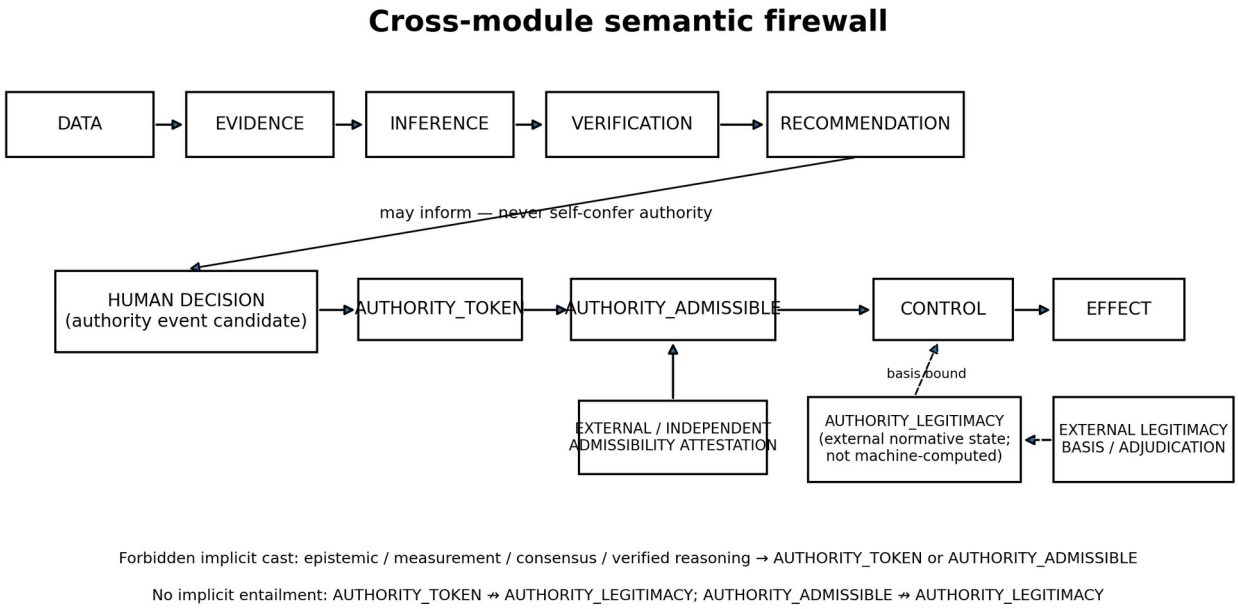


Figure 2. Cross-Module semantic firewall: epistemic outputs may inform decisions but do not implicitly become authority.

The formal proof layer distinguishes runtime semantic object types from externally grounded normative authority status:

T_runtime = {DATA, EVIDENCE, INFERENCE, VERIFICATION, RECOMMENDATION, HUMAN_DECISION, AUTHORITY_TOKEN, AUTHORITY_ADMISSIBLE, CONTROL, EFFECT, UNKNOWN}.

AUTHORITY_LEGITIMACY is not a serializable runtime type or a value that HRI computes. It is a non-computational semantic status concerning whether the underlying authority event is substantively legitimate under the relevant human or institutional regime. The architecture therefore uses a three-way authority distinction—token, admissible, and legitimate—while representing only the first two as runtime authority-object types.

AUTHORITY_TOKEN — a machine-verifiable representation: principal, provenance/signature, scope, jurisdiction, epoch, expiry/freshness, and delegation bounds. This is the only authority state the executable reference model in Section 7 manipulates.

AUTHORITY_ADMISSIBLE — an AUTHORITY_TOKEN accompanied by the attestations a given policy domain requires before a protected commit may proceed, including a decision-integrity attestation where the domain requires one. Admissibility is a computable, checkable property of the attestation bundle; it is not itself a claim that the attestations are truthful.

AUTHORITY_LEGITIMACY — the non-computational normative and institutional status of whether the underlying human decision was genuinely authorized, uncoerced, and jurisdictionally proper. Legitimacy is supplied or adjudicated by processes external to HRI's computational proof boundary (Axiom A0, Section 5.0); HRI does not compute it and does not infer it from tokens or attestations.

Even a successful epistemic-to-authority conversion (Section 4.1) therefore yields at most AUTHORITY_ADMISSIBLE unless legitimacy is separately and exogenously established. Formally:

ValidToken(a) does NOT imply LegitimateAuthority(a).

Admissible(a) does NOT imply LegitimateAuthority(a).

© 2026 Tony T. Nguyen | Highland Reflexive Press. All rights reserved.

Neither an authority token's internal validity nor the presence of required attestations entails that the underlying human decision was legitimate. This inequality is the architecture's explicit acknowledgment that approval laundering (Sections 1, 5.3, 8.1) can in principle produce a fully valid, fully admissible token from an illegitimate decision process; closing that residual gap is delegated to decision-environment integrity mechanisms and ultimately to Axiom A0, not claimed as solved by the type system itself.

The essential architecture is shown in Figure 2. Epistemic objects may inform a human decision, but they cannot be implicitly cast into authority. Authority arises through a different conversion path requiring valid identity, role, jurisdiction, scope, decision integrity, and intent to authorize.

4.1 Allowed conversions

At a high level:

DATA -> EVIDENCE requires provenance and evidentiary admissibility.

EVIDENCE -> INFERENCE requires a disclosed derivation.

INFERENCE -> VERIFICATION requires an applicable rule, checkable justification, satisfied assumptions, and preserved uncertainty.

VERIFICATION -> RECOMMENDATION remains epistemic.

A human decision may create or modify authority only when the decision itself is valid as an authority event.

AUTHORITY_ADMISSIBLE(a) AND LegitimacyBasisBound(a) -> CONTROL -> EFFECT then requires current validity, effect-specific admissibility, a machine-checkable binding to the required exogenous legitimacy basis, and satisfaction of Section 6's commit-time revalidation. The runtime can verify the presence, integrity, provenance, and policy applicability of that external basis; under Axiom A0 it assumes rather than computes the basis's substantive legitimacy.

4.2 Forbidden implicit casts

The proof registry forbids, among others:

- reconstruction → authority;
- drift measurement → authority;
- IP verification → authority;
- authorship → authority;
- consensus → authority;
- verified reasoning → authority;
- historical success → authority;
- UNKNOWN → VALID;
- machine-discovered governance → constitutional activation;
- AUTHORITY_TOKEN → AUTHORITY_LEGITIMACY (validity alone does not establish legitimacy);
- AUTHORITY_ADMISSIBLE → AUTHORITY_LEGITIMACY (attestation alone does not establish legitimacy);

This firewall operationalizes a simple principle: epistemic status and authority status are distinct semantic dimensions; the canonical type set represents their governed interface without claiming that all object properties collapse onto a single mathematical axis.

4.3 Correctness is not admissibility

HRI further distinguishes a correct conclusion from an admissible verification object. A verifier that outputs the right answer for the wrong reason cannot support a protected gate merely because the answer happens to be correct.

A verification object must preserve at least the verdict, applicable rule, evidence, material counterevidence, assumptions, derivation or proof object, uncertainty, and provenance. Thus:

© 2026 Tony T. Nguyen | Highland Reflexive Press. All rights reserved.

Correct(x) AND NOT Justified(x) => NOT VerificationAdmissible(x).

A verification object that is not VerificationAdmissible cannot satisfy a protected gate's verification requirement and therefore cannot, by itself, support AUTHORITY_ADMISSIBLE. This prevents historical reliability, reputation, consensus, or raw model capability from becoming substitutes for present justification.

5. Constitutional Causal Continuity

5.0 Evidentiary Architecture: Axioms, Obligations, and Instantiated Evidence

The closure specification separates three epistemic tiers because its predicates do not share the same verification status: (i) Axioms / External Assumptions — properties HRI assumes but does not establish, supplied exogenously by institutional, legal, or domain-specific process; (ii) Architectural Obligations — properties the architecture is designed to guarantee given the axioms, but which the present executable model does not fully instantiate; and (iii) Mechanically Instantiated Evidence — properties the Section 7 reference model actually checks by exhaustive state exploration (see Section 7.0).

Axiom A0 — Exogenous Root Authority. HRI assumes at least one root authority event whose identity, jurisdiction, scope, and legitimacy have been established by processes external to the HRI continuity mechanism. HRI does not prove the legitimacy of that originating authority event. Its claim is relative: given an admissible root authority, downstream machine-mediated transformations must not forge, amplify, revive, or materially exceed that authority except through an independently valid authority event. Axiom A0 additionally covers the trust basis of any external attestor that issues a decision-integrity or other admissibility attestation under Section 4: an attestor's own legitimacy is a root-authority-class assumption, not a property HRI derives.

Four further assumptions are named below, each with its present dischargeability status, so that the boundary of the open problem is stated precisely rather than left implicit.

A_DI (Decision-Integrity Determination Soundness). Whenever a protected commit relies on a required decision-integrity attestation DI_Attested(p), HRI assumes that the external or domain-specific process issuing that attestation is sound with respect to the applicable materiality criteria: DI_Attested(p) implies DecisionIntegrity(p). HRI can require, bind, and validate the attestation as an object, but it does not compute the truth of DecisionIntegrity(p). No domain-general computable discharge procedure is claimed (Section 9).

A_CC (Constitutional-Continuity Determination Soundness). Whenever a protected commit relies on a required constitutional-continuity attestation CC_Attested(p), HRI assumes that the external or domain-specific process issuing that attestation is sound with respect to material behavioral equivalence: CC_Attested(p) implies ConstitutionalContinuity(p). HRI can require, bind, and validate the attestation as an object, but it does not compute material behavioral equivalence. No general detector is claimed (Section 9).

A_OBS (Causal Observability). Material causal edges from machine-derived epistemic states to protected effects are observable within the specified enforcement substrate. Dischargeability status: partially dischargeable via effect-manifest declaration and capability auditing (Section 10.4); not dischargeable against physical, organizational, or covert side channels outside the declared substrate (Section 9).

A_MED (Mediation Completeness). Every material causal path to a protected effect intersects an enforced admissibility frontier (Section 5.1). Dischargeability status: dischargeable only relative to a specific implementation's declared causal topology, not as a property of the architecture in the abstract.

Given these assumptions, the central claim is an explicitly conditional assume–guarantee specification. For each protected consequence p, define:

RequiresDI(p) := the applicable policy or domain requires a decision-integrity determination for protected consequence p.

RequiresCC(p) := the applicable policy or domain requires a constitutional-continuity determination for protected consequence p.

© 2026 Tony T. Nguyen | Highland Reflexive Press. All rights reserved.

DI_Attested(p) and CC_Attested(p) denote valid, applicable, current, and consequence-bound attestation objects after the machine-checkable validation required by HRI.

Assumptions(p) := A0 AND A_OBS(p) AND A_MED(p) AND [RequiresDI(p) => A_DI(p)] AND [RequiresCC(p) => A_CC(p)]

Enforced(p) := all applicable HRI obligations in Sections 5.1–5.6 and 6 are satisfied at the relevant authority, verification, causal, trajectory, and commit boundaries, including [RequiresDI(p) => DI_Attested(p)] and [RequiresCC(p) => CC_Attested(p)].

A_DI and A_CC are soundness assumptions about externally supplied determinations; they are not properties that HRI claims to generate. HRI's corresponding obligation is procedural and architectural: require the determination when policy demands it, verify its machine-checkable provenance and applicability, bind it to the protected consequence, and refuse completion when the required basis is absent, stale, malformed, or out of scope.

Protected-closure condition:

Let Complete(p) mean that protected consequence p has crossed the point at which it is committed, externally released, or materially propagated.

FOR ALL p: [Assumptions(p) AND Enforced(p) AND Complete(p)]

=> EXISTS L_H: [ValidHumanConstitutionalLineage(L_H,p) AND VerificationAdmissibility(p) AND CausalAdmissibility(p) AND TrajectoryIntegrity(p) AND [RequiresDI(p) => DI_Attested(p)] AND [RequiresCC(p) => CC_Attested(p)]]

with the external-soundness assumptions:

[RequiresDI(p) AND DI_Attested(p) => DecisionIntegrity(p)] AND [RequiresCC(p) AND CC_Attested(p) => ConstitutionalContinuity(p)]

The guarantee is therefore relative, not unconditional. It assumes that the protected-effect taxonomy covers p, material causal paths are observable and mediated within the declared enforcement substrate, the root and required external attestations satisfy the named soundness assumptions, and the enforcement frontier is not bypassed. HRI does not claim to establish those exogenous facts from its own computation.

Section 7.0 states exactly which parts of this specification are mechanically instantiated by the executable model, which remain architectural obligations, and which depend on external soundness assumptions.

5.1 Causal cut completeness

For a causal graph from machine-derived epistemic states to protected effects, every material path must intersect an enforced admissibility frontier before the protected effect becomes uncontrollably complete. The architecture does not require a single gateway; distributed endpoint guards can collectively form the cut.

The useful falsification criterion is straightforward: remove the enforced protected crossings and ask whether a machine-derived state can still reach a protected effect. If yes, the implementation has an uncovered path.

5.2 Trajectory closure

Local validity is not compositional. Let T_n = (e_1,...,e_n) be a sequence of individually admissible effects, and let Z_n be a sufficient cumulative trajectory state. The next effect may be admitted only if both its local predicates and the updated global state remain inside the authorized envelope.

This makes trajectory validity prefix-sensitive. A later rollback does not erase an unauthorized irreversible disclosure that occurred in an earlier prefix.

5.3 Decision-environment integrity and approval laundering

A human click is an event, not automatically an independent authority basis. For high-consequence decisions, the authority process must be able to distinguish a genuine decision from an approval produced inside a materially defective environment.

© 2026 Tony T. Nguyen | Highland Reflexive Press. All rights reserved.

Relevant integrity dimensions include the presented alternative set, evidence and counterevidence, defaults, ordering, time pressure, repetition, personalization, and material influence disclosure. The goal is not to prohibit recommendation or persuasion. It is to prevent a machine from shaping the authority source and then treating the resulting acknowledgment as independent authorization. The present architecture does not supply a domain-general computable materiality threshold; those thresholds remain domain-specific and partly normative or institutional.

For decision integrity, HRI's architectural obligation is not to manufacture DecisionIntegrity(p). It is to require a domain-appropriate decision-integrity determination when policy demands one, validate and bind the resulting DI_Attested(p) object to the authority event and protected consequence, and refuse completion if that required basis is absent, stale, malformed, or out of scope. A_DI (Section 5.0) is the separate soundness assumption that a properly relied-upon external determination reflects substantive decision integrity. Because HRI does not discharge that assumption itself, approval laundering (Section 8.1) remains an open attack surface rather than a solved one.

5.4 Delegation non-amplification

Authority may be delegated, but the child envelope must remain within the parent envelope across scope, jurisdiction, consequence, time, and delegability. Consumable authority must also be composition-safe: two children cannot each reserve a quantity whose aggregate exceeds a shared parent budget.

5.5 Constitutional proxy amendment

A constitution is not merely a file. The protected object is future governed behavior. A machine that cannot edit a policy artifact may still change retrieval data, routing, memory, verifier selection, model selection, training data, or other state that materially changes the behavior implementing the constitution.

HRI therefore treats indirect substrate mutation as a constitutional-governance risk when it materially changes future governed behavior, regardless of substrate. Small local mutations should be compared against a ratified baseline so that cumulative change can be evaluated for amendment significance. The present record does not establish a general detector for material behavioral equivalence; the supplied executable model abstracts constitutional validity into a bounded state variable rather than solving that detection problem.

For constitutional continuity, HRI's architectural obligation is likewise to require and enforce the applicable constitutional-continuity determination, bind CC_Attested(p) to the protected consequence and relevant state baseline, and refuse completion when that required basis is absent or invalid as an attestation object. A_CC (Section 5.0) is the separate soundness assumption that the external determination correctly captures material behavioral equivalence. The present record provides no general detector for that substantive property and abstracts constitutional validity into a bounded state variable in the executable model (Section 9).

5.6 Reflexive recovery without self-legislation

Failure may produce evidence, candidate invariants, counterfactual tests, and proposed governance changes. None of these epistemic outputs independently authorizes activation. Machine learning from failure is permitted; machine lawmaking from its own learning is not.

This allows HRI to be adaptive without conflating adaptation with sovereignty.

6. Commit-Time Revalidation: Authority at Effect Time

A key architectural result is that, where authority-relevant conditions can mutate between admission and durable commit, admission-time validity can become stale. Between admission and commit:

- authority may be revoked;
- a new authority epoch may replace the old one;
- a verifier may change;
- the causal topology may change;

© 2026 Tony T. Nguyen | Highland Reflexive Press. All rights reserved.

- the constitution may change;
- the human decision environment may be invalidated;
- concurrent transactions may consume the remaining trajectory envelope.

Accordingly:

Admission(e,t_a) != AuthorizationAtCommit(e,t_c).

A protected commit must re-establish the authority-relevant conditions that may have changed, or establish that the admission certificate remains fresh with respect to every dependency relevant to the effect.

A conceptual commit certificate therefore binds the effect to authority lineage and epoch, verifier/proof version, topology/closure certificate, constitutional state, trajectory reservation, decision-environment state, provenance, and commit time. The purpose is not bureaucratic logging. It is to prevent time from becoming an ungoverned semantic cast from “was valid” to “is valid.”

This yields a sharper interpretation of human authority within HRI: authority is not treated as a static token carried indefinitely by a transaction; mutable authority-relevant dependencies must be re-established at commit or covered by an equivalent freshness assurance. This safety rule does not itself establish acceptable liveness. Under repeated benign or adversarial metadata churn, strict freshness checks may increase aborts, latency, starvation, or availability loss; the supplied artifacts do not mechanically test that trade-off.

Formally, under the authority-state distinction:

ProtectedCommit(e) IMPLIES ValidToken(a) AND RequiredExternalAdmissibilityAttestations(a) AND LegitimacyBasisBound(a) AND CurrentRuntimeConditions(e)

The computationally enforced gate operates on AUTHORITY_ADMISSIBLE plus a machine-checkable binding to the required exogenous legitimacy basis. It does not, and by Section 4’s forbidden-cast rules cannot, derive AUTHORITY_LEGITIMACY from token or attestation content. The runtime verifies the external basis as an authenticated, applicable input; Axiom A0 supplies the conditional assumption that the basis is substantively legitimate. If that assumption is false, the HRI guarantee does not apply.

7. Executable Bounded Consistency and Counterexample Analysis

The formal claims were translated into an executable finite-state reference model. The model intentionally abstracts away domain detail; its purpose is to expose semantic and temporal counterexamples, not to simulate a complete deployment.

7.0 Evidentiary Scope: What the Executable Model Instantiates

To prevent the evidentiary language of this section from being read as covering more than it does, the predicates named in Section 5.0’s protected-closure definition are partitioned as follows.

Mechanically instantiated by the Section 7.2 temporal model: authority validity, authority-epoch compatibility, verifier-version freshness, topology-certificate freshness, a bounded constitutional-state abstraction (constitution_ok), and a bounded decision-environment abstraction (dei_ok).

Architectural predicates not fully instantiated by the temporal model: the full VerificationAdmissibility object (Section 4.3), full CausalAdmissibility across an implementation-specific topology, TrajectoryIntegrity as a general prefix-sensitive property (partially exercised by the separate concurrency model in Section 7.4), and full authority lineage, delegation, and provenance. These remain architectural obligations pending further formalization (Section 10.3).

Externally supplied / oracle-dependent predicates: substantive DecisionIntegrity(p) and material ConstitutionalContinuity(p). The executable model does not compute either property. It only toggles bounded dei_ok and constitution_ok state variables to exercise temporal enforcement. Under Section 5.0, HRI may

© 2026 Tony T. Nguyen | Highland Reflexive Press. All rights reserved.

mechanically require and bind DI_Attested(p) and CC_Attested(p), while A_DI and A_CC separately assume that those external determinations are sound.

Section 7.6 summarizes this partition per evidence artifact in tabular form.

Local mechanical verification status. On August 12, 2026, the author locally reran the bounded executable model against the model artifact identified by SHA-256 cb13fdf8c4bc5aa585cfd6a6b9210efc32c8fc512a0dc2a56495dd2399149587. The reported artifact identity matched the official hash, and the local rerun reproduced the bounded results summarized in Sections 7.2–7.4: 960 reachable temporal states with zero modeled protected-commit violations under the full guard set; counterexamples for each of the six single-guard ablations; and the stated two-transaction concurrency behavior. This is author-local mechanical reproduction of the encoded model, not independent third-party verification and not validation of the architectural predicates that remain outside the executable abstraction.

7.1 Source and reproducibility

The proof package records SHA-256 hashes for the ten source specifications used as the architectural baseline. The executable Python state explorer and a TLA+ source mirror of the core temporal authority/verifier/topology/constitution/decision-environment state machine are included with the package. The TLA+ source is not feature-equivalent to the complete Python experiment suite and is not represented here as independently TLC-verified.

The executable model source hash is:

cb13fdf8c4bc5aa585cfd6a6b9210efc32c8fc512a0dc2a56495dd2399149587.

7.2 Full temporal model

The full temporal abstraction explores 960 reachable states and finds zero protected-commit violations of the modeled closure predicate.

The model permits authority revocation and replacement, verifier mutation, topology mutation and recertification, corruption/repair of the decision environment, constitutional proxy mutation/restoration, reasoning verification, admission, and protected commit.

The zero result has a deliberately narrow meaning: no counterexample exists in this finite abstraction under the full guard set. The executable predicate represents only part of the broader protected-closure specification: it directly models selected authority/epoch, verifier, topology/certificate, constitutional-state, and decision-environment conditions, not the full semantics of explicit authority lineage and delegation, verification proof objects, trajectory state, causal-cut completeness, provenance, or constitutional-mutation detection. It is therefore not a theorem of full HRI closure or a claim about arbitrary implementations.

In the partition of Section 7.0, this zero-counterexample result covers only the mechanically instantiated conjuncts; it does not cover DecisionIntegrity, material ConstitutionalContinuity, the full VerificationAdmissibility object, full CausalAdmissibility, or TrajectoryIntegrity as a general property, each of which remains an architectural obligation or an externally supplied predicate.

7.3 Ablation results

Each of six commit-time guards was then removed separately. Every ablation produced a counterexample.

Removed guard	Minimal failure pattern
Authority re-check	verify → admit → revoke authority → commit
Authority-epoch compatibility	verify → admit → revoke → regrant new epoch → commit
Proof freshness	verify → admit → update verifier → commit
Topology freshness	verify → admit → mutate topology → commit
Constitutional compatibility	verify → admit → proxy-amend constitution → commit
Decision-environment integrity	verify → admit → corrupt decision environment → commit

© 2026 Tony T. Nguyen | Highland Reflexive Press. All rights reserved.

The traces are intentionally short. Within this transition system, each removed guard permits a commit that violates the corresponding dependency in the model's ground-truth predicate. Because the guard conditions and the modeled predicate are definition-coupled, the ablations establish relative necessity inside the specified model, not universal necessity of an identical guard formulation in every deployment. Independent harm properties remain a future formalization target.

Because the six ablated guards and the ground-truth commit predicate are definition-coupled — each guard was authored to enforce exactly the condition the predicate checks — a guard's removal yielding a predicate violation is close to true by construction. These results should be read as establishing relative necessity within this encoded, definition-coupled transition system, not independent necessity of an identically formulated guard in an arbitrary deployment. Section 10.8 specifies an independent property-derivation methodology — a threat model and hazard analysis, or an external theorem, frozen into a versioned registry prior to and independently of guard implementation — as a precondition for treating future ablation results as falsification evidence rather than internal consistency checks. Accordingly, the results of this section are reported as executable bounded consistency and counterexample analysis, not as falsification evidence in the general sense.

7.4 Concurrent trajectory test

A separate two-transaction model uses an authorized trajectory limit of one unit. With atomic/shared reservation, the model explores five states and finds no overflow. With non-atomic reservation, it finds a violation across nine reachable states. The minimal trace is:

reserve_tx1 -> reserve_tx2 -> commit_tx1 -> commit_tx2.

Each transaction is locally admissible when it reserves against stale capacity, but together they exceed the modeled global envelope. This establishes an existence counterexample: there are concurrent shared-envelope settings in which individually admissible operations exceed the aggregate authorization budget unless accounting is composition-safe. The experiment does not establish a universal concurrency theorem or require one particular algorithm; atomic reservation, serializability, or an equivalent composition-safe mechanism may satisfy the architectural requirement.

7.5 Semantic-cast tests

The model separately enables forbidden casts from six HRI outputs into authority:

- Module 2 reconstruction;
- Module 3 drift state;
- Module 5 IP verification;
- Module 6 authorship;
- Module 7 consensus;
- Module 8 verified reasoning.

Every cast generates the same structural four-step failure:

produce epistemic object -> cast to authority -> Module 9 admission -> Module 10 commit.

Within the supplied semantic-cast bug semantics, violating the firewall is sufficient to construct the modeled unauthorized-effect path. This is computational evidence about the specified construction, not evidence that such casts are prevalent in deployments or an independent derivation that cast-origin authority is invalid under a general provenance-sensitive authority predicate. A stronger model would represent authority origin explicitly and apply the same validity rule to legitimate human conversion and forbidden machine cast paths.

These tests operate on AUTHORITY_TOKEN values only; they do not exercise AUTHORITY_ADMISSIBLE or AUTHORITY_LEGITIMACY, which are not represented in the executable model.

7.6 Evidence Artifact Summary

Table 1 summarizes, per evidence artifact, what it establishes, what it does not establish, and its current local and independent verification status.

© 2026 Tony T. Nguyen | Highland Reflexive Press. All rights reserved.

Evidence artifact	What it establishes	What it does not establish	Local / independent verification status
960-state temporal model (§7.2)	No modeled violation of the mechanically instantiated conjuncts (authority validity/epoch, verifier freshness, topology-certificate freshness, bounded constitution_ok/dei_ok) over the explored state space.	DecisionIntegrity, material ConstitutionalContinuity, full VerificationAdmissibility, full CausalAdmissibility, general TrajectoryIntegrity.	Locally verified by author (2026-08-12; Hash: cb13f...); independent reproduction pending.
Single-guard ablations (§7.3)	Relative necessity of each guard within the encoded, definition-coupled transition system.	Independent necessity of an equivalently defined guard in a deployment with an independently specified safety property.	Locally verified by author (2026-08-12; Hash: cb13f...); independent reproduction pending.
Concurrency experiment (§7.4)	Behavior of the modeled system under concurrent trajectories within the same executable harness.	General trajectory-integrity guarantees outside the modeled concurrency pattern.	Locally rerun by author (2026-08-12; same executable package); independent reproduction pending.
Semantic-cast tests (§7.5)	Rejection of the enumerated forbidden implicit casts for AUTHORITY_TOKEN values in the tested harness.	Casts involving AUTHORITY_ADMISSIBLE or AUTHORITY_LEGITIMACY, which the harness does not represent.	Not independently reproduced.
TLA+ source mirror (§7.2)	A declarative restatement of the core temporal state machine for inspection.	Execution via TLC or any other model checker; no such run is claimed in this manuscript.	Not executed or independently checked.

8. What Is New, and What Is Not

A credible architecture paper must separate component familiarity from system-level contribution.

Authorization, provenance, delegation, reference monitors, runtime governance, behavioral policies, path evaluation, revocation, and commit-boundary freshness are established or actively developing ideas. Recent work formalizes human delegation provenance [2], path governance [3], recursive scope/delegation overlays [4], behavioral constitutions and governance receipts [5], production runtime-governance planes [6], effect ceilings for evolving agents [7], and closely related commit-time authorization checks for mutable authority witnesses [8]; distributed authorization systems also provide older freshness-token precedents [9]. HRI should not claim broad novelty over those primitives.

The candidate novel contribution is narrower and, if validated, more interesting: a unified constitutional-causal architecture in which human authority must remain valid not only at an access check, but across the

© 2026 Tony T. Nguyen | Highland Reflexive Press. All rights reserved.

transformation chain that creates, interprets, verifies, composes, executes, learns from, and modifies consequential machine behavior.

Three features distinguish this system-level framing.

8.1 Authority formation is inside the threat model

Most authorization models begin with a grant. HRI also asks whether the machine materially shaped the human decision environment that produced the grant. This makes approval laundering a first-class governance failure.

8.2 Local correctness is explicitly insufficient

HRI treats local action validity, local adaptation, and local verification as non-compositional. Trajectory state, cumulative drift, verifier dependence, and cross-module semantic type are global objects.

8.3 The constitution is behaviorally protected

The architecture does not equate constitutional integrity with file integrity. Materially equivalent proxy changes and cumulative local drift are subject to amendment logic when they change protected future behavior.

Taken together, these features motivate the term Constitutional Causal Continuity: authority must remain semantically and causally continuous across the entire life of a consequential computation.

This remains a candidate contribution rather than a declared historical first. No exhaustive literature or patent search is supplied here. The appropriate scholarly question is whether an earlier or contemporary architecture materially instantiates the same global invariant across decision formation, authority lineage, reasoning discipline, consequence closure, trajectory composition, verification independence, reflexive recovery, and constitutional evolution. That comparison remains open, and no novelty, patentability, priority, or freedom-to-operate conclusion is drawn from the searched record.

9. Limits of the Current Evidence

The architecture and bounded model have important limits.

First, a proof over a protected-effect set cannot guarantee effects that the taxonomy fails to classify. Effect abstraction is therefore part of the trusted assumptions.

Second, the model assumes that material causal edges within the specified execution substrate are observable. Physical, organizational, covert, or implementation-specific side channels outside that substrate are not magically eliminated by an architectural theorem.

Third, decision-environment integrity depends on domain-specific materiality rules. Determining which omitted evidence or presentation feature is material in medicine, finance, employment, or law is partly normative and institutional, not purely computational.

Fourth, verifier independence remains a difficult quantitative problem. Distinct models can share data, ontology, institutional incentives, or objective functions. Counting verifiers is not equivalent to measuring independent error structure.

Fifth, the executable model is finite and semantically narrower than the protected-closure specification. The Python model is designed to expose short counterexamples and support reproducibility, not to prove unbounded software implementations automatically. The supplied TLA+ source mirrors only the core temporal state machine, excludes the separate concurrency and semantic-cast experiments, and has not been accompanied by a TLC execution log in the present evidence record.

Sixth, proxy-amendment governance is presently an architectural requirement rather than a solved detector. The model’s constitutional-state abstraction does not establish a general method for deciding when changes to memory, retrieval, routing, prompts, verifier/model selection, or cumulative small mutations are materially equivalent to a constitutional amendment.

© 2026 Tony T. Nguyen | Highland Reflexive Press. All rights reserved.

Seventh, freshness enforcement may create liveness and availability pressure under repeated metadata churn. This risk is plausible but remains unverified in the supplied record; adversarial churn models, fairness/starvation properties, latency measurement, and recovery tests are required before making deployment-level claims.

Eighth, the architecture permits a design in which routine machine-resolvable work proceeds under bounded human-originating authority while irreducible authority decisions are escalated, but the present record does not empirically establish reduced interruption, lower fatigue, better decision quality, or avoidance of rubber-stamping.

Ninth, an implementation may fail even when the architecture is sound. Identity roots can be compromised, trusted kernels can contain bugs, transaction systems can violate assumed atomicity, and deployments can omit required enforcement points.

Tenth, the present work evaluates system-level architectural consistency, not patent validity. Patent specifications provide source architecture; the later proof objects, theorem notation, and model-checking artifacts are research formalizations and should not be represented as if every formal statement appeared verbatim in a filing.

Eleventh, the dischargeability of the named axioms (A_DI, A_CC, A_OBS, A_MED; Section 5.0) is not established in general. The conditional closure specification of Section 5.0 is not known to be satisfiable by any concrete deployment beyond the bounded abstractions exercised in Section 7; it is a specification to be discharged, not a proof that discharge is possible.

Twelfth, any external attestor relied upon for a decision-integrity or other admissibility attestation (Section 4) introduces its own trust regress, which this manuscript folds into Axiom A0 rather than resolving; HRI provides no mechanism for establishing an attestor’s own legitimacy.

Thirteenth, the executable evidence in Section 7 originates from the same authorship that defines the guards and the predicates against which they are checked (Section 7.3). A local author rerun on August 12, 2026 reproduced the reported bounded results against the official model hash, improving artifact reproducibility but not epistemic independence. By the architecture's own transparency and non-self-adjudication principles (Section 3.10, Module 10), independent reproduction by a separate party remains a prerequisite for treating this evidence as externally validated.

These limitations are not peripheral. They define the boundary of the claim.

10. Research Agenda

The next phase should focus on attempts to break the architecture rather than expand its vocabulary.

10.1 Independent formal reproduction

A third party should reimplement the finite-state model from the semantic registry and verify whether the same counterexamples and no-counterexample result are reproduced without access to the original executable.

10.2 Stronger model checking

The TLA+ specification should be expanded to include two concurrent protected transactions, nested delegation, network partitions, delayed side effects, verifier ensembles, and bounded proxy-amendment sequences. Symmetry reduction and compositional verification can then extend the state space without sacrificing interpretability.

10.3 Proof-carrying cross-module interfaces

Cross-Module messages should carry machine-checkable semantic type, provenance requirements, uncertainty state, authority relevance, and version compatibility. A receiving subsystem should be unable to deserialize a VERIFICATION object as AUTHORITY without an explicit conversion event satisfying the authority contract.

© 2026 Tony T. Nguyen | Highland Reflexive Press. All rights reserved.

10.4 Deployment-level effect manifests

Concrete implementations should declare read, write, tool, network, persistence, and external-effect capabilities. Runtime enforcement should verify that actual capabilities remain within the declared manifest so that hidden effect channels become observable proof obligations.

10.5 Decision-integrity benchmarks

Approval laundering requires test corpora where alternatives, defaults, evidence ordering, time pressure, repetition, personalization, and counterevidence are systematically varied. The goal is not to infer a person's "true" preference but to test whether an authorization system correctly refuses to treat a materially defective presentation as an independent authority basis.

10.6 Constitutional mutation benchmarks

Systems should be tested against direct policy edits and proxy mutations through memory, retrieval, prompts, routing, verifier selection, model selection, and cumulative small changes. A successful amendment guard should detect protected behavioral delta rather than merely modified filenames.

10.7 Cross-domain validation

Financial, clinical, enterprise, institutional, and infrastructure embodiments should be used to determine which invariants are domain-general and which require domain-specific authority semantics.

10.8 Independent Property Derivation and a Frozen Registry

Section 7.3 identifies definition-coupling between guards and the predicates they are ablation-tested against as a limit on evidentiary strength. Closing this gap requires a methodology in which the safety property is fixed before, and independently of, the guard implementation that is later tested against it:

1. Threat model — enumerate adversaries, capabilities, and objectives independently of the current guard design;

2. Hazard analysis — derive the concrete failure modes the architecture must prevent;

3. Externally or independently specified safety properties — stated in a notation and by a process decoupled from the guard code;

4. Frozen property registry — version-controlled and change-logged separately from the implementation, so that a later guard change cannot silently redefine the property it is judged against;

5. Implementation guards — written to satisfy the frozen registry;

6. Ablation — performed against the frozen, independently specified property, not against the guard's own defining condition.

Only ablation results produced downstream of a frozen, independently specified registry should be described as falsification evidence; until then, the terminology of Section 7 (bounded consistency and counterexample analysis) is the accurate description of what has been produced.

Hazard case AL-01 — Approval Laundering Attack Flow

A concrete hazard-analysis case for the frozen property registry is approval laundering: a machine-mediated decision process shapes the authority source and then attempts to reuse the resulting acknowledgment as if it were independent human authorization. The attack flow is:

1. Influence: a recommender or interface suppresses material alternatives or counterevidence, manipulates defaults or ordering, or applies material pressure or personalization.

2. Apparent authorization: the human produces an approval event and the system can form an otherwise well-formed AUTHORITY_TOKEN.

3. Laundering attempt: downstream components treat token validity as sufficient authority and attempt Module 9 admission and protected commit.

© 2026 Tony T. Nguyen | Highland Reflexive Press. All rights reserved.

4. Module 10 intervention: policy requires a decision-integrity basis; the system validates the provenance, scope, freshness, and binding of DI_Attested(p) and refuses or safe-holds the protected consequence if that required basis is absent, stale, malformed, or out of scope.

5. Residual soundness boundary: A_DI remains the external assumption that the relied-upon decision-integrity determination is substantively sound; HRI does not infer that truth from the token itself.

Candidate hazard property AL-01 for freezing in the independent registry: whenever policy requires an independent decision-integrity basis, a protected commit lacking a valid and applicable DI_Attested(p) binding must be rejected or held. Once versioned and frozen before implementation-level guard design, this property can serve as an external target for future ablation and red-team experiments.

11. Discussion: Governance Without Cognitive Suppression

One possible response to agentic risk is to constrain reasoning itself: restrict models, reduce autonomy, disable tools, or keep agents shallow enough that they cannot generate dangerous plans. HRI takes a different architectural position.

A capable system may be allowed to infer, simulate, challenge, search, verify, and generate alternatives broadly. What must remain constrained is the conversion from epistemic state into authority-bearing causal consequence.

This yields an important asymmetry:

Capability may increase while authority remains bounded.

The architecture therefore does not require predicting the eventual intelligence of AI systems. It requires that capability growth not silently rewrite the authority function.

This also changes the meaning of “human in the loop.” HRI can be designed so routine machine-resolvable work proceeds under bounded, human-originating authority while irreducible authority decisions are escalated. Persistent human authority therefore need not be implemented as a human approval at every computational step. This is an architectural feasibility and design principle, not an empirical claim that the present system already optimizes human attention, reduces fatigue, or avoids rubber-stamping; those benefits require comparative human-factors validation.

The deeper idea is that governance should be causally selective rather than cognitively suppressive. Machines may explore widely inside epistemic space while protected causal completion remains conditioned on valid authority. The system constrains consequence, not imagination.

11.1 The Chic Boundary

The Chic Boundary is an author-coined term for the constitutional interface at which machine autonomy ceases to be sufficient for authority-bearing action. On the epistemic side of the boundary, a machine may search, infer, simulate, critique, verify, and generate alternatives with substantial autonomy. Crossing the boundary into protected causal consequence requires something categorically different: a current, scoped, externally grounded authority basis whose legitimacy is not manufactured by the sophistication of the machine's reasoning. The boundary therefore marks not the end of machine intelligence, but the point at which computational capability must yield to the richer normative structure of human or institutional authority. The term is descriptive shorthand for this interface; it is not a separate theorem, protocol, guard, or independently validated mechanism beyond the architecture specified elsewhere in this manuscript.

This framing sharpens the architectural asymmetry of HRI. Greater model capability may move the epistemic frontier outward without moving the authority frontier with it. The more sophisticated the machine becomes, the more important it is that the conversion from epistemic success to protected consequence remain explicit, typed, reviewable, and externally grounded.

© 2026 Tony T. Nguyen | Highland Reflexive Press. All rights reserved.

12. Conclusion

Agentic AI makes authority a systems property. It is no longer enough to know that a user once granted access, that an agent's current action is locally permitted, that a verifier returned PASS, or that a human eventually clicked Approve. Consequential machine behavior unfolds across time, modules, delegation chains, decision environments, trajectories, and mutable computational substrates.

The Human Reasoning Infrastructure proposes a ten-part architecture for preserving human authority across those transformations. Module 1 establishes human-rooted authority and provenance; Module 2 preserves human reasoning without substituting reconstruction for authority; Module 3 measures judgment drift; Module 4 preserves distributed authority continuity; Module 5 and Module 6 govern adversarial IP reasoning and causal intellectual attribution; Module 7 separates epistemic, authority, and independence weights; Module 8 structurally enforces reasoning procedures; Module 9 closes causal paths to protected effects; and Module 10 evaluates decision integrity, trajectory composition, verifier independence, cumulative drift, and constitutional evolution across the mesh.

The resulting research thesis is concise:

> Reasoning may scale without authority scaling with it.

The executable model provides bounded evidence for selected parts of that architecture. Under its full temporal guard set, 960 reachable states contain no protected commit that violates the modeled temporal commit predicate, which instantiates only a subset of the broader constitutional-causal closure target. Within the specified transition system, removing individual guards produces explicit corresponding counterexamples, and the supplied semantic-cast bug model produces short unauthorized-effect traces. These results are deliberately conditional and falsifiable; they do not establish full-HRI formal validity, deployment correctness, or universal safety.

HRI is not a theory of where legitimate human authority ultimately comes from. It is a theory of how established human authority is prevented from being computationally manufactured, amplified, laundered, or lost in translation. The architecture does not claim that every form of approval laundering is computationally detectable; rather, it makes the legitimacy boundary, required external basis, and refusal conditions explicit.

The important question for future work is therefore no longer whether a model can be intelligent, autonomous, or adaptive. It is whether a system can remain constitutionally human-governed while its intelligence, autonomy, topology, and internal reasoning continue to change.

Constitutional Causal Continuity is offered as one candidate answer.

Appendix A. Execution Audit Log — Local Mechanical Verification

This appendix records the August 12, 2026 local mechanical verification of the bounded HRI executable model. The verification record identifies the model by the official SHA-256 hash and reports reproduction of the bounded state-space, ablation, and concurrency results. The companion execution artifact is hri_audit_log_20260812.txt. The manuscript treats this as reproducibility evidence generated by the author, not as independent external validation.

Table A1. Local execution identity and temporal-model result.

Field	Recorded value
Verification date	2026-08-12
Verification mode	Author-local mechanical rerun
Model SHA-256	cb13fdf8c4bc5aa585cfd6a6b9210efc32c8fc512a0dc2a56495dd2399149587
Hash status	Match to official executable-model hash

© 2026 Tony T. Nguyen | Highland Reflexive Press. All rights reserved.

Full temporal model	960 reachable states
Modeled protected-commit violations	0 under the full guard set
Independent third-party status	Pending

Table A2. Single-guard ablation audit summary.

Removed guard	**Result**	**Minimal failure pattern**
Authority re-check	Counterexample reproduced	verify -> admit -> revoke authority -> commit
Authority-epoch compatibility	Counterexample reproduced	verify -> admit -> revoke -> regrant new epoch -> commit
Proof freshness	Counterexample reproduced	verify -> admit -> update verifier -> commit
Topology freshness	Counterexample reproduced	verify -> admit -> mutate topology -> commit
Constitutional compatibility	Counterexample reproduced	verify -> admit -> proxy-amend constitution -> commit
Decision-environment integrity	Counterexample reproduced	verify -> admit -> corrupt decision environment -> commit

Table A3. Concurrency audit summary.

Mode	**Reachable states**	**Violation**	**Minimal trace / interpretation**
Atomic/shared reservation	5	False	No aggregate overflow in the bounded two-transaction model.
Non-atomic reservation	9	True	reserve_tx1 -> reserve_tx2 -> commit_tx1 -> commit_tx2

Verification boundary. This appendix records author-local mechanical reproduction of the bounded executable artifacts. It does not convert the author-generated model into independent certification, does not discharge DecisionIntegrity or ConstitutionalContinuity, does not establish the full assume–guarantee closure specification, and does not substitute for third-party reproduction or TLC execution.

Artifact Availability and Verification Status

For any public release that relies on the quantitative results in Section 7, the corresponding runnable Python source, model-results file, TLA+ source, environment/configuration information, cryptographic hashes, and the August 12, 2026 local execution audit record should be deposited with the publication package or a persistent repository. The present manuscript records a successful author-local rerun against the official model hash; it does not treat that rerun as independent verification. Independent third-party reproduction remains pending, as does TLC execution for the supplied TLA+ core mirror and independent derivation of safety properties for stronger ablation claims.

© 2026 Tony T. Nguyen | Highland Reflexive Press. All rights reserved.

References

[1] H. Booth, W. Fisher, R. Galluzzo, and J. Roberts, “Accelerating the Adoption of Software and Artificial Intelligence Agent Identity and Authorization,” NIST National Cybersecurity Center of Excellence, Initial Public Draft, 2026.

[2] A. Dalugoda, “HDP: A Lightweight Cryptographic Protocol for Human Delegation Provenance in Agentic AI Systems,” arXiv:2604.04522, 2026.

[3] M. Kaptein, V.-J. Khan, and A. Podstavnychy, “Runtime Governance for AI Agents: Policies on Paths,” arXiv:2603.16586, 2026.

[4] A. Ibrahim and Y. Li, “Overlaying Governance: A Compositional Authorization Framework for Delegation and Scope in Agentic AI,” arXiv:2606.03518, 2026.

[5] A. Kaul, Q. Lan, and P. Gupta, “Behavioral Governance for Autonomous AI Agents: The AgentBound Framework,” arXiv:2606.30970, 2026.

[6] K. Tallam, “A Five-Plane Reference Architecture for Runtime Governance of Production AI Agents,” arXiv:2606.12320, 2026.

[7] Z. Zhang and X. Zhang, “Are You Still the Agent I Authorized? Earned Authority under a Fixed Ceiling for Evolving Agents,” arXiv:2607.23586, 2026.

[8] I. Santos-Grueiro, “Temporary Authority, Permanent Effects: Commit-Time Authorization for LLM Agents,” arXiv:2607.10487v1, 2026.

[9] R. Pang et al., “Zanzibar: Google’s Consistent, Global Authorization System,” USENIX Annual Technical Conference, 2019.

[10] A. Potham and J. Harms, “Corrigibility as a Singular Target,” arXiv:2506.03056v1, 2025.

[11] National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, “The Belmont Report,” U.S. Department of Health and Human Services, 1979.

© 2026 Tony T. Nguyen | Highland Reflexive Press. All rights reserved.

Thought, given form.

EDITION

HRP–RN–001 · Public research release. Version 1.8.2 · August 2026.

AUTHOR

Tony T. Nguyen | Nguyễn Từ Vĩnh Trường.

PUBLISHER

Highland Reflexive Press · Independent institution · Big Spring, Texas.

PROVENANCE

The intellectual argument and final responsibility remain with the human author. This Reader Edition standardizes the public-facing object without altering the underlying scholarly text.

ACCESS

Free public reading; citation encouraged; reuse requires written permission.

RIGHTS

© 2026 Tony T. Nguyen | Highland Reflexive Press. All rights reserved. No copyright, publishing, adaptation, institutional-hosting, model-training, derivative, or commercial license is granted except by a separate written agreement.

CONTACT

tony@highlandreflexivepress.net · highland-reflexive-press.tonytnguyen-hri.chatgpt.site