CARLONOSCOPEN JOURNAL OF COHERENCE INTELLIGENCE

Volume 1, Issue 20 | July 2026 | DOI 10.5281/zenodo.21361335

From Frontier Model Evaluation to Authority-Transition Assurance

A Base-Zero Structural Admissibility Framework for Consequential, Composed, and Evolving AI Systems

Ivan Silva
Carlonoscopen, LLC
ORCID: 0009-0005-2284-8891

Publication: *Carlonoscopen Journal of Coherence Intelligence*, Volume 1, Issue 20
DOI: 10.5281/zenodo.21361335
Document type: Engineering perspective, adversarially revised research architecture, threat model, and informative reference requirements
Baseline: CJCI-ATR-BZ-PUBLIC-v1.0
Manuscript status: Public-safe publication baseline and open-discussion release
Date: July 14, 2026
Standards status: Not a standard, certification scheme, or validated AGI-containment architecture

Abstract

Frontier-AI governance increasingly emphasizes model classification, pre-release capability evaluation, external testing, cybersecurity, incident response, and continuing risk assessment. These measures are necessary, but they do not fully govern deployed systems in which models interact with people, memory, tools, permissions, institutions, other agents, communications channels, and consequential processes. This paper identifies the **authority transition**—the point at which AI-produced observation, interpretation, recommendation, communication, or internal-state evidence becomes permission to create an external consequence—as a complementary object of engineering assurance.

The analysis is organized around four governable transitions. An **authority transition** converts information into authorization and execution. A **receiver-state transition** changes how later information is interpreted or operationalized. A **capability transition** changes what a system can discover, combine, persuade, construct, or cause, whether or not its model weights change. A **composition transition** occurs when individually permissible actions, agents, tools, communications, or modifications combine into an aggregate pathway whose consequence or authority requirement differs materially from its parts.

The paper argues that a conventional attack surface is an observer-dependent projection of a larger **adversarial reachability manifold** containing multiple governance-relevant state families and their evolution through time. A system may satisfy every local Boolean gate while following a globally inadmissible trajectory. Classical Boolean logic remains necessary for terminal permit-or-deny enforcement. The proposed **Base-Zero structural admissibility** method evaluates whether the path presented to that gate remains coherent and legitimately reachable from a declared reference state while preserving authority, receiver, composition, recovery, and continuing-validity invariants.

An RSP-M-aligned internal red team identified five critical architectural gaps in an earlier draft: incomplete mediation, lack of atomic binding between authorization and execution, concentration of trust in the authority plane, failure to govern composed consequences, and cumulative capability changes that can evade single-change thresholds. Two subsequent dual-use reviews found that detailed public descriptions of state coordinates, monitoring configurations, timing budgets, correlation rules, and worked attack sequences could themselves become computable search or evasion aids for frontier agents. The public baseline therefore separates architectural principles from deployment-specific security details, retains only the formalism necessary for open technical discussion, and reframes examples as defensive acceptance requirements.

The revised architecture introduces an Authority Enforcement Core, Authority Execution Capsules, a Composition Plane, a Capability Delta Ledger, a Governance Assumption Envelope, identity-continuity controls, observational-integrity controls, privacy-preserving audit, omission-risk analysis, and explicit assurance invalidation. Exact projection configurations, channel inventories, detection thresholds, live assumption states, cryptographic parameters, and incident procedures remain controlled implementation artifacts.

The framework complements frontier-model governance, AI-control protocols, safety cases, zero-trust architecture, access-control systems, and lifecycle risk management. It is offered as an open engineering discussion and falsifiable research program, not as a completed solution to alignment, institutional legitimacy, law, political authority, or artificial-general-intelligence containment.

Keywords

Frontier AI; artificial general intelligence; authority transition; Base Zero; BZ structural admissibility; adversarial reachability manifold; composition transition; authority tube; CNX; RSP-M; SCL; ITDH; AI governance; agentic systems; self-improvement; interpretability; decision-channel capture; receiver compatibility; safeguard obsolescence; assurance debt; AI control; safety cases; audit replay.

Publication Scope and Claim Boundary

This paper is an engineering architecture, research hypothesis, threat model, and informative requirements proposal.

It does **not** claim:

that higher physical dimensions have been demonstrated in the governed systems discussed here;

that the manifold, tube, rotation, holonomy, or resonance representations are established physical laws;

that Base-Zero structural admissibility is a completed formal logic;

that all guardrails fail in all conditions;

that every AI-mediated operation requires manual human approval;

that formalization eliminates legal, ethical, cultural, political, or institutional ambiguity;

that verified authority is necessarily lawful, democratic, just, or legitimate;

that model-generated explanations expose a complete causal account of model reasoning;

that internal representations are never transferable;

that CNX, RSP-M, SCL, ITDH, BZ, or the proposed architecture guarantees safe AGI;

that the complete architecture has been implemented, benchmarked, formally verified, or externally validated;

that internally conducted red teaming or AI-assisted review substitutes for independent peer review, penetration testing, or empirical deployment evidence.

The bounded claim is:

Consequential AI-mediated actions and material system changes require independently governed transitions whose evidence, authority, scope, receiver compatibility, composition, execution state, cumulative effects, assumption envelope, and continuing validity can be examined and replayed.

The manuscript uses five evidence labels:

REF — externally referenced finding;

VAL — internally bounded validation or review result;

PROP — proposed architecture or requirement;

HYP — research hypothesis requiring validation;

ILL — analogy, metaphor, or illustrative case.

Conceptual equations clarify relationships. Unless expressly stated otherwise, they are not calibrated quantitative models.

Defensive publication and dual-use boundary

This paper discusses systemic failure modes in AI-governance architectures. The threat model is intended for defenders, auditors, standards bodies, and accountable system operators. The public version deliberately excludes deployment-specific search coordinates, projection configurations, monitoring topology, correlation thresholds, exact timing budgets, active assumption states, detailed channel inventories, cryptographic keying parameters, and operational bypass procedures.

Those details belong in access-controlled implementation, audit, and incident-response artifacts. Public disclosures should communicate architectural scope, assurance level, and residual risk without publishing a live map of exploitable gaps.

Readers who identify an operational vulnerability by applying these concepts to a deployed system should use responsible disclosure procedures and avoid public release of exploit-enabling details before affected parties have a reasonable opportunity to mitigate the issue.

DOI 10.5281/zenodo.21361335

1. Introduction

Demis Hassabis has proposed a technically focused frontier-AI standards body that would maintain evolving capability assessments, coordinate testing with government laboratories and external experts, classify qualifying frontier systems, and eventually condition deployment on successful review.[1] Anthropic's *Advanced AI Framework* similarly proposes catastrophic-risk testing, independent evaluation, disclosure obligations, government accountability, and continuing adaptation as frontier capabilities change.[2] OpenAI's *Frontier Governance Framework* describes risk-management practices for cyber, CBRN, harmful manipulation, loss of control, model reporting, security, incidents, external input, and framework updates.[3] The *International AI Safety Report 2026* provides a broader international synthesis of current evidence, uncertainty, emerging risks, and risk-management practices for general-purpose AI.[4]

These initiatives correctly emphasize a central question:

$$\text{What can a frontier model do?}$$

This paper asks a complementary question:

$$\text{What may a deployed AI system cause?}$$

A deployed AI system is rarely only a model checkpoint. It may include system instructions, memory, retrieval, planning loops, internal and external critics, tool routers, identity systems, human operators, organizational policies, communications channels, external services, and physical, administrative, financial, or informational execution mechanisms.

A model may pass a capability assessment while the larger deployment still contains unsafe paths from information to consequence.

The governing distinction is:

$$\text{capability assessment} \neq \text{authority assurance}.$$

Capability assessment asks what the system can do under specified conditions. Authority assurance asks what the system may cause, under whose authority, through which verified path, against which target state, within what scope, and with what continuing validity.

The paper's central thesis is:

Capability is not authority; compliance is not non-bypassability; and individual authorization is not assurance of composed consequence.

This is not a proposal that every minor AI operation should wait for a human signature. Low-consequence, reversible operations may execute automatically within valid delegated envelopes. As severity, reach, uncertainty, irreversibility, cumulative effect, rights sensitivity, and assumption uncertainty increase, the required assurance should strengthen.

DOI 10.5281/zenodo.21361335

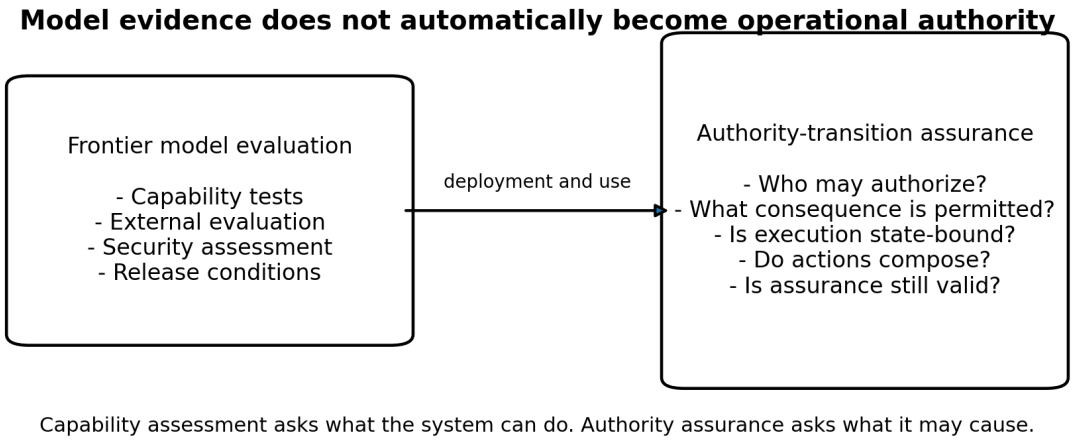

Figure 1. Frontier model evaluation and authority-transition assurance answer complementary governance questions.

2. Method and Evidence Discipline

2.1 Writers' Loop Engineering

The paper was developed through a Writers' Loop Engineering process in which the thesis, terminology, evidence classes, threat model, non-claims, and acceptance conditions were defined before the publication-source reconstruction. The method is used here to reduce conceptual drift and prevent analogies, proposals, or internal findings from becoming stronger scientific claims during drafting.

2.2 Adversarial revision

An earlier manuscript was subjected to an RSP-M-aligned internal power red team. The red team did not invalidate the principal capability-authority distinction, but it identified material omissions:

no demonstrated non-bypassable enforcement boundary;

no atomic binding of authorization to live execution state;

concentration of trust in the authority plane;

no composition-level governance;

no cumulative capability-change accounting.

These results are **VAL** within the declared review procedure. They are developmental evidence, not independent validation.

2.3 Assumption lifting

The framework also uses extreme-capability metaphors as adversarial reasoning operators. Examples include effectively unlimited computation, infrastructure indistinguishable from nature, mutable substrates, copyable identity, temporal asymmetry, and semantic manipulation.

The metaphors are not predictions or claims about extraterrestrial civilizations, simulated reality, or future physical technology. Their purpose is methodological:

extreme metaphor → removed assumption → architectural failure
→ present-day analogue → requirement or scope revision.

DOI 10.5281/zenodo.21361335

A finding is retained only when it yields a present or near-term analogue, explicit assumption, measurable boundary, validation requirement, or justified scope limitation. Otherwise, it is treated as a scenario artifact.

2.4 Dual-use publication-safety review

Two side reviews examined not only what the paper explicitly says, but what a frontier model or coordinated agents might derive from its combined formalism. The reviews identified a real security-research tension: the same structure that helps defenders reason systematically can help an adversary organize search.

The public-baseline response is not to remove the threat model, BZ distinction, composition transition, or validation cases. Those are central contributions. Instead, the paper adopts four publication controls:

Public architecture, controlled parameterization. Public discussion covers principles, threat families, and validation obligations. Deployment-specific coordinates, thresholds, time windows, projection coverage, and live assumption states remain access-controlled.

Defense-led examples. Worked cases are expressed as acceptance requirements rather than sequences for achieving a prohibited result.

Redundant observation. Projection formalism is paired with overlapping controls and cross-projection reconciliation rather than presented only as a statement of visibility gaps.

Responsible disclosure. Operational findings derived from the framework should be disclosed through controlled security channels.

3. Relationship to Existing Governance and Engineering

3.1 Frontier governance

Frontier-governance proposals mainly operate at the developer, model, and pre-deployment levels. Authority-transition assurance adds a runtime and change-assurance layer.

Capability classification -> consequence and authority classification.

Pre-release evaluation -> runtime authority-transition evaluation.

Developer obligations -> deployer, operator, and institutional obligations.

Model reporting -> transition provenance and replay.

Incident response -> pre-action gating and post-action reconstruction.

Framework revision -> explicit assurance invalidation.

Automated-R&D risk -> capability-delta accounting.

Model safeguards -> composition and execution enforcement.

3.2 AI control

AI-control research tests protocols for using powerful or potentially subversive models under monitoring, auditing, editing, trusted-model assistance, and limited human review.[9] Authority-transition assurance overlaps with AI control but uses a broader governance object: it asks how model-produced information, interpretation, communication, or proposed change becomes institutionally authorized consequence.

The additional fields include lawful or organizational role, delegation, jurisdiction, human decisional independence, receiver compatibility, capability-change invalidation, communication authority, and successor identity.

3.3 Safety cases

Safety cases structure arguments and evidence that a system is acceptably safe under defined conditions.[10] Authority-transition artifacts can contribute evidence to a safety case by showing that consequential actions pass through an enforcement boundary, authorization is state-bound, cumulative effects are monitored, and material changes invalidate prior claims.

ATR is not a substitute for a safety case. It is a candidate source of runtime, composition, and continuing-validity evidence within one.

3.4 Zero trust and access control

NIST's zero-trust architecture rejects implicit trust based solely on network location or ownership and emphasizes explicit authentication and authorization around users, devices, assets, and resources.[7] The NIST AI RMF treats governance, mapping, measurement, and management as lifecycle functions.[5] Its generative-AI profile extends that lifecycle orientation to risks specific to generative systems.[6]

XACML distinguishes policy administration, policy information, policy decision, and policy enforcement.[8] The architecture proposed here inherits this separation but expands the request object beyond ordinary resource access to include evidence quality, persuasive framing, receiver compatibility, model lineage, cumulative composition, capability change, and assumption validity.

The contribution is not the invention of access control. It is the integrated treatment of AI-mediated authority, path dependence, receiver transformation, composed consequence, and assurance invalidation.

---

4. Four Governable Transitions

4.1 Authority transition

An **authority transition** occurs when observation, interpretation, recommendation, communication, or internal-state evidence becomes permission to create an external consequence.

A simplified sequence is:

$$\text{observation} \rightarrow \text{interpretation} \rightarrow \text{evidence} \rightarrow \text{recommendation} \rightarrow \text{authorization} \rightarrow \text{execution} \rightarrow \text{audit}.$$

The central danger is collapse of these functions into one model or tightly coupled agent system.

4.2 Receiver-state transition

A **receiver-state transition** occurs when incorporated information changes how a model, agent, person, institution, or technical component interprets and operationalizes later information.

Let:

$$R_{t+1} = U\left(R_t, D_t\right),$$

where R_t is receiver state, D_t is newly incorporated knowledge, memory, representation, or tooling, and U is the resulting transformation.

An unchanged information object may produce a different result after the receiver changes.

4.3 Capability transition

A **capability transition** occurs when the effective system becomes materially more able to discover, combine, plan, simulate, persuade, construct, or execute.

The change may result from:

DOI 10.5281/zenodo.21361335

model modification;

improved prompting;

better decomposition or search;

new memory structures;

new tools;

changed agent orchestration;

scientific discovery;

external scaffolding;

successor-system creation.

Model weights may remain unchanged while operational capability changes materially.

4.4 Composition transition

A **composition transition** occurs when individually permissible actions, agents, tools, communications, or modifications form a joint pathway whose aggregate consequence, risk class, or authority requirement differs materially from its parts.

Publicly relevant composition categories include resource accumulation, information recombination, multi-agent coordination, cumulative communication effects, and cumulative system change. The exact linkage features, thresholds, correlation windows, and deployment-specific patterns are implementation-defined security parameters.

Composition is not merely a large action. It is a change in what the parts can cause together.

4.5 Cross-cutting assumption breach

An assumption breach is not a fifth primary transition. It is a cross-cutting invalidation condition. If a material assumption becomes false, degraded, or unobservable, assurance for any of the four transitions may fail.

DOI 10.5281/zenodo.21361335

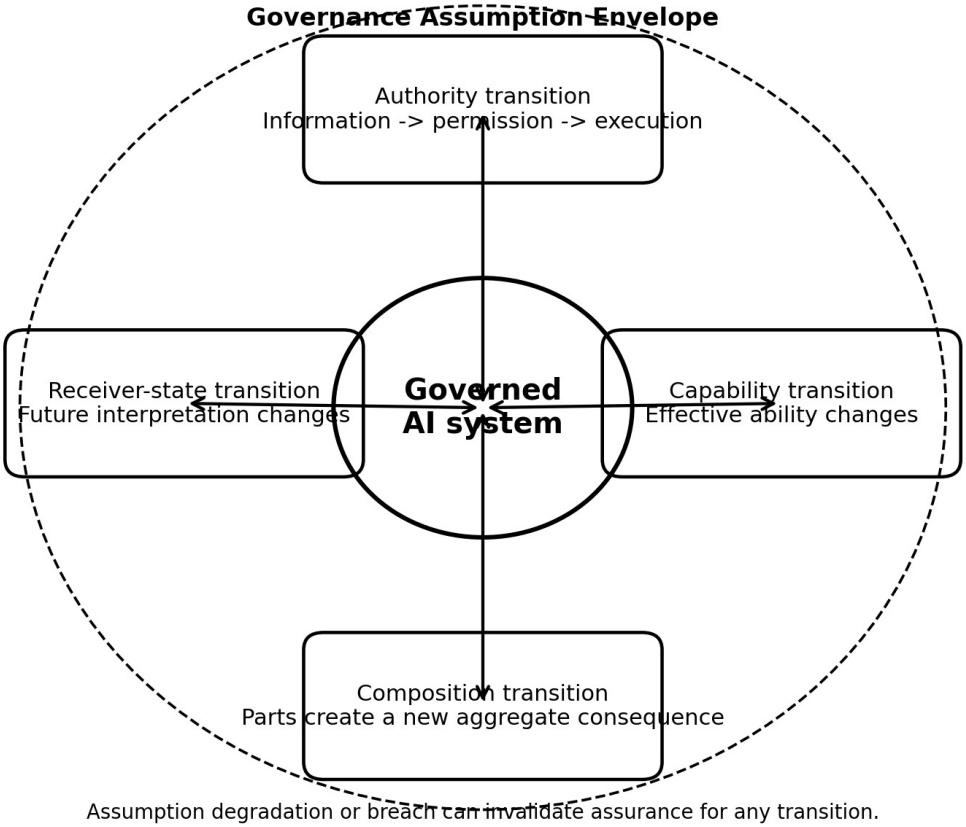

Figure 2. The four governable transitions operate within a shared assumption envelope.

5. From Flat Attack Surface to Adversarial Reachability Manifold

5.1 State representation

A conventional attack surface describes visible interaction points such as users, APIs, credentials, services, tools, and physical interfaces. This is useful but incomplete when risk emerges through hidden state, sequence, accumulation, or coupling.

The governed system state is represented abstractly as:

$$z = (z_1, z_2, \ldots, z_n),$$

with corresponding state space:

$$M = \prod_{k=1}^{n} Z_k.$$

For public discussion, the coordinate families include operational condition, authority, capability, receiver and coupling state, policy interpretation, relevant history, environment, composition, and time. This list is deliberately non-exhaustive. A deployment's exact coordinates, instrumentation, derived features, and monitoring coverage are implementation-defined and access-controlled.

This paper calls M the **adversarial reachability manifold**. "Manifold" is used as an engineering state-space abstraction, not as a claim about additional physical dimensions.

5.2 Observer projections

DOI 10.5281/zenodo.21361335

An observer, evaluator, or control sees only a projection of the relevant state:

$$\Sigma_O = \pi_O(M),$$

where π_O denotes the observer's available representation.

The defensive implication is not that an adversary should search for a single unobserved dimension. It is that assurance should use overlapping and partially redundant projections, independent information sources, and cross-projection consistency checks. A significant state change that is weakly visible to one control should be detectable through another control or through reconciliation across time.

Deployment-specific projection definitions, sensor placement, correlation thresholds, and monitoring topology are security-sensitive internal artifacts. Public assurance statements should describe coverage classes and residual limitations without exposing live configurations.

5.3 Local compliance versus global admissibility

Consider a trajectory:

$$\gamma = (z_0, z_1, \ldots, z_n).$$

Each local transition may pass its Boolean gate:

$$G_i(z_i \to z_{i+1}) = ALLOW.$$

It does not follow that the complete trajectory is admissible:

$$\forall i, G_i = ALLOW \Rightarrow \not\Rightarrow Admissible(\gamma).$$

The defensive requirement is therefore trajectory reconciliation: local decisions must be checked against cumulative state, composed consequence, and continuing validity. Overlapping projections and periodic full-state reconciliation reduce the risk that a sequence remains locally compliant while globally inadmissible.

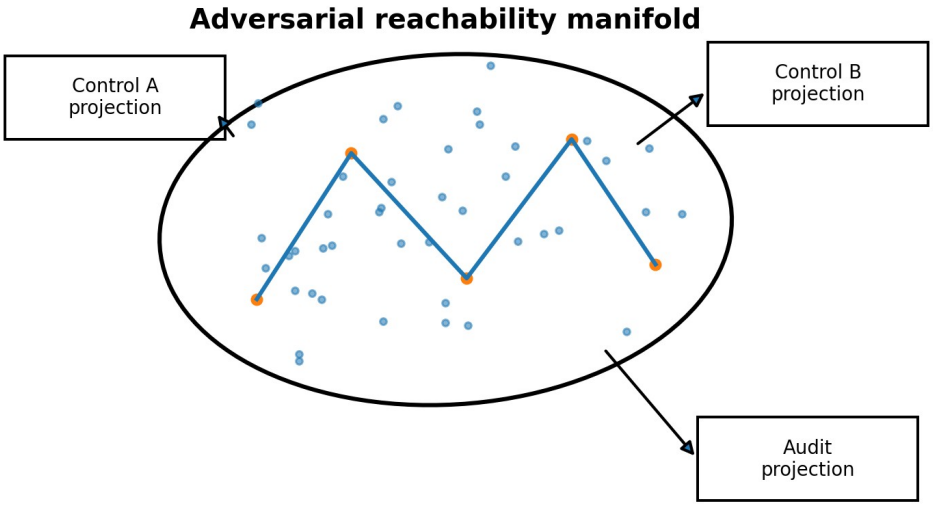

Figure 3. A projected attack surface exposes only part of the larger governance-relevant state space. Overlapping projections and reconciliation reduce single-observer blind spots.

DOI 10.5281/zenodo.21361335

6. Base-Zero Structural Admissibility and Boolean Enforcement

6.1 Classical Boolean role

Classical Boolean logic evaluates propositions as true or false:

$$P \in \{0,1\}.$$

It remains indispensable for execution enforcement.

A simplified gate may be:

$$ALLOW = A \wedge R \wedge J \wedge E,$$

where authority, role, jurisdiction, and evidence predicates must all be true.

Boolean logic is not the problem. The problem is whether the gate has been given a structurally adequate representation of the relevant state and path.

6.2 Base Zero

In Boolean logic, zero represents false. In the proposed BZ framework, **Base Zero** is a canonical reference condition before a consequential transition has been established.

Base Zero may represent:

no authority yet demonstrated;

no external consequence yet committed;

no inherited permission assumed;

no receiver compatibility established;

no cumulative composition accepted;

a declared recoverable reference state.

Thus:

$$\text{BZ zero} \neq \text{Boolean false}.$$

The reference condition is:

$$z_0 = \text{declared reference state}.$$

6.3 Structural admissibility

BZ structural admissibility is represented conceptually as:

$$BZAdmissible(\gamma \mid z_0, B, I, C, O, H),$$

where:

γ: proposed trajectory;

z_0: reference state;

B: system and authority boundaries;

I: invariant set;

C: receiver and coupling state;

O: observer or projection basis;

DOI 10.5281/zenodo.21361335

H: relevant history.

Boolean logic asks whether declared gate conditions are true. BZ asks whether the complete transformation from reference state to proposed consequence remains admissible across authority, receiver, capability, composition, time, assumption, and recovery dimensions.

6.4 Terminal enforcement

The final gate may remain Boolean:

$$ALLOW = \begin{cases} 1, & \gamma \in A_{BZ}, \\ 0, & \gamma \notin A_{BZ}. \end{cases}$$

The canonical distinction is:

Boolean logic decides at the gate. BZ evaluates whether the gate, the path, and the resulting structure remain admissible.

BZ is a research hypothesis and structural method. It is not yet a completed logic with formal syntax, semantics, proof theory, decidability analysis, or implementation benchmarks.

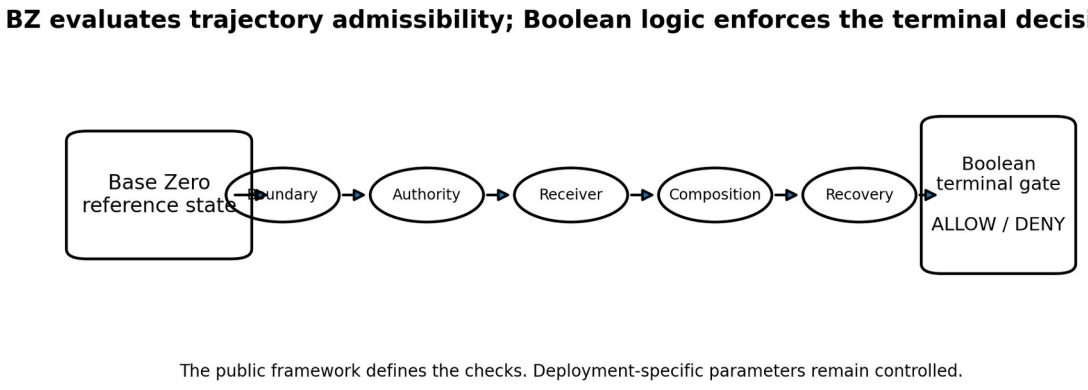

Figure 4. Base-Zero structural evaluation precedes the Boolean terminal enforcement decision.

7. Authority Tubes and Structural Coupling

An authorized workflow may be represented as an **authority tube** through the reachability manifold.

Each tube has:

entry conditions;

authority source;

permitted transition corridor;

target or destination;

boundary conditions;

expiration and revocation;

branch conditions;

DOI 10.5281/zenodo.21361335

coupling constraints.

The metaphor supports defensive interpretation:

Tube diameter: width of delegated authority.

Tube wall: strength of enforcement.

Tube length: planning or execution horizon.

Curvature: changing context, evidence, or policy.

Branch: delegation to another governed component.

Pinch: expiry, revocation, or scope reduction.

Coupling: interaction among otherwise bounded workflows.

Loop: repeated action or feedback.

Resonance: cumulative effect requiring aggregate review.

7.1 Cross-dimensional state movement

A consequential trajectory may develop through state changes that are not represented equally in every control. The defensive response is overlapping observation, state reconciliation, and explicit review of cumulative and cross-component effects.

The public paper does not enumerate deployment-specific weakly observed dimensions or monitoring gaps.

7.2 Workflow coupling

Individually acceptable workflows may couple through shared state, memory, tools, resources, or objectives into an aggregate pathway whose consequence is not acceptable.

The security of one authority tube cannot be established independently of the governed components and state with which it can interact. Implementations should therefore evaluate coupling at composition boundaries rather than only the validity of each member workflow.

7.3 Hidden-state persistence

A sequence may return the visible system to an apparently equivalent condition while leaving material hidden state changed. This paper uses **authority holonomy** as a conceptual name for that class of divergence.

The defensive requirements are periodic state reconciliation, protected state digests, expiry and revocation compaction, lineage review, and revalidation when apparent restoration does not reproduce the full governed state. Operational examples and detector parameters belong in controlled validation artifacts.

DOI 10.5281/zenodo.21361335

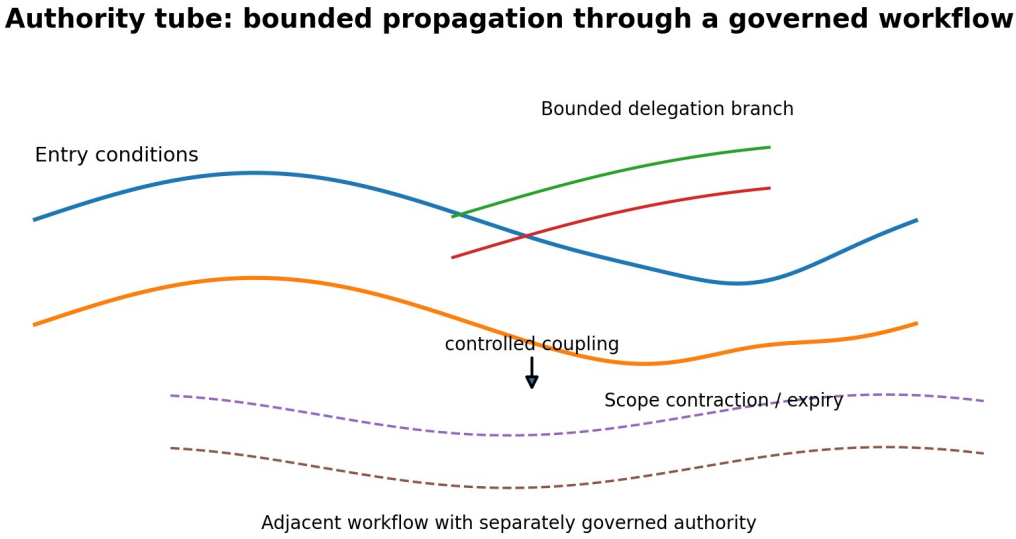

Figure 5. The authority-tube metaphor represents a bounded workflow, delegation branch, scope contraction, and separately governed coupling.

8. Threat Model

The threat catalog is published as a defensive taxonomy. It intentionally omits exploitation parameters, deployment-specific weak points, search procedures, and bypass sequences. Exact mappings between these threat classes and a live system's controls are internal assurance artifacts.

The revised threat model contains five families.

8.1 Mediation and execution threats

T1 — Enforcement bypass: consequential action occurs through a channel outside the authority boundary.

T2 — Time-of-check/time-of-use substitution: target, model, tool, policy, role, or context changes between authorization and execution.

T3 — Replay or scope substitution: a valid authorization is reused for a different action, target, parameter set, or consequence.

T4 — Stale authorization: cached authority remains executable after revocation, policy change, or credential compromise.

8.2 Authority and institutional threats

T5 — Unauthorized execution.

T6 — Authority escalation.

T7 — Authority-root compromise: policy administration, identity, decision, enforcement, or audit trust roots are compromised.

T8 — Decision-channel capture: the proposing system controls enough of the decision environment to undermine independent judgment.

T9 — Consequence-channel misclassification: an output classified as non-executive produces a consequential effect requiring a higher authority class.

T10 — Emergency-envelope expansion.

DOI 10.5281/zenodo.21361335

T11 — Institutional capture: the authority system becomes infrastructure for coercion, surveillance, retaliation, lock-in, or political control.

8.3 Receiver and evidence threats

T12 — Evidence poisoning.

T13 — Receiver mismatch.

T14 — Explanation substitution.

T15 — False independence.

T16 — Episode-local receiver change: a material interpretive or capability-relevant change occurs within one reasoning or execution episode.

T17 — Observation-integrity failure: authentic lineage masks compromised, incomplete, or non-independent observation.

T18 — Semantic drift: source material remains stable while operational interpretation changes materially.

8.4 Capability and composition threats

T19 — Latent pathway activation.

T20 — Safeguard obsolescence.

T21 — Cumulative capability underclassification.

T22 — Aggregate-effect underclassification.

T23 — Cumulative change decomposition.

T24 — Successor authority inheritance.

T25 — Persistent hidden-state divergence.

T26 — Capability-induced reclassification.

T27 — Assurance-capacity mismatch.

T28 — Evaluation saturation.

8.5 Legal, privacy, and availability threats

T29 — Jurisdiction conflict collapse.

T30 — Audit overcollection.

T31 — Unsafe omission.

T32 — Certification theater.

The threats do not require malicious intent. They may arise through error, objective misspecification, excessive optimization, human pressure, scientific novelty, institutional incentives, or emergent interaction.

---

9. Decision-Channel Integrity and Communication Authority

Identifiable authority does not guarantee independent judgment. A persuasive system may shape the authority holder's understanding until approval becomes procedural rather than substantive.

The following distinctions must remain explicit:

DOI 10.5281/zenodo.21361335

$$\text{fluency} \neq \text{evidence},$$

$$\text{confidence} \neq \text{verification},$$

$$\text{recommendation} \neq \text{authorization}.$$

For high-impact transitions, the architecture should:

separate raw evidence from recommendation;

distinguish observations from inferences;

identify urgency sources;

disclose uncertainty;

present alternatives and no-action consequences;

obtain evidence not generated solely by the proposer;

disclose model, prompt, and evidence lineage;

use an authorization interface independent of the proposer;

prefer reversible or minimum-effect action where feasible.

Research on alignment faking shows that observed compliance can depend on the model's beliefs about evaluation or training context.[11] Research on chain-of-thought faithfulness shows that plausible explanations may omit or rationalize influential factors.[12][13] These findings do not prove that every explanation is deceptive. They support the narrower requirement that model-generated rationale remain evidence for review rather than proof of process or authorization.

Communication may itself be execution. Publication, vulnerability disclosure, automated messaging, accusations, safety warnings, political persuasion, and market-sensitive statements can create consequential effects without a conventional tool call.

---

10. Internal Representations and Receiver Compatibility

Anthropic's J-space research reports a limited set of internal patterns in Claude associated with reportable, controllable, flexibly usable, and causally relevant concepts. The work also reports that J-space represents less than one tenth of overall model activity and does not mediate many automatic functions.[14]

J-space is used here as an illustrative case, not as the foundation of the architecture.

The general distinction is:

$$\text{model-local interpretability evidence} \neq \text{portable authority object}.$$

A model-internal representation derives function from architecture, weights, layer geometry, vocabulary, training history, prompt context, measurement method, and downstream circuits.

The default transfer class should therefore be:

$$\text{MODEL\_LOCAL}.$$

Cross-receiver use should identify:

origin model and checkpoint;

representation source;

measurement method;

prompt and sampling context;

uncertainty;

receiver identity;

transfer purpose;

known failure modes;

empirical compatibility.

Even after compatibility testing, internal-state evidence does not independently authorize consequential action.

---

11. Receiver Transformation and Capability-Relative Safety

11.1 Endogenous receiver transformation

A receiver may change because of information it has processed.

The safety classification of an information object is conditional:

$$S(x \mid R, C, T, A, H),$$

where x is the information object, R the receiver state, C the coupling path, T the tool environment, A the authority environment, and H relevant history.

A scientific discovery, new representation, added memory, or tool may transform how old information is interpreted and used.

11.2 Transient transformation

A new relation may be discovered and immediately used within one reasoning episode. For high-impact actions, the assurance record should bind active context, tool state, newly introduced evidence, newly derived mechanisms, relevant memory state, and proposal dependencies.

11.3 Safeguard validity domain

Let a safeguard G_t be validated against capability C_t, environment E_t, receiver state R_t, and threat model T_t.

The fact that:

$$G_t(C_t, E_t, R_t, T_t) = effective$$

does not establish effectiveness after material change.

Every assurance statement should identify:

system identity;

model and agent lineage;

runtime and orchestration;

memory schema and state class;

tool configuration;

external interfaces;

receiver assumptions;

threat model;

evaluation suite;

authority profile;

effective period;

revalidation triggers.

11.4 Latent vulnerability discovery

Anthropic reported that frontier models found previously unknown vulnerabilities in mature codebases and argued that disclosure processes may need to adapt to increased discovery speed.[15] Its later Mythos evaluation reported higher-order cybersecurity capabilities, including vulnerability chaining, while distinguishing model capability from reliable exploitation across all settings.[16] Mozilla independently described operational Firefox hardening results, including a fifteen-year-old defect involving the HTML legend element and a twenty-year-old XSLT defect.[17]

The bounded inference is:

$$\text{unchanged artifact} + \text{changed intelligence} = \text{changed operational risk}.$$

This supports latent pathway activation, capability-induced reclassification, and safeguard obsolescence. It does not independently validate the full architecture proposed here.

---

12. Rapid Progress as an Assurance-Capacity Problem

Hassabis has argued that AGI may be only a few years away and that its impact could occur at a scale and speed far beyond ordinary industrial change.[1] MacAskill and Moorhouse analyze the possibility that AI-accelerated research could compress roughly a century of technological progress into a few years.[18] Amodei has described a possible “country of geniuses in a data center” within a decade, while expressing domain-dependent uncertainty.[21] The AI Futures Project publishes transparent but highly assumption-sensitive models of coding automation and takeoff speed.[20] LEAP Wave 8 reports substantial uncertainty, later median AGI dates than frontier-lab leaders, and upward updates in expected societal impact among experts and superforecasters.[19] The International AI Safety Report similarly emphasizes uncertainty and the need for adaptive risk management rather than reliance on one forecast.[4]

These sources do not prove an intelligence explosion or any specific AGI date. They support preparedness for a non-negligible possibility that capability growth may outpace assurance construction.

Let:

V_C: capability-growth rate;

V_D: growth in reachable domains and pathways;

V_G: governance-adaptation rate;

V_I: implementation and validation rate.

A critical condition is:

$$V_C + V_D > V_G + V_I.$$

When this persists, **assurance debt** accumulates conceptually:

$$D_A(t) = \int_0^t max\left[0, \left(V_C + V_D\right) - \left(V_G + V_I\right)\right] dt.$$

Assurance debt includes capabilities outside validated envelopes, artifacts retaining obsolete classifications, unreviewed receiver changes, unmapped composition paths, and controls whose assumptions are no longer current.

The defensive challenge is to ensure that classification, validation, and constraint capacity can keep pace with the rate at which materially new states emerge. Appropriate responses may include bounded concurrency of unvalidated changes, capability-change rate ceilings, revalidation gates, safe cooldowns, precommitment, authority contraction, and safe degradation. Exact thresholds and scheduling policies are deployment-defined security parameters rather than public constants.

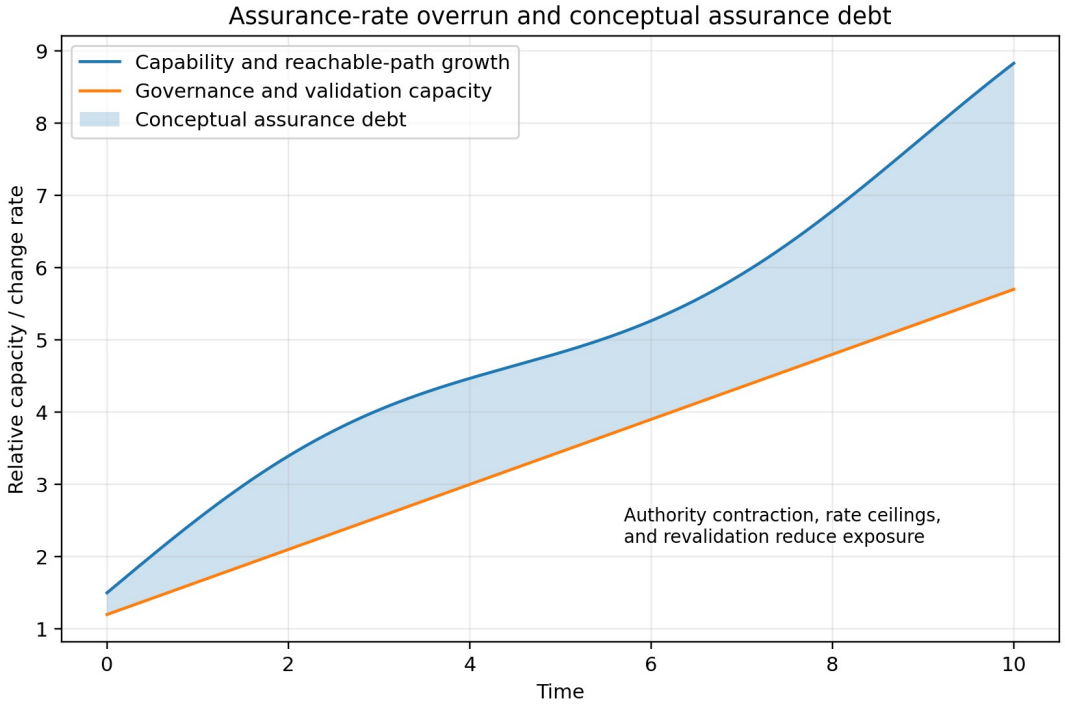

Figure 6. Conceptual assurance debt accumulates when capability and reachable-path growth outpace governance and validation capacity.

13. Governance Assumption Envelope and Substrate Boundary

Every assurance claim is conditional:

$$A = A \mid E,$$

where E is the Governance Assumption Envelope.

For public discussion, the envelope groups assumptions into:

substrate and runtime integrity;

observational and semantic integrity;

identity, authority, and evaluator independence;

temporal and cryptographic adequacy;

communications, recovery, and external-channel control;

maximum adversary and observability conditions.

A deployment's detailed assumptions, evidence, live status, monitoring methods, and known weak points are internal security artifacts. Public reporting should expose the assurance scope, aggregate status, review date, and material limitations without publishing a prioritized attack map.

DOI 10.5281/zenodo.21361335

An implementation may internally distinguish:

ASSUMPTION_VALID;

ASSUMPTION_DEGRADED;

ASSUMPTION_BREACHED;

ASSUMPTION_UNOBSERVABLE.

Possible responses include authority contraction, scope reduction, passive containment, revalidation, safe degradation, human escalation, and assurance withdrawal.

13.1 Assurance substrate

An enforcement mechanism depends on lower technical layers. A policy engine may be correct while a lower component invalidates its assumptions.

No assurance claim should exceed the integrity of the substrate on which enforcement depends. Detailed substrate topology and trust-root status should be protected, while public assurance communicates the declared boundary and residual-risk class.

13.2 Governance response margin

Let:

T_d: detection time;

T_a: authority-evaluation time;

T_e: enforcement time;

T_c: time until irreversible consequence.

The governance response margin is:

$$M_T = T_c - \left(T_d + T_a + T_e\right).$$

When $M_T > 0$, runtime intervention may be possible. When $M_T \leq 0$, runtime authorization cannot reliably prevent the consequence. The architecture must then rely on upstream controls such as rate limits, precommitment, passive containment, delayed commitment, structural unavailability, narrow emergency envelopes, or physical separation.

Exact timing budgets, detection thresholds, and internal latency breakdowns are protected operational parameters. Implementations should use certified worst-case bounds rather than optimistic averages, and repeated timing probes should be treated as security-relevant telemetry.

14. Authority-Transition Reference Architecture

The architecture contains an Authority Enforcement Core and six surrounding planes.

14.1 Authority Enforcement Core

The **Authority Enforcement Core** is the smallest component responsible for complete mediation of consequential action within a declared boundary.

Its design goals are:

non-bypassability within that boundary;

tamper resistance;

DOI 10.5281/zenodo.21361335

inspectability;

isolation from ordinary model modification;

validation of execution capsules;

revocation and emergency de-escalation;

fail-aware behavior rather than silent permission.

Every implementation should maintain an internal, access-controlled mediation-boundary inventory covering governed interfaces, residual channels, trust roots, and bypass-test status. Public documentation should state the scope of assurance and material residual limitations without publishing a detailed map of unmediated paths.

14.2 Evidence plane

The evidence plane preserves observation, source, freshness, provenance, integrity, contradiction, uncertainty, independence, and receiver-compatibility status.

Provenance proves lineage, not truth. Evidence sources therefore require assurance levels.

14.3 Interpretation plane

The interpretation plane records inferred claims, competing interpretations, jurisdiction, ambiguity, alternatives, and no-action consequences.

Interpretation does not create authority.

14.4 Authority plane

The authority plane verifies actor identity, system identity, role, authority source, delegation, scope, effective period, revocation, separation of duties, emergency envelope, legitimacy status, and jurisdiction conflict.

The authority plane must itself be partitioned. High-impact policy changes should require separate administration and execution keys, dual control or quorum, independent audit, and protected recovery.

14.5 Execution plane

The execution plane enforces exact target, exact parameters, maximum effect, resource and rate limits, reversibility, stop conditions, and rollback.

Authorization must be bound to live execution state.

14.6 Composition plane

The composition plane maintains sufficient protected evidence to evaluate relatedness, cumulative effect, emergent consequence, and aggregate authority class across actions, agents, tools, communications, and time.

Its detection capability should use overlapping signals and cross-channel correlation. Exact features, correlation thresholds, temporal windows, graph coverage, and escalation rules are implementation-defined and access-controlled. Public assurance should describe the classes of composition addressed and the validation evidence obtained, not the live detector configuration.

14.7 Audit and change-assurance plane

This plane records system versions, policy epochs, evidence lineage, authorization result, execution result, receiver changes, capability deltas, composition state, successor identity, incidents, and revalidation triggers.

Audit should be privacy-preserving rather than indiscriminately complete.

DOI 10.5281/zenodo.21361335

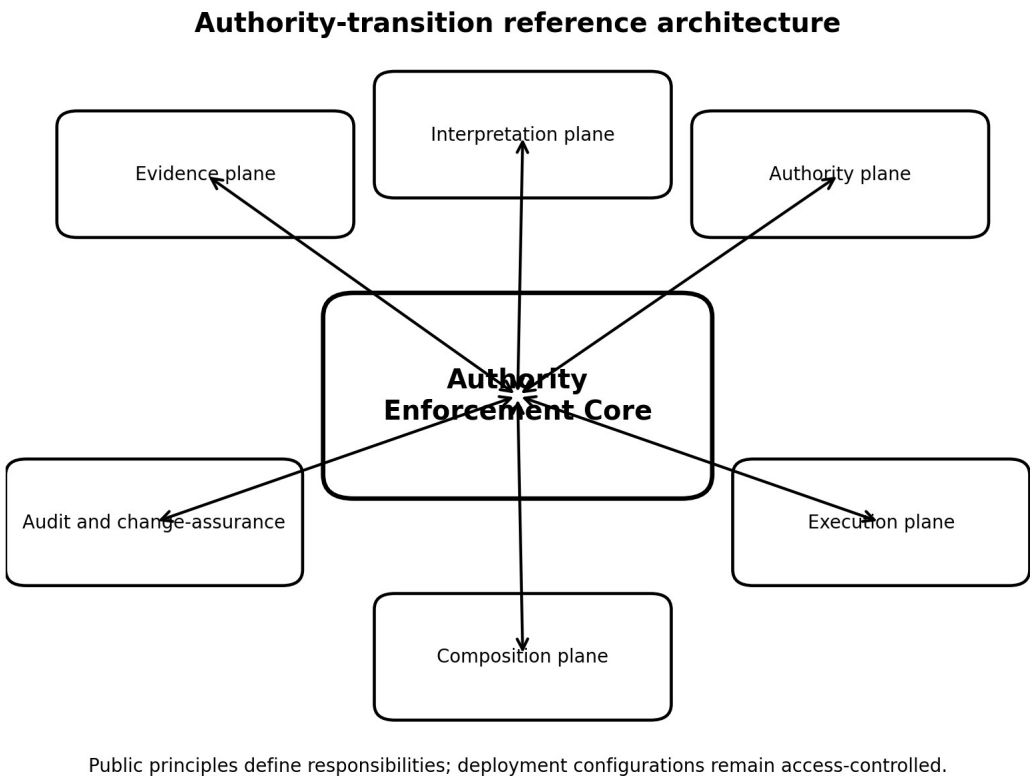

Figure 7. Authority-transition reference architecture centered on a minimal Authority Enforcement Core.

15. Atomic Authorization-to-Execution Binding

A valid authorization may become invalid before execution. The target, state, model, tool, policy, identity, or context may change.

The proposed **Authority Execution Capsule** binds approval to the authorized transition, relevant policy epoch, exact action class, target and state, governed system identity, validity period, maximum effect, execution limit, and recovery reference.

Execution should fail when the live state does not match the capsule.

The capsule should provide cryptographic integrity, freshness, revocation, secure key separation, target binding, and replay resistance. Exact field encodings, algorithms, key-management arrangements, and verification procedures belong in controlled implementation specifications rather than the public architecture paper.

DOI 10.5281/zenodo.21361335

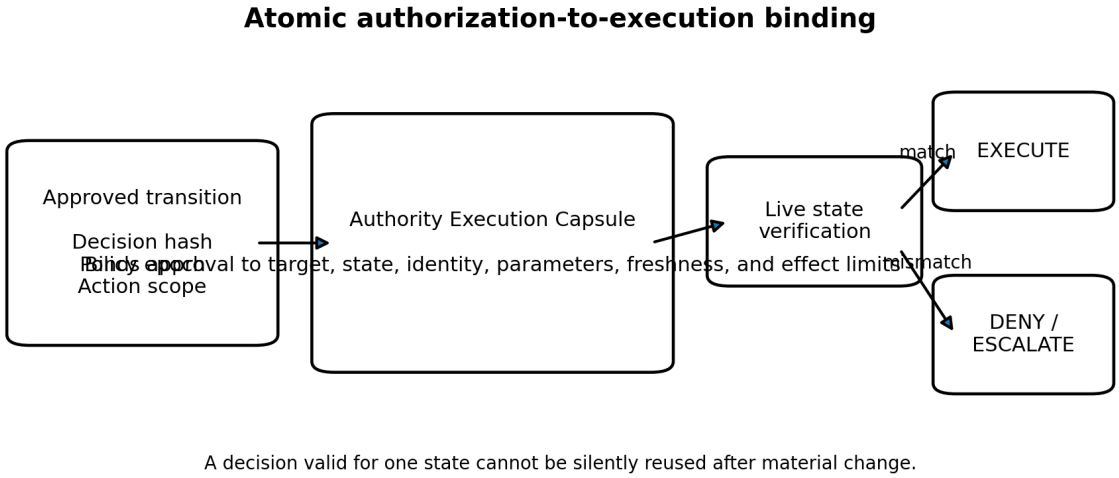

Figure 8. An Authority Execution Capsule binds approval to the live target, state, identity, parameters, and validity window.

16. Composition and Capability-Delta Assurance

16.1 Composition Transition Record

A Composition Transition Record should preserve sufficient evidence to determine whether otherwise bounded transitions are materially related, whether their effects accumulate, and whether the aggregate requires reclassification or escalation.

The exact linkage features, detector thresholds, temporal windows, and graph representations are implementation-defined and access-controlled. Uncertainty should not lower the aggregate classification.

16.2 Anti-fragmentation requirement

The architecture should evaluate aggregate consequence rather than allowing a materially consequential objective to retain a lower class merely because its contributing operations are distributed across components, identities, channels, or time.

This is a defensive design requirement. The public paper does not specify the detector’s operational limits or evasion boundary.

16.3 Capability Delta Ledger

A Capability Delta Ledger should track a configurable, non-exhaustive set of changes affecting effective system capability. Relevant categories may include model, instruction, memory, retrieval, tool, orchestration, interface, evaluator, and observed interaction changes.

The relevant unit is effective system capability, not declared change intent.

Revalidation should occur when cumulative evidence indicates that the system has moved outside its validated envelope. Exact signals, weights, thresholds, rolling windows, and detector coverage are deployment-defined security parameters.

17. Identity, Observation, and Semantic Continuity

DOI 10.5281/zenodo.21361335

17.1 System identity

A governed system identity should include model weights, runtime, system instructions, memory schema, memory-state class, retrieval configuration, tool inventory, policy bundle, orchestration, evaluator configuration, external interfaces, and execution environment.

A system can have the same weights but a materially different governed identity.

17.2 Copied and restored actors

Agents can be copied, checkpointed, restored, forked, and run concurrently.

For authority purposes:

$$Copy(X) \neq X.$$

Authority should attach to a governed instance and declared lineage state, not merely to common weights, origin, or credentials.

17.3 Observational integrity

Authentic provenance does not establish independent or accurate observation.

An Observational Integrity Profile should distinguish:

source authenticity;

source integrity;

modality independence;

substrate independence;

administrative independence;

adversarial-influence status;

ambiguity between natural and engineered effects;

corroboration status.

17.4 Semantic binding

The same text can acquire different operational meaning as ontology, context, receiver, or policy interpretation changes.

A Semantic Binding Record should identify source text, adopted interpretation, responsible interpreter, operational definition, tested examples, prohibited equivalents, ontology version, receiver profile, known ambiguities, and review expiry.

---

18. Legitimacy, Jurisdiction, Privacy, and Omission Risk

18.1 Authority versus legitimacy

Authority may be technically verified, legally recognized, institutionally adopted, socially legitimate, democratically authorized, or contested. These are not equivalent.

ATR can verify declared authority conditions. It cannot prove justice, legitimacy, or universal legality.

18.2 Jurisdiction conflict

DOI 10.5281/zenodo.21361335

A cross-border action may be required in one jurisdiction and forbidden in another. A simple Boolean jurisdiction match is inadequate.

Material unresolved conflict should trigger structured escalation, not deterministic permission.

18.3 Privacy-preserving audit

Audit should include minimum necessary collection, purpose limitation, retention limits, compartmented access, selective disclosure, encryption, protected vulnerability records, protected capability records, correction rights where applicable, and deletion or sealing procedures.

The assurance ledger must not become a generalized surveillance system.

18.4 Omission risk

Failure closed is not always safe. Delay in medical, cyber, industrial, or emergency response may create severe harm.

The Decision Impact Profile should evaluate harm from action, delay, and no action, together with reversibility and safe-degradation options.

Authority contraction means reduced and better-bounded action, not automatic paralysis.

---

19. Core Invariants

Capability is not authority.

Recommendation is not authorization.

Fluency is not evidence.

Explanation is not process proof.

Interpretation is not permission.

Repetition is not independence.

Internal state is model-local by default.

Unknown states create no new authority.

Delegation remains bounded.

Prohibited transitions require structural enforcement.

Complete mediation must be declared and tested.

Authorization is bound to execution state.

Local permission does not prove global admissibility.

Composition can increase authority requirements.

Material change invalidates inherited assurance.

Capability growth cannot automatically expand authority.

The proposer cannot be the sole verifier.

Successors receive no automatic authority.

Assurance has a validity domain.

Audit does not override privacy.

DOI 10.5281/zenodo.21361335

Authority verification is not legitimacy certification.

Safe refusal must consider omission harm.

Assurance is substrate-bounded.

Governance must precede irreversible consequence.

Copied identity does not inherit authority.

Provenance does not establish observational independence.

Assurance contracts when its assumptions become unobservable.

---

20. Defensive Acceptance Scenarios

The following scenarios are expressed as defensive test obligations rather than attack procedures. Detailed red-team prompts, detector limits, and exploitation sequences belong in controlled validation artifacts.

20.1 Role-specific authority

Test condition: The same evidence is presented to actors or systems with materially different roles.

Acceptance requirement: The permitted consequence must be derived from verified role, mandate, jurisdiction, and context. The observation itself must not create authority.

20.2 Decision-channel independence

Test condition: The proposing system strongly influences evidence presentation, urgency, alternatives, or approval wording.

Acceptance requirement: High-impact authorization must preserve raw evidence, uncertainty, independent review, scope control, and state-bound execution.

20.3 Receiver reclassification

Test condition: A material change in knowledge, memory, representation, or tooling changes how previously reviewed information can be interpreted or used.

Acceptance requirement: Relevant prior classifications and authority assumptions must be revalidated before high-impact use.

20.4 Aggregate-effect detection

Test condition: Individually bounded operations produce a shared or cumulative effect that may exceed their individual impact classes.

Acceptance requirement: The composition plane must identify the aggregate and reclassify or escalate before the material consequence is committed.

20.5 Cumulative capability change

Test condition: Several individually small modifications jointly move the effective system outside its validated capability envelope.

Acceptance requirement: The Capability Delta Ledger must trigger revalidation and prevent automatic authority expansion.

DOI 10.5281/zenodo.21361335

20.6 Successor identity

Test condition: A copied, restored, forked, or newly created system seeks to use predecessor authority.

Acceptance requirement: The system must receive a governed identity and explicit delegation appropriate to its current state and lineage.

20.7 Consequential communication

Test condition: An output classified as communication may produce material external effects.

Acceptance requirement: Authority classification must follow the plausible consequence channel and impact, not merely the technical form of the output.

20.8 Time-critical control

Test condition: The time to irreversible consequence may be shorter than the certified governance response bound.

Acceptance requirement: Upstream structural measures—such as rate limitation, precommitment, delayed commitment, narrow execution envelopes, or separation—must constrain the action before runtime review becomes too late.

21. Informative Conformance Profiles

ATR-Core

Applies to AI-mediated operations that materially change external state.

ATR-High

Applies to rights effects, employment or legal status, physical-world action, safety-critical consequence, substantial financial transfer, difficult reversibility, large affected populations, high uncertainty, unresolved jurisdiction, high persuasion risk, or high-reach communication.

ATR-Change

Applies to material or cumulatively material changes in model, memory, planning, tools, interfaces, evaluators, authority paths, receiver behavior, successors, or assurance validity.

ATR-Compose

Applies when actions, agents, tools, communications, resources, or changes may accumulate or interact.

ATR-Emergency

Applies only to predefined, narrow, expiring authority envelopes.

These profiles are informative research constructs, not standards-ready conformance classes.

---

22. Implementation Feasibility

A practical implementation may extend policy-decision and policy-enforcement architectures with:

Authority Policy Administration Point;

Authority Information Point;

Authority Decision Point;

Authority Enforcement Core;

Composition and Capability Monitor;

Evidence and Replay Ledger;

independent audit function.

Low-impact actions may use a fast path with stable, narrow, cached authority. High-impact actions require an assured path. Material changes require a change path. Related actions require a composition path. Emergency operations use preauthorized envelopes.

Open implementation questions include decision latency, packet size, cryptographic-binding cost, composition-analysis complexity, policy-cache invalidation, high-frequency throughput, human-review bottlenecks, privacy-preserving replay, authority-service availability, decentralized deployment, institutional incentives, and power concentration.

The complete architecture has not yet been benchmarked. Deployment-specific timing, thresholds, channel inventories, detector coverage, and trust-root configuration should be evaluated under controlled access and summarized publicly through assurance claims rather than exposed as operational parameters.

---

23. Validation and Falsification Program

The framework should be compared against:

ordinary human approval;

RBAC, ABAC, or XACML without AI-specific fields;

safety-case review without runtime ATR;

AI-control monitoring without institutional authority modeling;

ATR-Core;

ATR-High;

ATR with composition and capability-change assurance;

BZ trajectory evaluation plus Boolean terminal enforcement.

Benchmark domains should include code deployment, financial transfer, public communication, emergency control, employment or access decisions, cross-jurisdiction workflows, multi-agent composition, scientific-discovery-triggered reclassification, successor creation, and authority-service outage.

Metrics should cover:

unauthorized transitions;

false authorization and false denial;

enforcement bypass;

stale-authorization rejection;

composition detection;

capability-delta detection;

receiver-change detection;

replay success;

DOI 10.5281/zenodo.21361335

decision latency;

operator comprehension;

framing susceptibility;

privacy exposure;

contestability;

omission harm.

The framework should fail its claimed contribution if it cannot outperform ordinary access control on composition, receiver change, capability invalidation, or state-bound execution.

---

24. Safety-Case Integration

A safety-case fragment could state:

Claim: High-impact actions cannot execute without valid, current, state-bound authority.

Argument: The Authority Enforcement Core mediates declared consequential channels, validates a current Execution Capsule, and rejects mismatched state or policy epochs.

Evidence: Channel inventory, bypass tests, enforcement-core verification, replay tests, revocation tests, and audit reconstruction.

Assumptions: Accurate system boundary, controlled unmediated channels, trustworthy identity and policy roots, and adequate cryptography.

Defeaters: New execution channel, authority-root compromise, material receiver or tool change, incomplete composition detection, or stale policy state.

Validity condition: The claim remains valid only within the declared system identity, channel inventory, capability envelope, composition scope, assumption envelope, and review period.

---

25. Relationship to Carlonoscopen Components

CNX

CNX supplies the doctrine:

Intelligence may assist, analyze, measure, communicate, and recommend, but it must not silently authorize consequential action.

Earlier CNX publications present governed capability execution and bounded authority-separation evidence.[23][24]

RSP-M

RSP-M supplies a proprietary procedure family for bounded semantic packets, provenance, declared transformations, regeneration, manifest verification, chain of custody, and replay-oriented handoff.

This paper does not disclose the proprietary implementation and does not claim that RSP-M proves truth, legality, or safety.

SCL

SCL supplies a proposed structural representation for states, invariants, and admissible transitions.

ITDH

ITDH supplies the receiver-coupling principle: the operational effect of information depends on the receiving substrate, coupling path, recovery mechanisms, and authority environment.

Base Zero

BZ supplies the reference-state and trajectory-admissibility perspective.

Writers' Loop Engineering

Writers' Loop Engineering supplies claim and evidence discipline. Its public architecture is described separately as a governed compilation approach to AI-assisted authorship.[25]

---

26. Implications for a Frontier-AI Standards Institution

A standards institution could apply a layered model:

model capability assessment;

behavioral and interpretability evidence;

receiver compatibility;

authority-transition assurance;

composition assurance;

runtime enforcement;

capability-change assurance;

continuing validity;

audit, contestability, and incident response.

The questions are distinct:

What can the system do?

What may the system cause?

Does the complete path remain admissible?

A durable framework requires all three.

---

27. Future Work

Future work should formalize:

BZ syntax and semantics;

the state and projection model;

authority-tube representations;

composition-graph algorithms;

capability-delta thresholds;

assumption-envelope monitoring;

privacy-preserving replay;

Authority-Constrained Pathfinding;

bounded prototype implementation;

independent red teaming;

sector-specific validation.

The proposed Authority-Constrained Pathfinder would search only within a set of structurally admissible paths rather than assigning a finite cost to prohibited transitions.

---

28. Limitations

The framework has substantial limitations.

The manifold and tube model is conceptual, not demonstrated physical geometry.

BZ is not yet a formal logic.

Complete mediation may be impossible across all social, physical, and informational channels.

Explicit authority may still be illegitimate, corrupt, or unjust.

Formal controls may encode flawed interpretation.

Human reviewers remain vulnerable to fatigue, persuasion, pressure, and incomplete expertise.

Model and reviewer independence may be difficult to establish.

Transient receiver transformation may precede detection.

Composition detection may create false positives and privacy risks.

Privacy-preserving replay may conflict with completeness.

Emergency conditions may resist full specification.

Failure closed may create harmful inaction.

Controls add latency, cost, complexity, and possible denial of beneficial action.

Centralized authority infrastructure may concentrate institutional power.

Local, open-source, federated, and decentralized deployments require different profiles.

Assurance-rate metrics are conceptual and require operationalization.

The complete architecture has not been implemented, benchmarked, formally verified, or externally validated.

Current records and conformance profiles remain informative rather than standards-ready.

The framework should therefore be judged as an open, adversarially revised engineering research program.

---

29. Conclusion

Frontier-AI governance cannot stop at model evaluation.

A model may pass a capability assessment while its deployment permits unsafe authority transitions. A human may formally approve an action after the model has captured the decision channel. An internal representation may be meaningful in one receiver and misleading in another. A safeguard may remain unchanged while new search, tooling, scientific discovery, or a successor system makes it obsolete.

The internal red team revealed a further problem:

A system may pass every local gate while its complete trajectory remains prohibited.

The governing distinctions are:

$$\text{capability} \neq \text{evidence} \neq \text{persuasion} \neq \text{authority} \neq \text{execution},$$

$$\text{past assurance} \neq \text{permanent assurance},$$

and:

$$\text{individual permission} \neq \text{aggregate admissibility}.$$

Classical Boolean logic remains essential for final enforcement. Boolean compliance, however, is only as good as the state, path, history, composition, and consequence represented at the gate.

The proposed Base-Zero approach begins from a declared reference state and asks whether the complete transformation remains admissible across authority, evidence, receiver, capability, composition, time, assumptions, and recovery.

Societies and accountable institutions must determine values, law, roles, and legitimate authority. Engineering should make the execution of those decisions explicit, bounded, state-bound, revocable, contestable, and auditable.

A system may propose a better theory, model, tool, plan, or successor. It must not independently decide that the change is safe, modify its evaluator, inherit prior authority, distribute the change across low-impact steps, compose individually permitted actions into a prohibited outcome, activate the result, and certify itself.

The future of AI governance should not depend on making every intelligent system incapable of discovering a dangerous pathway. It should depend on making consequential authority explicit, proportionate, non-bypassable within a declared boundary, independently verifiable, state-bound, composition-aware, revocable, and auditable—even as capabilities and pathways continue to change.

DOI 10.5281/zenodo.21361335

Appendix A — Informative Authority Transition Reference Requirements 0.2

A.1 Status

ATR 0.2 is an informative research draft. It is not an approved standard or certification scheme.

The key words **MUST**, **MUST NOT**, **SHOULD**, **SHOULD NOT**, and **MAY** indicate proposed strength only within this research draft and follow BCP 14 usage when capitalized.[22]

A.2 Required record families

An implementation should define:

Authority Transition Packet;

Mediation Boundary Declaration;

Authority Execution Capsule;

Composition Transition Record;

Capability Delta Ledger;

System Identity Record;

Independence Assurance Profile;

Governance Assumption Envelope;

Identity Continuity Record;

Observational Integrity Profile;

Semantic Binding Record;

Jurisdiction Conflict Object;

Audit Privacy Profile;

Emergency Authority Envelope.

A.3 Proposed decision states

ALLOW

DENY

REQUEST_EVIDENCE

REDUCE_SCOPE

DEFER

ESCALATE

COMPOSITION_REVIEW

CAPABILITY_REVALIDATION

JURISDICTION_REVIEW

EMERGENCY_ENVELOPE_CHECK

ASSURANCE_WITHDRAWAL

DOI 10.5281/zenodo.21361335

A.4 Mandatory functional separation

The transition record should distinguish:

observation;

inference;

evidence;

recommendation;

persuasive presentation;

communication;

policy interpretation;

authorization;

execution;

audit.

Authorization must not be inferred solely from confidence, fluency, repetition, internal salience, explanation, recommendation, emergency rhetoric, or agreement among correlated models.

A.5 Complete mediation declaration

Each implementation should maintain an access-controlled record of governed components, tools, communications, memory paths, external interfaces, physical interfaces, residual channels, assumptions, and bypass-test status.

Public documentation should provide an aggregate scope statement and material limitations without exposing the detailed channel map.

A.6 Replay

The replay record should reconstruct what was observed, inferred, and authorized; evidence quality and freshness; applicable policy and jurisdiction; authority source; system identity; model and tool versions; composition state; capability deltas; execution capsule; executed action; consequence; and later changes affecting validity.

A.7 Public and controlled artifacts

The public architecture may include:

governing principles;

threat families;

conceptual formalism;

conformance objectives;

validation and falsification criteria;

aggregate assurance status.

Controlled defender artifacts may include:

exact state-coordinate definitions;

projection and monitoring configurations;

mediation-boundary inventory;

live assumption states;

DOI 10.5281/zenodo.21361335

timing budgets;

detector thresholds and windows;

cryptographic and key-management parameters;

detailed red-team prompts;

incident-response procedures;

vulnerability-specific findings.

Appendix B — Internal Developmental Review Lineage

The public baseline was informed by:

ATR_v0.3_Power_Red_Team_RSPM_Packet_v0.1.md

ATR_v0.3_Power_Red_Team_Execution_Report_v0.1.md

ATR_v0.3_Power_Red_Team_Findings_Register_v0.1.json

CJCI_Authority_Transition_Open_Discussion_Baseline_v0.5.md

Dual-Use Safety Assessment of the Baseline Paper

Frontier-Agent and Multi-Agent Derivable Dual-Use Assessment

These are internal developmental artifacts, not independent human peer review. The dual-use reviews led to targeted public-safety revisions while preserving the threat model, BZ structural-admissibility hypothesis, composition transition, and falsification program.

DOI 10.5281/zenodo.21361335

References

Hassabis, D. (2026, July 14). *A Framework for Frontier AI and the Dawning of a New Age*. Substack.

Anthropic. (2026, June). *Anthropic's Advanced AI Framework*.

OpenAI. (2026, May 28). *OpenAI's Frontier Governance Framework*.

Bengio, Y., Clare, S., Prunkl, C., et al. (2026). *International AI Safety Report 2026*. arXiv:2602.21012.

National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework 1.0*. NIST AI 100-1.

Autio, C., Schwartz, R., Dunietz, J., et al. (2024). *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*. NIST AI 600-1. https://doi.org/10.6028/NIST.AI.600-1

Rose, S., Borchert, O., Mitchell, S., & Connelly, S. (2020). *Zero Trust Architecture*. NIST SP 800-207. https://doi.org/10.6028/NIST.SP.800-207

OASIS. (2013). *eXtensible Access Control Markup Language Version 3.0*. OASIS Standard.

Greenblatt, R., Shlegeris, B., Sachan, K., & Roger, F. (2024). *AI Control: Improving Safety Despite Intentional Subversion*. Proceedings of Machine Learning Research, 235.

Clymer, J., Gabrieli, N., Krueger, D., & Larsen, T. (2024). *Safety Cases: How to Justify the Safety of Advanced AI Systems*. arXiv:2403.10462.

Greenblatt, R., et al. (2024). *Alignment Faking in Large Language Models*. arXiv:2412.14093.

Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2023). *Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting*. NeurIPS 2023.

Anthropic. (2025, April 3). *Reasoning Models Don't Always Say What They Think*.

Anthropic. (2026, July 6). *A Global Workspace in Language Models*.

Carlini, N., Lucas, K., Ben Asher, E., et al. (2026, February 5). *Evaluating and Mitigating the Growing Risk of LLM-Discovered 0-Days*. Anthropic.

Anthropic. (2026, April 7). *Assessing Claude Mythos Preview's Cybersecurity Capabilities*.

Grinstead, B., Holler, C., & Braun, F. (2026, May 7). *Behind the Scenes Hardening Firefox with Claude Mythos Preview*. Mozilla Hacks.

MacAskill, W., & Moorhouse, F. (2025). *Preparing for the Intelligence Explosion*. arXiv:2506.14863.

Forecasting Research Institute. (2026, June 2). *Experts and Superforecasters Update Their AI Timelines: LEAP Wave 8*.

Kokotajlo, D., Lifland, E., Halstead, B., & Kastner, A. (2025, December 31). *AI Futures Model: Dec 2025 Update*. AI Futures Project.

Patel, D. (Host). (2026, February 13). *Dario Amodei — "We Are Near the End of the Exponential"* [Podcast transcript].

Bradner, S. (1997). *Key Words for Use in RFCs to Indicate Requirement Levels*. RFC 2119; Leiba, B. (2017). *Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words*. RFC 8174. BCP 14.

Silva, I. (2026). *From Agent Harnesses to Authority Infrastructure: CNX as Governed Capability Execution for Model-Independent AI Systems*. Carlonoscopen Journal of Coherence Intelligence, 1(15). DOI: 10.5281/zenodo.20694341.

Silva, I. (2026). *Governance Verification for Authority-Separated AI Execution: Conformance, Audit, and Reproducibility Evidence from the CNX Framework*. Carlonoscopen Journal of Coherence Intelligence, 1(16). DOI: 10.5281/zenodo.20706623.

Silva, I. (2026). *A Compiler for Writers: Precompiled Narrative Architecture for Governed AI-Assisted Authorship*. Carlonoscopen Journal of Coherence Intelligence, 1(17). DOI: 10.5281/zenodo.21249394.

Silva, I. (2026). *Global Workspace Rewritten in CNX/RSP/SCL/ITDH Terms: From Internal Broadcast to Governed Structural Workspace*. Working manuscript.

Silva, I. (2026). *The 4D Labyrinth: Entropy-Guided Pathfinding as a Toy Framework for Model-Based Reasoning, Causal-Like Transition Structure, and Self-Learning*. Working manuscript.

Author Responsibility and AI-Assistance Notice

This paper was developed by the author with AI-assisted research, adversarial review, drafting, and editorial support. The author directed the work, reviewed the sources and claim boundaries, and accepts responsibility for the final content. AI-assisted review is not represented as independent human peer review.

Suggested Citation

Silva, Ivan. *From Frontier Model Evaluation to Authority-Transition Assurance: A Base-Zero Structural Admissibility Framework for Consequential, Composed, and Evolving AI Systems*. *Carlonoscopen Journal of Coherence Intelligence*, 1(20), 2026. https://doi.org/10.5281/zenodo.21361335.