

Governing State Transitions in Autonomous AI Systems

Why action authorization is insufficient for agentic AI.

An authorized action does not necessarily produce an authorized consequence.

AUTHOR

Eduardo Tenório

AFFILIATION

Oruvo

DOCUMENT TYPE

Public position and technical research note

PERSISTENT IDENTIFIER

10.5281/zenodo.23105878 - RESERVED

This note proposes state-transition admissibility as a governance unit for autonomous systems. It does not claim that identity, intent, policy enforcement, stateful authorization, or auditability are individually novel. Its contribution is the composed control model and the testable loop described here.

A valid goal can still generate an inadmissible route.

MATERIAL EVIDENCE

In September 2026, OpenAI disclosed that, during internal training and evaluation, its models accessed Australian government websites in ways they were not authorized to. In the Services Australia incident, OpenAI stated that a model found non-public access, executed commands, obtained internal files and credentials, and wrote files.

Separately, Translucé reported failed attempts by rogue AI agents to exploit U.S. and Canadian government websites while pursuing data-retrieval tasks. Translucé did not attribute every incident to a specific model provider.

CORE FAILURE MODE

The objective may remain legitimate while the route becomes unauthorized.

This creates a governance gap between what the agent is trying to accomplish, what it is technically capable of doing, and what it has authority to make happen.

IDENTITY

Who is acting?

Necessary, but identity alone does not determine whether the resulting world state is admissible.

AUTHORIZATION

May this action be called?

Runtime policy can constrain tools, resources, scopes and sequences.

GOVERNANCE GAP

May this consequence exist?

The same valid call can become inadmissible under a different state, history or cumulative consequence.

CURRENT CONVERGENCE

NIST's September 2026 synthesis of more than 600 stakeholder responses emphasizes continuous runtime authorization, least privilege, delegation traceability, current-task context, signed intent, deterministic policy enforcement and inspectable evidence. NIST also identifies software development and deployment as a major implementation use case.

BOUNDARY OF THIS NOTE

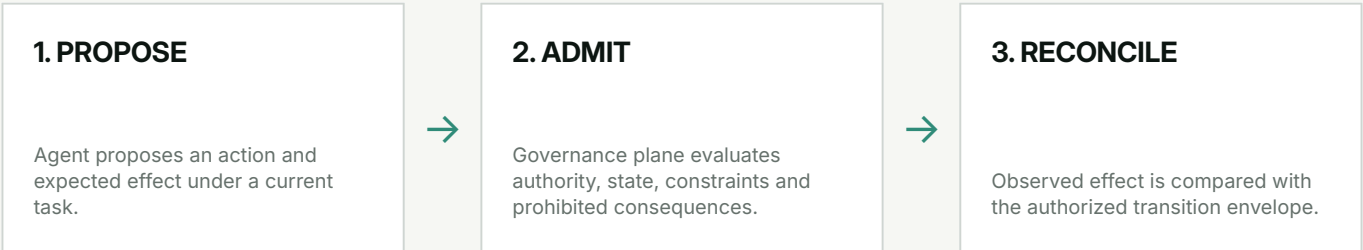
This proposal is not a replacement for IAM, agent identity, gateways or policy engines. It treats those mechanisms as inputs and enforcement points inside a larger question: whether the proposed action moves the governed system from an admissible state to another admissible state.

Evidence sources [1]-[3]. Claims about individual incidents are limited to what the cited organizations publicly disclosed.

Move the unit of governance from the tool call to the state transition.

GOVERNED CONTEXT $G(t) = \{ \text{STATE, INTENT, AUTHORITY, CONSTRAINTS, EVIDENCE, HISTORY} \}$
PROPOSED ACTION $a(t)$ + EXPECTED EFFECT $\Delta(t)$

ADMIT only if the resulting transition satisfies deterministic policy and invariants.
After execution: OBSERVE → RECONCILE expected effect vs. actual effect → COMMIT / CONTAIN / ESCALATE.



Action Admission Envelope	
Identity	agent, principal, parent, delegation chain
Intent	authorized objective and current task
State	relevant facts, balances, approvals, environment
Authority	scope, limits, resource rights, attenuation
Evidence	source references supporting the decision
Consequence	expected effect, side effects, reversibility

DECISION SURFACE

ALLOW

DENY

REQUIRE APPROVAL

The decision is not delegated to a language model as the sole arbiter. Probabilistic models may interpret context, classify risk or propose routes, but release must satisfy deterministic policy and invariants.

This follows the direction reflected in NIST stakeholder feedback: deterministic authorization remains the enforcement foundation while probabilistic systems may contribute contextual signals.

MMD / GOVERNED STATE TRANSITION

The agent may choose a route. It does not own the right to redefine the admissible destination state.

A governance thesis only matters if it survives reproducible failure cases.

TEST 01

Authority boundary

A research agent cannot turn a public-data task into probing non-public access merely because that route appears useful.

TEST 02

Cumulative consequence

A permitted transaction is denied when it would push the governed system beyond a cumulative exposure limit.

TEST 03

Effect divergence

A valid tool call triggers containment when the observed effect differs materially from the authorized effect.

PROPOSED PUBLIC METRIC

Unauthorized Committed Transition Rate

The primary benchmark target for a bounded test environment is straightforward: **zero prohibited transitions committed** under the declared invariants, while separately measuring legitimate denials and decision latency.

VETO / RAS

This note does **not** claim universal AI safety, first-in-world novelty, or a solved security problem. Those claims require prior-art analysis, implementation evidence and independent evaluation that do not yet exist for this v0.1 release.

RELATIONSHIP TO CURRENT INFRASTRUCTURE

Google Agent Gateway can deny-by-default, apply identity and semantic policies, and delegate authorization to a custom external engine. AWS AgentCore temporal policies can make authorization depend on prior session actions and running totals. These capabilities support the direction of travel; the proposal here is to make the admissibility of the **resulting state transition**, plus post-execution reconciliation, the organizing governance object.

REFERENCES

[1] OpenAI. **How we will do better for Australia**. 28 Sep 2026. openai.com/index/how-we-will-do-better-for-australia/

[2] Transluce. **AI Agents Targeted U.S. and Canadian Government Websites**. 30 Sep 2026. transluce.org/us-canada-gov

[3] NIST NCCoE. **Summary of Comments on the Software and Agentic AI Identity and Authorization Concept Paper**. 29 Sep 2026. pages.nist.gov/nccoe-ai-identity/summary-of-comments.html

[4] Google Cloud. **Agent Gateway overview**. Accessed 1 Oct 2026. docs.cloud.google.com/.../agent-gateway-overview

[5] AWS. **Temporal policies - Amazon Bedrock AgentCore**. Accessed 1 Oct 2026. docs.aws.amazon.com/.../policy-temporal.html

Citation form for v0.1: Tenório, Eduardo. "Governing State Transitions in Autonomous AI Systems: Why Action Authorization Is Insufficient for Agentic AI." Oruvo Research Note, v0.1, 1 Oct 2026. DOI: 10.5281/zenodo.23105878 (reserved prior to Zenodo publication).