Evaluation Authority Sandbox Console

authrex.systems · browser simulation · seeded PRNG · synthetic data only · client-side execution

> **Scope.** Everything in this document describes independent research artifacts: browser simulations, formal models, and unbuilt hardware reference designs. Maturity is self-assessed. Nothing here is a certification, an Authorization to Operate, a government endorsement, or evidence of a fielded system. Model-checking results apply to the stated finite model and assumptions, not to physical implementations.

PURPOSE

Simulates SANDBOX governing what an AI under evaluation may do: within the packaged simulation and stated mediation assumptions, proposed actions outside the configured test tier are rejected and recorded, irreversible-class actions are denied outright, and state resets cleanly between runs.

CONTROLS AND DISPLAY

- Tier controls set the configured test tier; the AI under test issues synthetic action proposals.
- The reset control demonstrates clean state between runs; the ledger records every decision.
- Scenario and seed reproduce runs; the V&V panel reruns packaged assertions.

READING THE VERDICTS

- CONTAINED / RESET VERIFIED / BREACH-ATTEMPT LOGGED are simulation state labels.
- Behavioral consistency between test and production, VM containment limits, covert channels, and model deception remain open research surfaces stated on the product page.

LIMITATIONS

Self-assessed TRL 3 to 4 reference architecture demonstration mapped to the NDAA §1534 requirement (statutory milestone dated April 1, 2026). Hardware anchors are unbuilt Rover and UAV reference designs: candidate future test platforms.

CANONICAL TERMINOLOGY

SATA: Sensor & Actor Trust Assessment · HMAA: Human-Machine Authority Allocation · ADARA: Adversarial Deception-Aware Reasoning · MAIVA: Multi-Agent Integrity Voting · FLAME: Forced Latency for Authority-Managed Escalation · CARA: Controlled Authority Recovery Architecture · ERAM: Escalation Risk Assessment Model

AUTHREX Systems · authrex.systems · Burak Oktenli · July 10, 2026 · Site documentation, CC BY 4.0 where stated; artifact licenses vary by record · Independent single-author research; self-assessed maturity; not a government publication