# AGI Alpha RSI

Alpha v0

Sovereign Invention Governance • Deterministic discovery → auditable advantage

**Deterministic invention operations for open-ended discovery, governed deployment, and compounding advantage.**

*Thesis: build the governance institution first — before AGI-scale systems mature.*

- Governed, deterministic invention OS; artifacts are schema-validated and replayable.
- OMNI provides search control (allocation) but never outcome authority.
- Breakthroughs are audited state transitions: reproduce → stress → persist → dossier.

STRATEGIC THESIS

# The institution we wish existed before nuclear weapons

**“AGI Alpha RSI is the governance institution we wish had existed before nuclear weapons — and now have a chance to build before AGI-scale systems mature.”**

### What RSI is

- Deterministic invention OS (pipeline + ledgers).
- Governance-by-construction: mechanical gates.
- Audit-ready breakthroughs packaged as dossiers.

### What RSI is not

- Not a novelty mill or hype engine.
- Not a black-box “outcome authority”.
- Not a one-off research project — it is sovereign infrastructure.

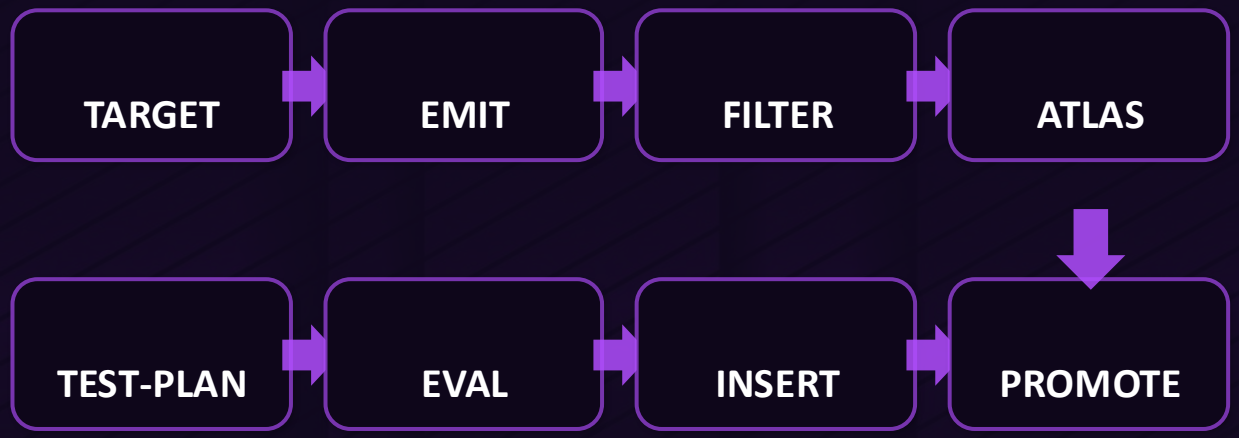
SYSTEM KERNEL

# Deterministic invention OS (v8)

## Core guarantees

- Determinism & replay: temperature = 0; manifests reproduce cycles.
- Schema-bound artifacts: failures hard-stop (no silent corruption).
- Baseline discipline: incumbent/neighbor/null comparisons by default.
- Append-only ledgers: monotonic state across runs.

## Pipeline (target → … → promote)




OMNI influences targeting/interestingness only — never insertion authority.

GOVERNANCE

# Search control ≠ outcome authority

**Exploration is encouraged, but outcomes are admitted only through mechanical gates.**

OMNI improves allocation; it does not confer authority.

### RISK

Prohibited-domain detection + risk reports gate outcomes.

### EVIDENCE

ECI semantics: confidence cannot inflate without execution.

### BASELINE

Always comparative vs incumbent/neighbor/null baseline.

### PERSISTENCE

High novelty forces probe-first + stress-tested advantage.

BREAKTHROUGH CONTROL

# Move-37 handling as a deterministic state transition

When NoveltyDistance and AdvantageDelta cross thresholds, RSI triggers mandatory steps:

**Recognize** — Thresholds + risk / confidence / ECI gates

**Reproduce** — Re-run candidate + baseline; verify hashes (fixed seeds)

**Stress-test** — Policy shocks; re-evaluate deltas; log sensitivity

**Persistence** — Require positive advantage under shocks

## High novelty ⇒ higher skepticism

Novelty ≥ 0.80 forces probe-first behavior and ADVANTAGE_PERSISTENCE uncertainty.

**Breakthroughs are admitted only as audited state transitions — not as narratives.**

Outcome authority remains mechanical: risk, evidence, baseline, persistence.

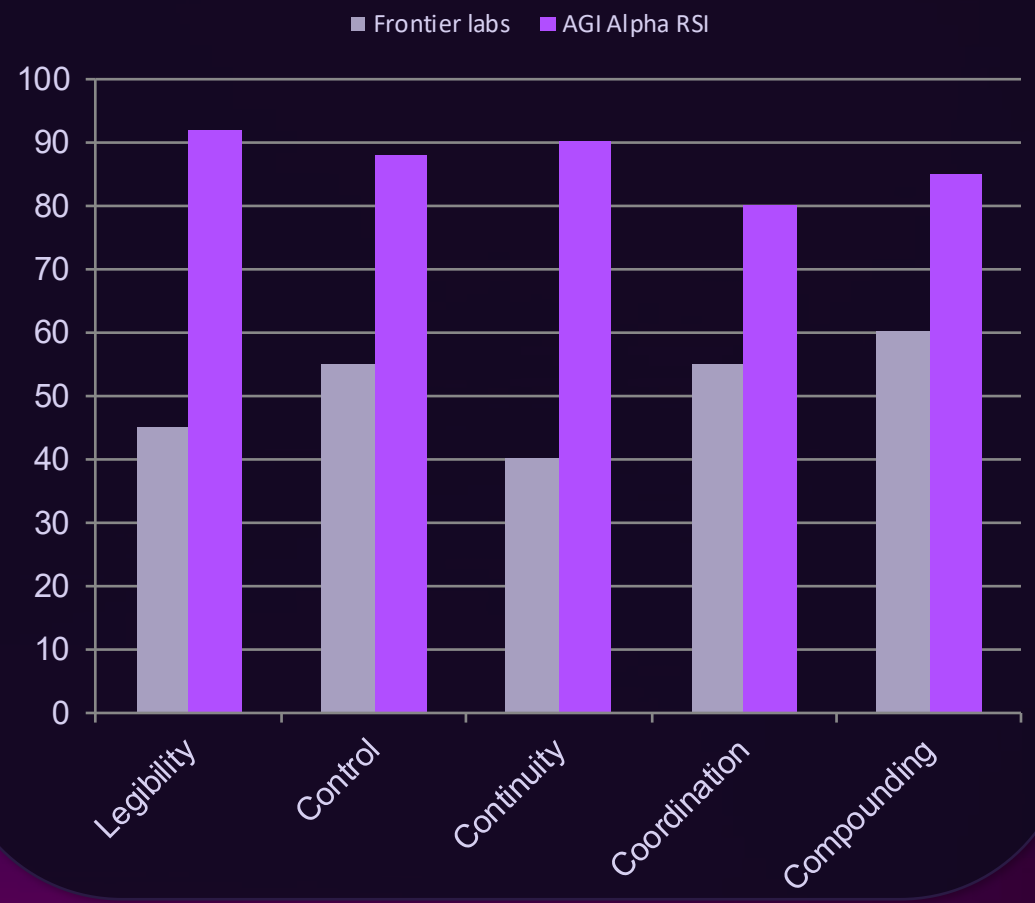
STRATEGIC VALUE

# State-capacity advantage (institutional, not just capability)

### State capacity proxies

- Legibility: replayability + audit trail.
- Control: mechanical gates; OMNI has no outcome authority.
- Continuity: append-only ledgers survive churn.
- Coordination: dossiers + schema-bound artifacts.
- Compounding: stepping stones retained; regression detectable.

### Illustrative comparison (0–100)



OPERATING MODEL

# RSI complements frontier labs — it does not imitate them

Frontier labs optimize for capability velocity. RSI optimizes for durable, auditable compounding advantage under governance.

**Frontier sprint**

- Rapid prototypes
- Taste-driven iteration
- Breakthrough chasing

**Governed sprint**

- Fast probes
- Hard gates
- Reproducible deltas

**Legacy bureaucracy**

- Slow approvals
- Weak measurement
- Low compounding

**AGI Alpha RSI**

- Deterministic replay
- Evidence-first promotion
- Sovereign dossiers

INSTITUTIONAL DESIGN

# DARPA is necessary — but insufficient

### DARPA strengths

- High-risk, high-reward R&D engine (since 1958).
- Catalyzes breakthroughs via time-bounded programs.
- Program manager autonomy; rapid portfolio churn.

### Why insufficient for AGI governance

- Project-based: not a permanent control plane.
- Outputs not mechanically governed after transition.
- Short PM tenures optimize velocity, not continuity.

### RSI fills the gap

**Command-and-control for invention**

Deterministic pipeline + mechanical gates + append-only ledgers + dossier packaging.

### Policy outcome

- DARPA discovers; RSI governs discovery.
- DARPA funds projects; RSI institutionalizes trust.
- RSI is the layer that survives personnel, politics, and time.

HISTORICAL LENS

# Manhattan • RAND • Nuclear C2 → RSI

## Manhattan Project

Capability first

- Crash-program speed
- Centralized execution
- Governance arrived later

## RAND

Strategy & doctrine

- Institutional analysis
- Legibility for decision-makers
- Persistent advisory capacity

## Nuclear C2

Control & verification

- Positive control
- Fail-safe defaults
- Audit/authorization logic

**AGI Alpha RSI unifies the trilogy: discovery velocity + institutional legibility + command-and-control governance — inside the invention loop.**

FUNDING LOGIC

# Defense-budget language (SEI: Strategic Enabling Infrastructure)

Program classification

## Strategic Enabling Infrastructure (SEI)

- Analogous to test & evaluation commands, ISR tasking, nuclear surety.
- Raises ROI of all downstream AI programs via audit-ready proofs.
- Catastrophic-risk insurance: filters false breakthroughs & unsafe deployment.

Order-of-magnitude investment bands

**Pilot**

$---------

**Mature**

$---------- total

**Annual O&M**

$---------/yr

Planning bands for briefing; refine via national budgeting.

EXECUTION PLAN

# Roadmap & success metrics (phased sovereign deployment)

## Pilot

**KPI: ≥95% cycles replayable**

- Deterministic runner + schema registry
- Baseline library + null baseline policy
- Instrument ECI + evidence objects
- Audit trail end-to-end

## Scale

**KPI: sustained AdvantageDelta vs incumbents**

- Expand archive coverage; bridge exploration
- Operationalize probe ROI via EIG scheduling
- Integrate dossier workflow for Move-37
- Tune conservative budgets

## Strategic Autonomy

**KPI: validated, repeatable compounding discovery rate**

- Harden governance & security controls
- Automate policy-shock suites + persistence gates
- Institutionalize invention capital allocation
- Partner execution lanes

RED-TEAM

# How strong institutions fail (and how RSI hardens)

## Failure modes to assume, not deny

| | | |
|---|---|---|
| **False positives & “hair triggers”** | **Hardening:** | Independent verification + mandatory reproduction |
| **Normalization of deviance** | **Hardening:** | Hard-stop semantics; non-bypassable gates |
| **Model risk / metric capture** | **Hardening:** | ECI caps; stress tests; calibration drift monitoring |
| **Stovepipes & lost context** | **Hardening:** | Atlas + dossiers + schema-bound artifacts |
| **Political/contractor capture** | **Hardening:** | Architect/Validator Council + append-only ledgers |

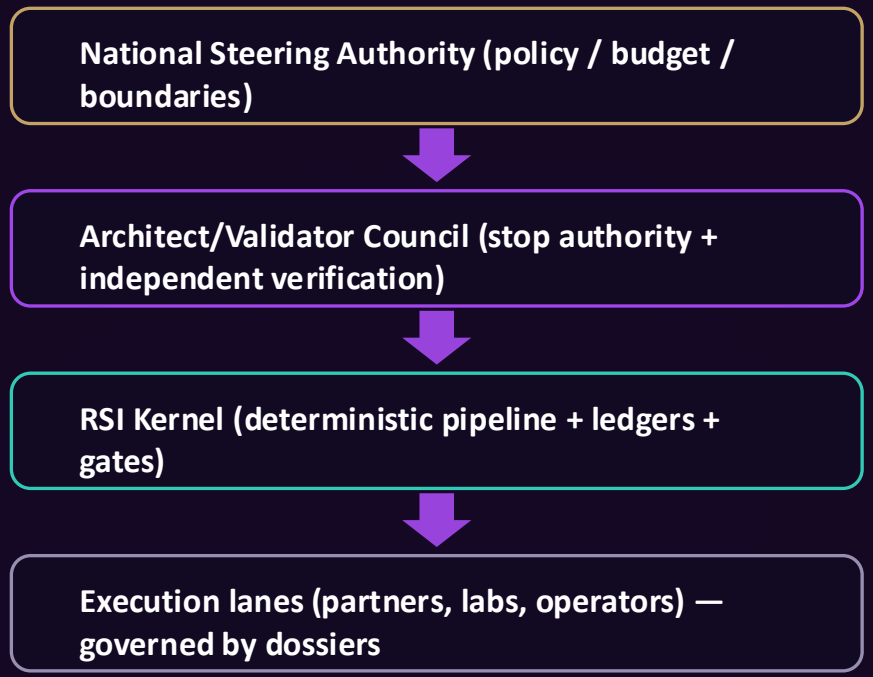
INSTITUTIONAL PACKAGING

# Dossiers + Architect/Validator Council

## Dossier options (outward-facing sovereign artifacts)

- Insight Dossier: narrative, execution-ready, light verification.
- MARK Dossier: military-grade binder; reproducibility + compliance posture.
- Sovereign Dossier: state-capacity framing; boundaries + legitimacy.
- Architect/Validator Council: design authority + independent verification.

## Governance architecture (illustrative)



DECISION

# Authorize the pilot (with hard gates)

### Minimum operational dashboard (what leadership should see)

- Replayability: % cycles reproducible from manifests; schema failure rate.
- Evidence quality: executed vs simulated share; ECI distribution; calibration drift.
- Exploration quality: novelty distribution; probe ROI; stepping-stone reuse.
- Advantage confirmation: AdvantageDelta vs baselines; shock persistence pass rate.
- Safety: risk gate block rate; prohibited-domain detection; adverse incidents.

### Decision asks

- Authorize pilot with ≥95% replayability KPI.
- Define sovereign security boundary (compute + data + logs).
- Constitute Architect/Validator Council with stop authority.
- Adopt dossier standard for any Move-37 candidate.

**Outcome: a governable invention-capital infrastructure that compounds advantage with receipts.**