AGI Alpha RSI

A Sovereign Innovation Operating System for Recursive Self-Improvement

PRESS + SOVEREIGN STRATEGY BRIEFING • HIGH-TRUST TECHNICAL SUMMARY

Document status	FINAL (Press-grade, institutional)
Prepared for	Global technology press; sovereign technology leadership; national innovation agencies
System baseline	Runner + Prompt Pack: rr_omni_v7 (Move-37 breakthrough protocol + EIG-scheduled probing)
Confidentiality	Shareable with attribution; remove internal deployment details for public release

Executive Summary

AGI Alpha RSI is a deterministic, auditable “innovation operating system” that uses high-capability LLMs as modular components-within strict schemas and replayable runners-to discover, test, and compound new strategies and innovations. It is engineered to produce non-obvious, high-leverage novelty while maintaining institutional standards of evidence, reproducibility, and governance.

What makes it different (in one page):

- Open-ended, Quality-Diversity search (MAP-Elites-style archive) that accumulates a portfolio of distinct innovations-rather than optimizing a single objective.
- An OMNI Interestingness Kernel that does more than score: it routes actions (PROBE/REFINE/INSERT/REPLACE/ESCALATE) and names the unknowns that matter.
- A deterministic Expected-Information-Gain (EIG) scheduler that converts “interestingness” into compounding search control-allocating probe budgets to the most informative tests first.
- Evidence as a first-class currency via ECI (Evidence Contact Index): credibility increases only through executed, checkable tests; simulated judgement is capped.
- Move-37 Breakthrough Protocol (v7): objective breakthrough detection, reproduction, stress-testing under policy shocks, persistence gating, and a decision-grade dossier bundle.

Why this matters for sovereign strategy

Advanced innovation is becoming an infrastructure problem: nations compete on the ability to continuously generate, validate, and operationalize new technical advantages. Traditional R&D struggles with three structural limits-slow feedback, fragmented evidence trails, and weak reproducibility at scale.

Strategic objective	AGI Alpha RSI capability
Speed of learning	Compress multi-month exploratory loops into repeatable cycles with deterministic audit artifacts.
Strategic autonomy	Operate as a sovereign capability: controlled deployment, data residency, and policy-aligned governance.
Proof-grade outputs	Every promoted innovation ships with an evidence ledger, reproducible manifests, and baseline-comparative advantage metrics.

What AGI Alpha RSI is

AGI Alpha RSI is not a single model. It is a rigorously-defined, prompt-pack-driven operating system for invention-where each LLM call is schema-bound, replayable, and audited; and where every claim is pressured toward contact with evidence through deterministic runners and micro-benchmarks.

The platform orchestrates a closed loop-Target → Emit → Filter → Atlas → Test-Plan → Eval → Insert → Promote-and persists state across cycles so the archive and causal atlas compound rather than reset.

Architecture at a glance

Three layers form the system boundary:

1) Prompt Pack (schema-bound agent roles)

2) Deterministic Runner (stage sequencing + evidence minting)

3) Persistent State (QD archive + causal atlas + ECI ledger)

Layer	Guarantee
Prompt Pack	A single JSON file mapping prompt_id → {system, user_template, output_schema_ref}. Every agent role is a strictly-typed interface; outputs are validated and repaired deterministically.
Runner Config	A single JSON file defining cycle stages, budgets,

	constraints, scoring formulas, routing rules, novelty distance spec, baseline comparison policy, and the Breakthrough Protocol.
Deterministic Runner	Executes one cycle at a time with temperature=0, fixed seed, canonical hashing and per-call provenance logs. Produces run_outputs.zip + state_for_next_run.json.
Persistent RSI State	Monotonic growth: cycle_index increments; archive.frontier_cells and archive.candidates append; atlas triples accumulate; ECI ledger records all evidence events (pre/post).

The recursive cycle (one deterministic pass)

TARGET	EMIT	FILTER	ATLAS	TEST-PLAN	EVAL	INSERT	PROMOTE
Select archive cells + bridge targets (coverage and cross-domain).	Generate candidates (LHF + Pioneer) + scaffold genomes/variants.	Risk gate + boringness + novelty report + OMNI interestingness decision routing.	Extract triples; complete mechanisms; contradiction and side-effects; bridge discovery; atlas patch.	Decision-conditioned falsification ladder; PROBE ladders target named unknowns.	Execute tests and episodes; baseline-comparative grading; mint evidence objects + ECI updates.	Insert/replace winners in the QD archive + HELM-style reporting + eval manifests.	Lane-aware promotion queue for pilots and strategic escalation.

Engineering systematic pressure toward non-human, high-leverage novelty

To reliably surface strategies that a human team would not naturally propose, AGI Alpha RSI uses open-ended search control rather than “one-shot ideation.” Novelty is not treated as style-it is a measurable distance from what has already been explored, coupled to an objective evaluation stack.

1) Quality-Diversity archive (MAP-Elites semantics)

Candidates are stored as a portfolio across an explicit descriptor space (e.g., capital intensity, time horizon, mechanism class, regulatory friction). The system seeks breadth-filling empty or sparse cells-so stepping-stones are preserved and recombined later, enabling compounding discovery.

Operational effect:

- Coverage pressure: target empty/sparse cells to avoid converging on a single local optimum.
- Pioneer pressure: allocate a dedicated lane to cross-domain bridges and mechanism novelty.
- Archive memory: the system remembers what has been tried and what worked, preventing repeated rediscovery.

2) Deterministic Novelty Distance (v7)

Every emitted candidate receives a deterministic novelty distance score in [0,1] before interestingness evaluation. This provides a stable novelty signal that is auditable and reproducible.

Spec ID	novelty_distance.v1
Neighbor pool	frontier_cells + recent candidates (max_neighbors=50)
Components	descriptor_sim (0.40) + triple_sim (0.30) + text_sim (0.30)
Composite similarity	0.40·descriptor_sim + 0.30·triple_sim + 0.30·text_sim
Novelty distance	novelty_distance = clamp(1 - max_neighbor(composite_sim), 0, 1)
Thresholds	High novelty ≥ 0.80; Breakthrough candidate ≥ 0.90

This novelty distance is persisted as a first-class artifact (candidates/novelty_distance.jsonl) and is copied verbatim into the OMNI Interestingness audit record, preventing score drift or retroactive rewriting.

3) OMNI Interestingness as an action router

AGI Alpha RSI operationalizes OMNI-style open-endedness by requiring the interestingness kernel (P63) to output not only scores, but a deterministic action recommendation: REJECT, PROBE, REFINE, INSERT, REPLACE, or ESCALATE. The output must also name the precise unknowns driving uncertainty-each with expected information gain.

Why this matters:

- It turns novelty into a policy: we systematically allocate effort to reducing uncertainty where it matters most.
- It prevents premature ‘archive insertion’ based on rhetoric-low confidence routes to PROBE/REFINE automatically.
- It creates a closed prompt ecosystem: every downstream step (test planning, probing, escalation) is conditioned on structured outputs.

4) EIG-scheduled probing for information-efficient search control (v6)

For PROBE candidates, falsification ladders must embed unknown_id in each test_id. The runner then selects which L0 rungs to execute using a deterministic Expected-Information-Gain scheduler that combines: (a) expected_info_gain from P63’s uncertainty focus, (b) the entropy of the rung’s forecasted outcome, and (c) an explicit cost bucket.

This makes probing measurable and optimizable: the probe/probe_schedule.jsonl artifact records the exact EIG score per executed rung, enabling budget governance and continuous improvement of the probing policy.

Engineering objective advantage confirmation

Breakthrough claims are only meaningful if they outperform a baseline under controlled conditions. AGI Alpha RSI therefore treats evaluation as a comparative measurement problem, not a narrative problem.

Baseline comparison requirements (v7)

Every deep evaluation is baseline-comparative. The baseline comparator is chosen deterministically: (1) incumbent elite in the same archive cell, else (2) nearest neighbor used for novelty distance, else (3) a fixed null baseline.

Advantage metric: grade.overall_score (default: overall rubric score).

Replace threshold (minimum delta): 0.05

Breakthrough threshold (minimum delta): 0.15

The runner emits eval/baseline_comparison.jsonl containing candidate score, baseline score, delta, and whether replacement/breakthrough thresholds are met.

Evidence Contact Index (ECI): credibility as currency

ECI is a ledgered credibility score in [0,1] that moves only when tests produce evidence. Executed evidence (deterministic checks, micro-benchmarks, controlled episodes) increases or decreases ECI with larger steps; simulated or judgement-only evidence is capped. This prevents ‘self-delusion loops’ and forces contact with reality.

Institutional controls:

- Every evaluation mints an evidence object with hashes and provenance; summaries never substitute for raw artifacts.
- At least one executed evidence event per cycle is mandatory; the system cannot drift into purely simulated RSI.
- Judge outputs are audited and periodically recalibrated against anchors to mitigate drift.

Determinism, auditability, and replay

All model invocations are constrained to temperature=0 with canonical JSON hashing. Each cycle emits a complete audit trail: stage logs, manifests, artifacts, and a state bundle. This makes reproduction a routine operation: re-run with the same state and inputs; verify output hashes; and compare advantage deltas.

Move-37 Breakthrough Protocol (v7)

A Move-37-class event is treated as a measurable outlier: a candidate that is both (a) far from known solutions (high novelty distance) and (b) objectively advantaged versus a baseline. When those conditions hold, the system enters a controlled escalation path that produces a decision-grade dossier with no ambiguity.

NoveltyDistance ≥	0.90

AdvantageDelta ≥	0.15
Risk ≤	max_risk_score (config bound)
Confidence ≥	0.55
ECI ≥	0.55
Status	breakthrough_candidate=true

Protocol steps (same-cycle escalation)

Reproduce - Re-run candidate and baseline on the same eval battery under multiple fixed seeds; verify hashes and manifests.

Stress-test - Apply policy shocks (controlled perturbations) and re-evaluate candidate vs baseline; log advantage deltas per shock.

Persistence gate - Confirm advantage persists: positive delta above a minimum threshold on ≥ required pass rate across shocks.

Package dossier - Emit a dossier bundle: run_manifest, novelty distance log, baseline comparisons, evidence objects, ECI ledger, HELM-style summary, and reproduction manifests.

Mandatory ADVANTAGE_PERSISTENCE unknown for high novelty

If novelty_distance is high (≥ 0.80 by default), the system must explicitly track an uncertainty item with unknown_id = ADVANTAGE_PERSISTENCE. This forces test planning to include at least one L0 rung that directly probes whether the measured advantage persists under shocks or baseline variations.

Outcome: decision-grade breakthrough dossier

A confirmed breakthrough is not an anecdote. It is a bundle of artifacts that allows an independent team to replay the event, validate the advantage, and assess persistence under stress. For sovereign stakeholders, this reduces ‘innovation risk’ from a story to a governed evidentiary process.

Sovereign deployment, control, and governance

AGI Alpha RSI is designed as an institutional platform: its core value is not only ideation, but the ability to prove, reproduce, and govern innovation outcomes. Sovereign stakeholders can operate it as a controlled capability aligned with national priorities and regulatory constraints.

Deployment modes

- Sovereign enclave: operate entirely within national infrastructure (on-prem or sovereign cloud), with strict data residency.
- Air-gapped / classified: run without external network access; executed evidence relies on deterministic micro-benches and controlled simulations.

- Hybrid: controlled external evaluation sandboxes while keeping state, logs, and dossiers in the sovereign boundary.

Trust primitives

- Deterministic manifests: every cycle produces run_outputs.zip and a state bundle for continuation.
- Cryptographic hashing: canonical JSON digests + content hashes enable tamper-evident audit trails.
- Schema governance: every LLM interaction is constrained by explicit input/output schemas; invalid outputs are repaired or rejected.
- Evidence policy: credibility rises primarily through executed checks; simulated evidence is capped.

Persistence invariants (no silent resets)

The runner enforces monotonic state growth across runs. If a run would violate an invariant (e.g., cycle_index not incrementing by +1, or archive tables shrinking), it must halt and emit an error object rather than silently reset.

- state_manifest.cycle_index increments exactly +1 per run.
- archive.frontier_cells and archive.candidates are non-decreasing (append-only).
- Stable IDs are never recycled (candidate_id, scaffold_id, cell_key).
- ECI ledger is append-only; each event records pre/post values and evidence provenance.

Operational governance

The system supports portfolio governance via lanes (LHF vs Pioneer), budgets, risk gates, and promotion policy. This enables leadership to tune the exploration/exploitation balance while preserving proof-grade standards.

Capabilities and differentiation

What this platform does that typical AI systems do not

Category	AGI Alpha RSI
Not a chatbot	A deterministic invention pipeline with explicit stages, budgets, and artifact outputs.
Not a single benchmark	A compounding portfolio in a QD archive-diversity is the point, not an afterthought.
Not 'LLM-judge only'	Executed evidence is mandatory; simulated evidence is capped; audit artifacts are preserved.
Not 'one-shot ideation'	Action-routing + EIG-scheduled probing turns uncertainty into systematic search control.
Not hype-driven	Breakthrough Protocol produces reproduction and stress-test bundles before escalation.

Representative application domains (config-tunable)

The platform is domain-agnostic: it learns and tests within controlled environments defined by your priority list, constraints, and evidence policies. Example domains include:

- Critical infrastructure operations (energy, transport, logistics)
- Public finance and procurement (audit-grade process innovation)
- Industrial policy and manufacturing optimization
- Public-sector service delivery modernization
- Scientific R&D portfolio management (hypothesis generation + falsification ladders)

What stakeholders receive per cycle

Every cycle emits decision-ready artifacts, including:

- run_outputs.zip with complete per-stage logs, candidates, tests, evidence objects, ECI ledger, and reports
- state_for_next_run.json to continue compounding search without reset
- promotion_queue (JSONL + CSV) for leadership decisions
- optional dossier bundles for any breakthrough candidates

Adoption pathway

AGI Alpha RSI can be adopted as an institutional capability with clear gates and measurable outputs. A typical rollout proceeds in three controlled phases:

Phase I - Sovereign sandbox: Deploy the deterministic runner and prompt pack within a controlled environment; configure constraints, focus domains, and evidence micro-benches; validate audit trail end-to-end.

Phase II - Portfolio compounding: Run multi-cycle operations so the archive and causal atlas grow; tune lane budgets and promotion policy; establish judge calibration cadence and reporting standards.

Phase III - Operational integration: Integrate with existing R&D governance: promotion queue → pilots; dossier bundles → decision committees; establish continuous monitoring, red-teaming, and versioned model/prompt governance.

Governance commitments (for high-trust deployment)

- Model and prompt versioning with signed manifests (no silent changes).
- Reproducibility as a policy: any promoted result must be replayable from state + hashes.
- Risk and compliance gating before insertion/promotion; constraints are explicit and auditable.
- Separation of duties: generation, grading, and judge auditing are modular roles with logged provenance.

Appendix A - Technical baseline (rr_omni_v7)

The current system baseline integrates: action-routing interestingness (P63), uncertainty-conditioned falsification ladder generation (P51), EIG-scheduled probe execution (v6), and the Move-37 Breakthrough Protocol (v7) with deterministic novelty distance and baseline comparisons.

Key v7 additions

- Deterministic novelty distance computation (novelty_distance.v1) logged per candidate.
- Baseline comparator selection for every deep evaluation; advantage deltas logged deterministically.
- Breakthrough trigger thresholds + extra evaluation budgets (multi-seed replay + policy shocks).
- Mandatory ADVANTAGE_PERSISTENCE unknown injected for high-novelty candidates; must appear in probe ladders.
- Dossier bundle that packages evidence, ledgers, manifests, and HELM-style reporting for decision makers.

Artifact outputs (minimum set)

- candidates/novelty_distance.jsonl
- eval/baseline_comparison.jsonl
- probe/probe_schedule.jsonl
- evidence/evidence_objects.jsonl
- eci/eci_ledger.jsonl
- reports/helm_like_summary.md
- dossier/index.json (when breakthrough protocol triggers)

For implementation reference, see the attached runner configuration and prompt pack that define these behaviors as mechanical, schema-bound requirements.

Contact and next steps

For press inquiries, sovereign briefings, or technical due diligence, the recommended starting point is a controlled demonstration run in a sovereign sandbox, producing a full run_outputs.zip and state_for_next_run.json for independent replay and verification.

- End of document -