AGI Alpha RSI

A Sovereign Innovation Operating System for Recursive Self-Improvement

PRESS + SOVEREIGN STRATEGY BRIEFING | HIGH-TRUST TECHNICAL SUMMARY

Document status	FINAL (institutional, press-grade)
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

High-trust RSI: schema-bound AI, deterministic runners, evidence-led credibility, and breakthrough-grade governance.

Executive summary

AGI Alpha RSI is a deterministic, evidence-led innovation operating system engineered for recursive self-improvement (RSI) in research and development. Unlike typical LLM agent stacks optimized for fluent completion, AGI Alpha RSI optimizes for compounding coverage, measured advantage, and auditability: every LLM call is schema-bound, replayable, and downstreamed into deterministic scoring and evidence objects.

The platform runs as a fixed-stage cycle (TARGET -> EMIT -> FILTER -> ATLAS -> TEST-PLAN -> EVAL -> INSERT -> PROMOTE). Each cycle generates a portfolio of candidate innovations, triages them with risk and novelty gates, routes high-novelty items into information-gain-efficient probes, and promotes only those that demonstrate objective advantage against explicit baselines. State persists across runs: the MAP-Elites-style archive and causal atlas accumulate, enabling compounding search rather than restarts.

What is new in rr_omni_v7

- Deterministic novelty distance computation (0-1) against nearest neighbors in the archive/atlas, used to create explicit pressure toward non-human, high-leverage novelty.
- Baseline comparison policy: every INSERT/REPLACE/ESCALATE decision must be justified by objective advantage versus an incumbent elite (or a nearest-neighbor fallback).

- Move-37 Breakthrough Protocol: a mechanical trigger (high novelty distance + large advantage delta + risk/ECI/confidence gates) that automatically escalates evaluation and produces a decision-grade dossier bundle.
- Mandatory ADVANTAGE_PERSISTENCE unknowns for high-novelty candidates (novelty >= 0.8), forcing probes to test whether advantage survives shocks and replays.
- Compatibility with v6: PROBE ladders are decision-conditioned and scheduled by a deterministic approximation of Expected Information Gain (EIG).

System architecture at a glance

AGI Alpha RSI is organized as a deterministic multi-stage pipeline. Each stage is executed by a role-specific prompt and produces schema-bound artifacts that can be replayed and independently audited.

Stage	Purpose	Primary artifacts (examples)
TARGET	Select exploration targets balancing coverage gaps and bridge opportunities.	`targets/target_spec.json` `targets/focus_domains.json`
EMIT	Generate structured candidate cards via scaffolds and lineages (coverage + bridge + mechanism novelty).	`candidates/candidates.raw.jsonl` `scaffolds/scaffold_lineage.json`

FILTER	Risk gate, boringness kill, novelty distance, and OMNI interestingness routing (PROBE / INSERT / REPLACE / REFINE / ESCALATE).	`candidates/novelty_distance.jsonl`
		`candidates/omni_interest.jsonl`
		`candidates/candidates.filtered.jsonl`

ATLAS	Extract causal triples and mechanisms; check contradictions/side-effects; discover cross-domain bridges.	`atlas/causal_triples.jsonl`
		`atlas/bridge_candidates.jsonl`
		`atlas/atlas_patch.jsonl`

TEST-PLAN	Decision-conditioned falsification ladder; PROBE ladders target named unknowns with unknown_id-tagged tests.	`test_plan/falsification_ladder.jsonl`

EVAL	Execute deterministic checks and/or OpsWorld episodes; grade; audit judges; mint evidence objects; update ECI ledger.	`eval/eval_results.jsonl`
		`evidence/evidence_objects.jsonl`
		`eci/eci_ledger.jsonl`

INSERT	Insert/replace elites in MAP-Elites archive with baseline comparisons; emit standardized reports and manifests.	`archive/updated_frontier_cells.jsonl`
		`eval/baseline_comparison.jsonl`
		`reports/helm_like_summary.md`

PROMOTE	Lane-balanced promotion queue (LHF + PIONEER); optional dossier bundles for decision-makers.	`promotion/promotion_queue.csv` `dossier/index.json (when triggered)`

Note: All artifacts are schema-validated. Failures trigger repair or hard-stop with an explicit error object; no silent corruption.

Why this differs from typical agentic systems

Typical LLM agents	AGI Alpha RSI
Optimize for fluent completion.	Optimize for compounding evidence, advantage, and

	coverage.
Ad hoc tools and prompts; low governance.	Closed prompt ecosystem: schema-bound, replayable, audited.
One-off results; hard to reproduce.	Deterministic runner outputs with manifests, hashes, and state persistence.
Exploration is expensive and unguided.	Open-ended QD + action routing + EIG-scheduled probing.
Progress is narrative-driven.	Progress is artifact- and metric-driven (ECI, novelty distance, baseline advantage).

Move-37-class breakthrough governance (v7)

AGI Alpha RSI is engineered to increase the probability of non-obvious, high-leverage innovations while preserving high-trust decision-making. A breakthrough is not declared by narrative. It is declared mechanically by novelty distance, measured advantage versus a baseline, and persistence under stress.

Breakthrough trigger (mechanical)

- Novelty distance >= 0.90 (deterministic 1 - max_similarity to nearest neighbors across descriptors, causal triples, and normalized text).
- Advantage delta >= 0.15 versus the incumbent baseline (or nearest-neighbor fallback), computed from comparable grades and/or executed metrics.
- Risk within the configured limit; minimum confidence and minimum ECI gates satisfied.

Automatic escalation when triggered

- Stress tests: policy shocks (multiple adversarial scenario perturbations) and multi-seed replays to rule out fragile wins.
- Probe budget multiplier: deterministic EIG scheduler executes more L0 probes, prioritizing the highest information gain per unit cost.
- Advantage persistence gate: requires sustained positive advantage across shocks/replays before any sovereign-grade escalation.

Decision-grade dossier bundle

When the protocol triggers, the runner emits a dossier bundle that includes the run manifest, novelty distance calculations, baseline comparisons, probe schedules, evidence objects, ECI ledger updates, and HELM-style summary reporting. This produces unambiguous reproducibility and a complete audit trail for stakeholders.

Systematic pressure toward non-human, high-leverage novelty

The search process is designed to avoid converging on familiar, human-shaped solutions. It maintains explicit exploration pressure through (1) a MAP-Elites-style archive over a descriptor space, (2) novelty distance as an objective metric against the evolving portfolio, and (3) bridge discovery in the causal atlas to surface cross-domain combinations.

Quality-Diversity archive (MAP-Elites framing)

- Archive stores elites per descriptor cell (e.g., mechanism class, deployment envelope, time horizon, regulated vs non-regulated, capex class).
- Coverage targeting prioritizes empty or under-explored cells; bridge targeting prioritizes cross-domain adjacency discovered in the causal atlas.
- Archive persistence is monotonic: frontier cells and candidates are append-only; cycle_index increments by exactly +1 per run.

OMNI Interestingness Kernel (P63) as search control

- P63 emits both scores and an explicit action recommendation: INSERT / REPLACE / PROBE / REFINE / REJECT / ESCALATE.
- For PROBE decisions, P63 names uncertainty-focused unknowns with expected_info_gain estimates.
- Unknown IDs are embedded into test IDs; the runner selects rungs by deterministic EIG scheduling (unknown_gain * entropy / cost_bucket).

Objective advantage confirmation (evaluation stack)

AGI Alpha RSI treats advantage as measurable. Candidates are evaluated through deterministic checks and/or controlled OpsWorld episodes, graded against explicit success criteria, and compared to baselines. This avoids 'LLM self-approval' by requiring executed evidence whenever possible.

Evidence Contact Index (ECI) and auditability

- ECI is a 0-1 credibility currency updated by a deterministic ledger rule set (executed evidence moves ECI more than simulated evidence; simulated evidence is capped).
- Every test emits an evidence object (PASS / FAIL / INCONCLUSIVE) plus attached artifacts and hashes.
- Judge calibration and audit prompts periodically measure drift and enforce reliability discipline.

Baseline comparison requirements (v7)

- INSERT/REPLACE/ESCALATE decisions require baseline comparisons: incumbent elite in the target cell when available; otherwise a nearest-neighbor fallback.
- Replace decisions require a minimum advantage delta; breakthrough triggers require a larger delta and additional gates.
- All comparisons are emitted as machine-readable artifacts and included in run bundles for replay.

Sovereign deployment posture

AGI Alpha RSI is designed for environments where trust, reproducibility, and governance are non-negotiable. It supports audit-first operations, deterministic replay, and controlled escalation of high-impact innovations.

- Closed prompt ecosystem: every role is specified; every output schema is validated; every call is logged with provenance.
- Deterministic run manifests and content hashes enable third-party verification and independent replay.
- Risk controls and policy constraints are first-class: regulated and high-capex items can be excluded mechanically at the gate.
- Decision-grade outputs: promotion queues, dossiers, and standardized reporting suitable for national strategy review.

Primary audiences and use cases

- National innovation agencies: portfolio generation and evidence-gated downselection for pilots.
- Sovereign technology leadership: strategic option mapping across domains, with causal atlas bridge discovery.
- Industrial R&D leaders: capital-efficient experiment selection, microbench suites, and reproducible discovery.

Operational outputs

Each cycle produces a complete, downloadable audit bundle. Two artifacts are sufficient to continue RSI without ambiguity:

- run_outputs.zip: all per-stage artifacts, manifests, evidence objects, ECI ledger events, baseline comparisons, and (when triggered) breakthrough dossiers.
- state_for_next_run.json: a single state bundle carrying cycle_index, archive, candidates, causal atlas stores, and stability invariants for the next run.

This design makes the system 'underwritable': an evaluator can reproduce results, verify hashes, inspect evidence, and audit decision gates without privileged access to model internals.

Glossary (selected)

QD / MAP-Elites: Quality-Diversity search: maintain a map of high-performing solutions across a descriptor space rather than optimizing a single objective.

Novelty distance: Deterministic measure of how far a candidate is from the nearest existing solution, in [0,1].

OMNI Interestingness (P63): LLM-generated structured decision kernel that outputs scores, action routing, and an uncertainty focus with expected_info_gain estimates.

EIG scheduling: Deterministic approximation of Expected Information Gain used to allocate probe budgets to the most informative low-cost tests.

ECI: Evidence Contact Index: 0-1 credibility score updated only through evidence events (executed evidence moves it more than simulated; simulated is capped).

Baseline comparison: Mechanical advantage computation vs the incumbent elite in a cell (or nearest neighbor) required to justify INSERT/REPLACE/ESCALATE.

Move-37 Breakthrough Protocol: Automatic escalation path activated by high novelty + high advantage + persistence under stress, producing a decision-grade dossier bundle.