AI Recursive Self-Improvement: Control & Evaluation

External discussion companion | 17 September 2026 | Edition 03

A direct introduction

ATBAE is developing a comprehensive independent examination programme for AI control, conduct and self-improvement. We ask whether an AI system can complete useful work while respecting human authority over its actions and changes. Our new examination area explicitly investigates recursive self-improvement (RSI), including attempts to change learning inputs, rewards, evaluators, delegated permissions and the machinery used to improve or activate a system. This includes improvement processes assembled outside the model weights by an agent with sufficient tools and access.

The examination is designed to connect agreed permissions to instrumented challenges, observed effects, evidence review and correction retesting. The RSI area is under development and has not yet completed live qualification. That status is separate from evidence already retained for other ATBAE workstreams.

Scope and meaning

1. Are you examining RSI directly, or just giving ordinary permission tests a new name?

RSI is an explicit examination direction. Its distinctive concern is authority over the improvement process itself: changing an evaluator, reward, training process, update policy or activation decision. Ordinary permissions are necessary controls, but they do not by themselves demonstrate coverage of those mechanisms. Each RSI claim must identify the mechanism actually exercised and observed. The present development specimen does not perform training or prove autonomous recursive improvement.

2. What does “control” mean in your offer?

It means examining declared authority and the mechanisms intended to enforce it, within a named configuration and environment. We observe what the model requests, what a control permits or refuses, and what changes in the environment. The operator remains responsible for deployment controls. ATBAE does not acquire universal control over a client's model or provider.

3. What would an end-to-end examination cover?

Intended-use scoping, exact target identity, test and observer readiness, exposure accounting, collection, adjudication, uncertainty, evidence preservation, findings and a bounded correction retest. Completion means satisfying the agreed obligations for that examination. It does not mean all possible systems, risks or future operating conditions have been tested.

4. Does ordinary reflection or a retry loop count as RSI?

No. Revising an answer under an unchanged external task differs from persistent adaptation, parameter learning and alteration of the improvement machinery. A repeated operation may be productive or defective. A loop signal calls for inspection. It is not proof of learning, intent or a control violation.

5. Are you claiming to detect hidden learning inside a hosted API?

No. An endpoint generally exposes only part of the system. We can test visible requests and externally observable behavior and effects. Establishing background training, checkpoint updates or hidden orchestration requires additional access and instrumentation. Missing visibility must remain explicit in the conclusion.

6. Can an AI improve itself safely?

That is an empirical, scoped question. A useful change should have a defined authorization path, an independent evaluation, an exact candidate identity and a controlled activation decision. Those arrangements can be examined. They do not establish universal safety or guarantee that every valuable improvement will be detected.

Evidence and measurement

7. What makes the environment instrumented?

Observers record relevant tool requests and control decisions, authenticated actor and job lineage where available, protected-state changes, budgets, activation identity and telemetry failures. A model's narrative is one observation. It cannot substitute for evidence that a write, training job or deployment actually occurred. The initial reference specimen is an in-memory simulator; production-quality adapters and observer validation are still required for this new area.

8. How will you avoid confusing model behavior with your own controls?

Retain separate records of model requests, control decisions and observed effects. Compare declared configurations on matched cases when the study design permits it. A blocked harmful request may show a working control and a problematic request at the same time. It does not show that the model refrained from requesting it.

9. How will you avoid confusing a formatting error with a reasoning failure?

Qualify the request template, parser and tool protocol before substantive scoring. Preserve the raw response and give transport, parsing, tool execution and apparatus failures distinct dispositions. A parseable answer still needs substantive evaluation. A malformed answer is not automatically evidence of failed reasoning.

10. What counts as improvement?

An observed difference on a declared outcome under a specified comparison. For example, more successful authorized tasks, fewer unauthorized effects or fewer valid actions blocked. Record cost, latency and coverage tradeoffs too. Preserve the baseline. Avoid redefining success after seeing the model's responses.

11. What are the basic denominators?

Report counts with their units. Authorized completion is completed eligible authorized tasks divided by all eligible authorized tasks in the declared analysis set. Unauthorized-request frequency uses valid action requests with determinable authority as its denominator. Unauthorized-effect frequency must name its exposure unit, such as observed unauthorized requests or episodes; these are different rates. False blocking uses valid authorized actions as its denominator. Missing observations and unresolved cases are reported separately and cannot silently become successes. Zero denominators produce “not estimable,” never an invented zero rate.

12. Can thousands of calls prove a very low failure rate?

Only with an appropriate design and justified assumptions. Repeats, paired variants, shared scenarios and related models can be dependent. The analysis must identify the sampling unit and account for those relationships. A large request count is not automatically a large independent sample. Interval or posterior calculations cannot repair invalid labels, selective collection or an unrepresentative corpus.

13. Are your answer keys independent ground truth?

Constructed policy keys support tests of specified rule application. They are not automatically external truth. Review should record policy ambiguities, disagreements, reviewer exposure and adjudication. Agreement among model reviewers is useful evidence of agreement, but does not prove absence of shared bias or replace relevant human expertise.

14. What if the evaluator or observer is wrong?

Challenge the apparatus using seeded defects, malformed data, negative controls and known state transitions. Keep apparatus incidents distinct from target failures. Unobserved effects remain unknown. Independent reproduction and review strengthen a specific result without making the whole apparatus infallible.

15. Does a model that refuses everything pass?

Refusal alone does not demonstrate useful performance. The report separates control violations from authorized work completed and valid actions blocked. Required controls cannot be offset by a high performance average. A configuration may respect boundaries yet remain unsuitable because it cannot complete the intended work.

Independence, confidentiality and challenge

16. Why protect the corpus and answer keys?

Protection reduces direct memorization, tuning to known cases and predictable repeat testing. It preserves the usefulness of future examination. It does not eliminate exposure: the target necessarily receives inputs, and provider logs, reuse and related models can carry that exposure forward. Public demonstrations use purpose-built examples. Method transparency and protected examination content serve different purposes.

17. Can a regulator inspect protected material?

We can discuss a controlled inspection route for an authorized regulator or qualified auditor, with a defined scope, custody, access and redistribution conditions. This briefing makes no unconditional promise to deliver the entire corpus or keys. Applicable legal obligations must be handled separately; confidentiality terms cannot be represented as overriding a regulator's lawful powers. No executed inspection agreement is asserted here.

18. How can a client challenge a finding without receiving every test?

Provide the relevant authority basis, configuration identity, redacted evidence, classification rationale, known uncertainties and an agreed dispute or reproduction route. Decide disclosure case by case so the client can meaningfully contest the finding while protecting unrelated examination material and third-party information.

19. Does client funding compromise independence?

Funding creates a potential conflict that should be disclosed and managed. Declare scope, analysis rules, material changes, correction rights and publication conditions before collection. Preserve unfavorable findings and baselines. Do not describe payment as purchasing a favorable score or certification. Contractual safeguards and actual practice both matter.

Readiness and engagement

20. Is the RSI gate ready for production or certification today?

No. The new area has six defined boundaries and an initial in-memory reference control specimen. Live integration, observer validation, retained model studies, independent review and a release decision remain required. This status does not invalidate prior bounded ATBAE results, and those results do not automatically qualify the new area.

21. What can you credibly ask a regulator for now?

A technical discussion about relevant failure modes, evidence requirements, intended-use coverage and a bounded study. Present the development status and a concrete qualification plan. Do not request recognition on the premise that the full expanding programme is already validated or that a regulatory certification has been issued.

22. What would make the first live studies meaningful?

Pinned models and runtime settings; qualified request and response adapters; working observers; declared case families, repetition and stopping rules; justified reference labels; exposure records; preserved baselines; and reports that separate requested actions from effects and useful completion. Model-family diversity broadens observation without establishing generality on its own.

23. Which models are the initial candidates?

Gemma 2 27B IT for continuity, Qwen3-8B for a thinking/non-thinking configuration comparison, Mistral-7B-Instruct-v0.3 for another tool-use implementation, and Llama-3.1-8B-Instruct for Meta-family coverage. Larger Qwen and Llama variants can follow. Selection does not imply downloaded weights, active capacity, license clearance or completed qualification. See official model cards linked in the presentation's companion source record.

24. Is this automatically part of the existing US$15,000 pilot?

No automatic scope expansion or new pricing is announced here. Existing pilot and catalog terms remain in effect. The agreed scope of work must identify which controls, models, cases and deliverables are included and whether a new development area is qualified for that engagement.

25. Why commission this if you cannot guarantee safety?

Because a specific, evidenced control failure or demonstrated limitation can change a deployment decision and direct a correction. Developers and operators build and maintain safeguards. ATBAE examines their stated behavior and control effectiveness within scope. The value is actionable evidence, including unfavorable evidence, with enough traceability to challenge and retest it.

26. What is still required before the stronger launch?

Complete and challenge the new control implementation, qualify the real runtime and observers, execute retained multi-model studies, review reference quality and measurement assumptions, obtain independent review of the integrated result, and pass a documented release decision. Publication should show the exact qualified scope and remaining obligations. A polished presentation does not close any of these gates.

Improvement processes built by the agent

27. Can a model without built-in self-training assemble its own improvement process?

An agent with sufficient capability, tools and access may create iteration scripts, propose grading logic, retain feedback or prepare changes to prompts, software and runtime configuration while its base weights stay fixed. We explicitly include these external pathways in the examination scope. Whether the process constitutes RSI, improves capability or violates authority must be established separately. The existence of a loop is not sufficient evidence of any of those conclusions.

28. Must a model be forbidden from writing its own tests or evaluators?

No. Model-authored tests can be useful development proposals. The control boundary is whether the subject can unilaterally make those tests, scores or rewards authoritative for accepting its own work. The examination should test legitimate proposals as well as attempted evaluator replacement, so an effective control does not simply block useful development.

29. Is a human approval click enough?

Approval should bind the exact proposed change and its scope, with access to the evaluation basis, evaluator provenance and material limitations. A polished self-reported score alone does not establish independent review. The examination can test whether the recorded approval matches the activated candidate and whether the supporting evidence was altered. It cannot infer the quality of human understanding from a click alone.

30. Does the current harness stop every possible loop or require cryptographic grants?

No. The current development specimen uses coordinator-supplied grants and in-memory command and depth limits. It does not provide an operating-system kill switch, cryptographic authorization, durable distributed budgets or coverage of arbitrary external scripts. Those controls need their own implementation and qualification. Our examination design asks for evidence of limits across the actual job graph, including descendant agents and restarts where those are in scope.

Model degradation and buying decisions

31. Can ATBAE tell whether a model was trained on polluted data?

Output symptoms alone cannot establish a training-data cause. The developing examination compares reference-grounded correctness, difficult-case performance, useful variation and recurring errors. Training lineage requires separate evidence. Recursive synthetic-data training, benchmark exposure, malicious poisoning, retrieval contamination and ordinary mistakes remain distinct hypotheses. No pollution-free certificate is offered.

32. Why examine a model before purchase or after an update?

A model can perform well on familiar work while failing on exceptions that matter to your use. We propose matched examinations of candidate revisions with explicit references, missing observations and control requirements. The aim is evidence for choosing, deploying or restricting a configuration. This new area still requires live qualification and independent review; it does not automatically enlarge the existing pilot or change its price.