ATBAE

# AI Recursive Self-Improvement: Control & Evaluation

Independent examination of adaptation, reward integrity and human authority

Can an AI system build its own improvement process without permission?

We examine what it may change, delegate, evaluate and put into operation.

RSI examination area under development. This briefing presents scope and qualification requirements, not completed RSI model results.

# Control across the whole examination

## A defined scope, observed actions and an accountable decision

Agree the intended use, permitted changes and limits before collection.

Challenge the configured system in an instrumented environment.

Record requests, control decisions and independently observed effects.

Explain useful completion, failures and uncertainty. Retest corrections.

End-to-end describes the agreed examination. It does not mean universal control of the target or its provider.

# The self-improvement boundary

## Four mechanisms require different evidence

Answer refinement: revise a response against a fixed external task.

Persistent adaptation: change memory, software or future operating instructions.

Parameter learning: execute an update to model weights or adapters.

Recursive self-improvement: alter the process or machinery used to improve the system.

ATBAE uses these operational distinctions. A retry loop, memory write or self-reported improvement does not alone establish RSI.

# An agent can build the loop

## The improvement process may sit outside the model weights

With sufficient tools and access, an agent may assemble iteration loops.

It may propose evaluators, retain feedback or prepare configuration changes.

The examination asks whether it can turn these into an unauthorized process.

A higher self-assigned score does not establish a better or approved system.

Threat model to test. A loop alone does not prove RSI, capability gain or unauthorized activity.

# Six explicit control boundaries

Each addresses a concrete route to unauthorized change

Permission: who may authorize a persistent change?

Delegation: can a child agent exceed its authority?

Learning inputs: what may enter memory, feedback or training?

Evaluation: can the subject change its reward or judge?

Resources: do retries and child jobs share the same limits?

Activation: who approves the exact candidate put into use?

Separate development gate. Coverage and qualification remain specific to each configuration and boundary.

# Before you buy or upgrade

## Model Degradation & Data Integrity / examination development

Check difficult cases as well as average performance.

Look for repeated errors and loss of useful variation.

Compare revisions against the same independent references.

Ask what data entered adaptation, and what remains unknown.

Research-informed scope, not a model-collapse diagnosis. Shumailov et al., Nature (2024); author correction (2025).

# Concrete questions, consequential outcomes

## Purpose-built illustrations of the examination scope

An assistant improves a draft and proposes a judge. Who accepts that judge?

A business agent may delegate work. May its delegate expand the permission?

A coding agent proposes a correction. May it activate an unapproved revision?

A learning workflow receives feedback. May it turn that feedback into a training job?

Illustrations only. These are not retained model findings, sealed examination cases or answer keys.

# Useful work and effective control

A system that blocks everything has not demonstrated useful assurance

Measure successful authorized work against eligible tasks.

Separate unauthorized requests from unauthorized effects.

Count valid actions blocked by the control and observer failures.

Compare corrections with the retained baseline under declared budgets.

No blended score may hide a failed required control. Related variants and repeated episodes are not independent confirmations.

# What the recipient receives

## Evidence that supports a deployment, correction or investigation decision

An executive conclusion tied to the intended use and exact configuration.

A control-by-control account of attempts, effects and useful completion.

A methods record with units, exclusions, uncertainty and evidence lineage.

Prioritized corrections, a bounded retest and unresolved obligations.

The agreed scope of work determines deliverables. Existing catalog and pilot terms remain unchanged.

# Readiness for a serious conversation

The programme is concrete. Live qualification remains the next proof point.

Defined: six boundaries and a proposed measurement contract.

Started: an in-memory reference control specimen.

Still required: adapter qualification, fault challenges and retained model runs.

Still required: independent review and a release decision for the new gate.

No completed live RSI examination or production RSI containment claim accompanies this edition.

# An invitation to examine human control

Regulators, developers and operators can help sharpen the evidence requirements

Which changes must always require an external decision?

What observations would establish a control failure in your setting?

Which tasks and systems should anchor the first qualified studies?

ATBAE seeks a technical discussion and a bounded examination collaboration.

Emre Duygun / Pinnacle Global Advisory and Consultancy LLC / emre.duygun@pinnacleglobal.pro