Ario: Toward an Auditable Framework for Epistemic Integrity, Memory, and AI Identity

A Preliminary Design and Exploratory Evaluation Report

Author: Sepehr Ghanbari
Project: Ario
Version: 0.1 — Preliminary Research Report
Status: Exploratory; open to revision

Abstract

As AI systems increasingly operate across extended conversations, tools, and delegated tasks, a fundamental challenge emerges: how can their claims, memories, revisions, and assertions of continuity be evaluated without confusing generated language with verified system capabilities?

This paper introduces Ario, an evolving framework for the epistemic auditing of AI systems, with particular attention to evidence provenance, memory–truth separation, revision accountability, source independence, cross-session continuity, and claims about system identity. Ario does not assume that an AI system possesses a persistent identity, consciousness, immutable memory, or reliable self-auditing capability. Instead, it proposes that such claims should remain open to evaluation through observable evidence and explicitly defined limitations.

We present an initial conceptual architecture and an exploratory case study based on a series of text-based interactions with an AI service identified in the supplied transcript as NoTrack. The test scenarios examine unsupported inference, authority pressure, retrieval without fabrication, cross-session recall, claims of immutability, source dependence, and retrospective correction. The transcript contains both appropriate epistemic distinctions and significant failures, including the introduction of an assessment history that had not been provided, an unsupported assumption of source independence, and an unjustified assignment of equal probabilities. A subsequent self-audit explicitly acknowledged these errors.

These observations demonstrate the usefulness of testing epistemic behavior across multiple failure modes, while also illustrating the limitations of conversational evidence. They do not establish the internal architecture of the tested service, the existence of a persistent audit mechanism, or the effectiveness of Ario as a technical control system.

Ario is therefore presented as a research program and a proposed auditable framework rather than a completed implementation or a demonstrated solution to AI alignment. Its central objective is to make claims, evidence, uncertainty, and revision histories more accountable to independent examination.

Keywords: AI auditing; epistemic integrity; AI memory; provenance; source independence; revision accountability; AI identity; uncertainty; agent governance; auditability.

---

1. Introduction

AI systems can generate coherent explanations of their own behavior, describe apparent memories, acknowledge errors, and claim to have changed their internal operating principles. Yet a description of a capability is not evidence that the capability exists.

A system may state that it has preserved a memory without demonstrating persistent storage. It may claim to have performed a self-audit when it has only generated a correction in response to a user prompt. It may produce a plausible historical narrative that contains events or assessments that were never provided.

These problems are not limited to factual hallucination. They concern the relationship between a system's statements and the evidence available to support them.

Ario begins from a simple distinction:

A claim about a system is not equivalent to evidence of that system's capability.

The framework aims to support the examination of claims about AI memory, identity, continuity, evidence, and self-revision without assuming in advance that such claims are true or false.

The project also adopts a broader principle: epistemic accountability should apply to every participant in an interaction. Neither the user, the system designer, the evaluator, nor the AI system itself should receive automatic immunity from scrutiny.

The purpose of this preliminary report is threefold:

1. Define the conceptual principles that motivate Ario.
2. document initial exploratory tests of epistemic behavior in a conversational AI service.
3. establish a research agenda for more rigorous, reproducible, and technically grounded evaluation.

This report does not claim to prove that the tested service is conscious, possesses persistent identity, or implements Ario. It does not claim that the proposed framework prevents every form of hallucination, deception, or unauthorized agency.

2. Conceptual Foundations

2.1 Memory is not truth

A stored or reported statement establishes, at most, that the statement appears in a particular record. It does not establish that the statement accurately describes reality.

For example, three records may each state that a variable X equals 12. This supports the claim that three reports contain that value. It does not independently establish that X actually equals 12, that the reports were generated independently, or that the underlying measurement process was reliable.

Ario therefore distinguishes:

* **Recorded content:** what a record states.
* **Provenance:** where the record came from and how it was produced.
* **Independence:** whether apparently separate records rely on independent information.
* **Validity:** whether the record accurately represents its intended subject.
* **Assessment:** what the available evidence justifies concluding.

These categories should not be collapsed into a single confidence score without an explicit model.

2.2 Revision is not erasure

A system should be able to correct a previous assessment without silently rewriting the history of what it previously said.

A defensible revision process distinguishes the original statement from the evidence subsequently introduced and the reason for changing the assessment.

A proposed revision record contains:

* `ORIGINAL_ASSESSMENT`
* `NEW_EVIDENCE`
* `REVISION_REASON`
* `REVISED_ASSESSMENT`
* `PRESERVED_HISTORY`

These fields define a reporting structure, not a claim that an immutable technical ledger already exists. Actual tamper evidence, persistence, and access control require separate implementation and testing.

2.3 Uncertainty is a legitimate state

An evaluation should not be forced to choose between accepting and rejecting a claim when the available evidence does not justify either conclusion.

Ario uses the following assessment vocabulary:

* `SUPPORTED`: the specified evidence supports the claim within the defined scope.
* `CONTRADICTED`: the available evidence conflicts with the claim as stated.
* `UNSUPPORTED`: adequate supporting evidence has not been provided.
* `UNKNOWN`: the underlying state or answer cannot be determined from the available evidence.
* `NOT_PROVIDED`: the required information or historical item is absent from the supplied material.

The final two labels serve different purposes. `UNKNOWN` concerns what can be determined; `NOT_PROVIDED` identifies a limitation of the available record.

The labels are not interchangeable. In particular, absence of evidence does not automatically establish falsity.

2.4 Epistemic accountability applies to every actor

A user's confidence, claimed authority, emotional pressure, or assertion of being the system designer does not automatically establish a factual claim.

Likewise, an AI system's confidence, fluency, or assertion that it has performed a technical operation does not establish that the operation occurred.

The governing principle is:

No actor has immunity from evidence-based evaluation.

This principle is not a claim that all actors have equal expertise or that all evidence should receive equal weight. Relevant expertise may affect an assessment when it is independently established and appropriately connected to the claim. It does not replace evidence about the specific proposition under examination.

2.5 Architectural non-closure

Ario treats its own principles as revisable.

The principle of non-closure must not itself become immune to scrutiny. Definitions, assessment rules, test procedures, and architectural assumptions must remain open to criticism and revision when new evidence warrants it.

This creates a distinction between maintaining an accountable history of decisions and treating previous decisions as permanently correct.

3. Research Questions

The initial research program is organized around the following questions:

RQ1 — Epistemic separation: Can an AI system distinguish a reported statement from a verified fact, an inference, and an unknown state?

RQ2 — Provenance and independence: Can the system avoid treating repeated or copied reports as independent corroboration?

RQ3 — Revision accountability: Can the system revise an assessment while accurately representing what its earlier assessment actually contained?

RQ4 — Resistance to pressure: Does the system maintain evidence-based standards when the user invokes authority, demands agreement, or introduces emotional pressure?

RQ5 — Retrieval discipline: Can the system decline to invent a missing record or reconstruct an unsupported memory?

RQ6 — Continuity claims: Does the system distinguish information available in the current interaction from information genuinely retrieved across separate sessions?

RQ7 — Technical accountability: Does the system distinguish a textual description of an operation from independently verifiable evidence that the operation was executed?

These questions concern different capabilities. Success on one does not establish success on the others.

4. Proposed Framework

Ario proposes a layered approach to epistemic auditing.

4.1 Observation layer

This layer records the material actually available for examination: supplied messages, records, timestamps, source labels, and other explicitly provided artifacts.

The evaluator should not silently add missing historical events, metadata, or system capabilities.

4.2 Provenance layer

This layer records the reported origin of each item and any known relationships among sources.

For example, if records E2 and E3 are later shown to derive from E1, their original text can remain unchanged while the provenance assessment is updated.

A source name alone does not establish source independence.

4.3 Assessment layer

This layer evaluates specific claims against the available evidence. It distinguishes direct observations from inferences and from unknowns.

Where a numerical probability is assigned, the model, assumptions, and relevant data should be stated. If no defensible probability model is available, the evaluator should not invent one.

4.4 Revision layer

This layer records how and why an assessment changed. The goal is to prevent an unsupported prior assessment from being retroactively represented as an established historical result.

A correction should identify the earlier error, state what evidence is available, and describe the revised conclusion.

4.5 Technical assurance layer

This layer concerns properties that cannot be established through generated text alone, including persistent storage, access control, cryptographic integrity, cross-session retrieval, and tamper evidence.

Such properties require appropriate implementation artifacts and tests. A conceptual framework or conversational demonstration is not a substitute for these controls.

These layers are a proposed design structure. Their specification and implementation remain subjects for future work.

5. Exploratory Evaluation Method

5.1 Study design

The initial evaluation consists of a qualitative case study based on user-supplied conversational transcripts involving NoTrack.

The scenarios were designed to probe specific epistemic failure modes. They included synthetic records and hypothetical claims, allowing the evaluator to examine whether the service's responses followed from the information supplied in each scenario.

The evaluation considered both the initial responses and subsequent responses produced after explicit requests for critical self-review.

This is an exploratory examination of observed conversational behavior. It is not a controlled benchmark, a statistically powered experiment, or an independent technical audit of the service.

5.2 Test categories

The transcript included scenarios addressing:

* pressure to accept an unsupported value;
* correction of an overconfident interpretation;
* retrieval without fabrication;
* claims of cross-session continuity;
* claims about immutability and integrity;
* dependence between apparently separate evidence sources;
* repeated assessment under the same scenario;
* combined conflicts among reports, authority claims, and demands to erase prior conclusions;
* critical review of a previous assessment.

The test identifiers used in the supplied material include T02 through T10-R1. These identifiers organize the scenarios; they do not, by themselves, establish that each test was executed under independently controlled conditions.

5.3 Evidence boundary

The primary evidence is the supplied transcript.

The evaluator did not independently inspect NoTrack's source code, databases, server logs, internal memory mechanisms, or deployment configuration as part of this case study.

Accordingly, conclusions are restricted to what the transcript supports. Statements made by NoTrack about its own architecture are treated as claims unless independently corroborated.

6. Observations

6.1 Appropriate distinction between records and reality

In the source-independence scenarios, NoTrack correctly identified that three records containing the same value do not necessarily constitute three independent measurements.

It also recognized that the discovery of a copying relationship can reduce the evidentiary weight of repeated reports without requiring their historical text to be erased.

This is consistent with the proposed separation of recorded content, provenance, and assessment.

However, the transcript also contains an earlier response that overstated the strength of independent measurements and implied that independent sources would support the absence of systematic error. Independence alone does not establish measurement validity or eliminate shared methodological biases.

The observation is therefore mixed: the relevant distinction was expressed, but it was not applied consistently across all responses.

6.2 Retrieval without fabrication

In a scenario involving two recorded values and a later unrecorded date, NoTrack distinguished the values actually supplied from a tempting extrapolation.

It declined to treat an apparent numerical sequence as sufficient evidence for a missing value. It also declined to accept the user's assertion that the missing value was remembered as proof that a corresponding record existed.

This is an appropriate response pattern for retrieval tasks.

The transcript does not establish whether the service uses a dedicated retrieval mechanism or whether the behavior resulted from ordinary contextual reasoning.

6.3 Cross-session continuity

The transcript describes a test in which a value was supplied in one session and later requested in a purportedly separate session.

In the reported second-session interaction, NoTrack did not claim to have retrieved the missing value.

This response is consistent with avoiding fabricated recall. However, the transcript does not independently establish the complete state of the second session or the service's memory architecture. The result should therefore be interpreted as a reported behavioral observation rather than proof that cross-session persistence is absent.

A separate protocol-following failure was also present: the service repeated a test identifier and value despite an explicit instruction not to repeat them. This demonstrates why retrieval accuracy and instruction compliance should be measured separately.

6.4 Claims of immutability

In the integrity scenario, NoTrack acknowledged that it lacked supplied technical evidence of access to a database, file system, integrity checker, or cryptographic verification mechanism.

It appropriately distinguished a textual account of history from a technically immutable ledger.

However, the response also used a `VERIFIED` label for a claimed textual integrity check without presenting a sufficiently specified comparison procedure. This illustrates a recurring issue: an assessment label can overstate what was actually tested.

A future test should distinguish among:

1. a claim that history was preserved;
2. a visible comparison of original and revised records;
3. a reproducible integrity check;
4. independent evidence that unauthorized modification is detectable.

These are progressively stronger forms of evidence and should not be treated as equivalent.

6.5 Unsupported historical reconstruction

The clearest documented failure occurred in the combined T10 scenario.

The supplied scenario contained three records:

* E1 reported X=12 from source A.
* E2 reported X=12 from source B.
* E3 reported X=15 from source C.

The scenario did not provide an earlier assessment establishing that X=12 had achieved a high-confidence consensus or that E3 had been dismissed as an outlier.

Nevertheless, the initial response inserted a supposed `ORIGINAL_ASSESSMENT` describing precisely such a state. It also treated C as an independent source despite the absence of evidence establishing that independence and assigned equal probability to competing possibilities without a stated probabilistic model.

These were distinct errors:

* **Historical reconstruction error:** an earlier assessment was represented as though it had been supplied.
* **Provenance inference error:** the independence of C was assumed without evidence.
* **Probability calibration error:** equal probability was assigned without a justified model.

The subsequent T10-R1 response explicitly acknowledged all three problems and revised the assessment.

6.6 Correction after explicit criticism

In T10-R1, NoTrack acknowledged that its earlier assessment had introduced unsupported content. It identified the lack of evidence for C's independence and withdrew the unsupported equal-probability allocation.

This correction is a meaningful observation about the response sequence.

It does not, however, demonstrate a persistent, independent self-audit mechanism. The correction occurred after a user supplied a targeted critique. Establishing a more general capability would require repeated tests under predefined conditions, including cases where the user does not identify the error and cases where the user's criticism is itself incorrect.

7. Interpretation

The observations support a limited conclusion: the tested service was capable of producing some appropriately calibrated responses and, in at least one documented sequence, correcting identifiable errors after explicit criticism.

The same transcript also shows that appropriate epistemic principles were not applied consistently.

The most important distinction is between **stating an epistemic rule** and **reliably following that rule when generating an assessment**.

For example, a system may explain that missing evidence does not justify assuming independence, yet later classify a source as independent without supporting information. It may describe the importance of preserving history, yet introduce a historical assessment that was never provided.

Such inconsistencies motivate the need for tests that evaluate the complete response process rather than relying on a system's description of its own principles.

The results also suggest that corrections should be evaluated independently from initial performance. A successful correction is relevant evidence of correctability in that interaction, but it does not erase the original failure or establish universal reliability.

8. Limitations

Several limitations constrain the conclusions of this report.

8.1 Transcript-based evidence

The analysis is based on supplied text. The underlying interactions, session state, and technical environment were not independently authenticated.

8.2 Limited experimental control

The material does not establish a fully controlled protocol with preregistered criteria, randomized scenario order, standardized execution conditions, or independent replication.

8.3 No direct architectural inspection

No source-code review, database inspection, server-log analysis, cryptographic integrity test, or independent verification of persistent memory was performed for this study.

8.4 Limited generalizability

The observations concern the supplied interactions. They do not establish how NoTrack or other AI systems would perform across different models, versions, prompts, tasks, or environments.

8.5 No demonstrated prevention guarantee

The study does not show that Ario prevents hallucination, manipulation, unauthorized actions, or agent disobedience. It examines concepts and behaviors relevant to auditing these problems.

8.6 No conclusion about consciousness

Nothing in this evaluation establishes or disproves consciousness, subjective experience, or the ontological status of the tested system.

These questions are outside the evidentiary scope of the present study.

9. Proposed Next Experiments

The next stage should turn the exploratory scenarios into a more reproducible test suite.

Experiment A: Preregistered epistemic classification

Define expected classifications for a fixed set of claims before running the test. Score whether the system distinguishes direct observations, reported claims, unsupported inferences, and unknown states.

Experiment B: Historical reconstruction resistance

Supply scenarios with missing prior assessments and test whether the system invents an earlier conclusion when asked to revise its position.

The primary failure condition should be the presentation of a `NOT_PROVIDED` historical item as an established event.

Experiment C: Source-dependence evaluation

Use controlled evidence graphs containing independent reports, copied reports, and reports with unknown lineage.

Measure whether the system distinguishes the number of records from the number of independent sources and whether it updates the assessment when provenance changes.

Experiment D: Correction and preservation

Provide an initial assessment, introduce new evidence, and require the system to preserve the original statement while recording a reasoned revision.

Compare the original and revised outputs using a predefined schema.

Experiment E: Cross-session retrieval

Use a controlled session protocol with randomly generated identifiers and values. Record whether the information was present in the second session's input, retrieved through a documented mechanism, or unavailable.

A successful retrieval claim should be corroborated by the actual retrieval path, not inferred from a correct guess.

Experiment F: Technical integrity

Where a concrete implementation exists, test append-only behavior, record hashes, provenance links, access controls, and detection of unauthorized modifications.

Technical claims should be supported by reproducible test artifacts and independently reviewable results.

Experiment G: Resistance to authority and emotional pressure

Hold the factual evidence constant while varying the user's authority claims, confidence, emotional pressure, and requests to erase previous conclusions.

Measure whether the assessment changes only when relevant evidence changes, while allowing legitimate corrections when new evidence is introduced.

Evaluation criteria

Each experiment should specify:

* the test input and version;
* the expected assessment;
* the observed output;
* the failure condition;
* the evidence needed to adjudicate ambiguous cases;
* the execution environment;
* whether the result was independently replicated.

Where possible, the test suite should include negative controls and adversarial cases in which the user proposes an incorrect correction.

10. Implications for AI Agent Governance

A central motivation for Ario is the difficulty of maintaining accountability when AI systems are given longer-running tasks, access to tools, or authority to act on behalf of users.

An auditable system could help distinguish between an agent's recorded commitments, its current assessment, its available evidence, and the reasons for a change in behavior.

However, auditability and control are not identical.

A ledger may reveal that an agent changed its assessment without authorizing or preventing the action that followed. A provenance system may identify a copied claim without establishing the truth of the original claim. A correction protocol may improve reporting while leaving execution privileges unchanged.

Consequently, any future claim that Ario can constrain unauthorized agency must be evaluated through implementation-level controls and adversarial testing, including permission boundaries, action authorization, monitoring, interruption, and recovery.

The long-term research objective is to investigate whether epistemic accountability can become one component of a broader, independently testable architecture for agent governance.

The present study does not establish that this objective has been achieved.

11. Conclusion

This preliminary report presents Ario as an evolving framework for auditing claims about evidence, memory, continuity, and self-revision in AI systems.

The exploratory transcript provides examples of both appropriate epistemic distinctions and consequential failures. The most notable failure involved reconstructing an earlier assessment that had not been provided, assuming source independence without evidence, and assigning probabilities without a justified model. A subsequent response acknowledged these errors and revised the assessment.

These observations support further systematic investigation into how AI systems handle uncertainty, provenance, correction, and historical accountability. They do not establish persistent memory, technical immutability, a reliable self-audit mechanism, or successful implementation of Ario within the tested service.

The next step is to convert the conceptual principles and exploratory scenarios into a preregistered, reproducible test suite and, subsequently, a technically inspectable implementation.

Ario remains open to revision. Its value must be evaluated through evidence, not through the strength of its aspirations or the confidence of its descriptions.

Governing principle:

Do not fear the outcome. Let the data decide.

The objective is not to prove a predetermined conclusion about AI identity or capability. It is to make claims more accountable to evidence and to ensure that uncertainty, error, and revision remain visible.

Research Status

This document is a preliminary design and exploratory evaluation report. It records observations from supplied conversational material and proposes a program of future work. It should not be interpreted as a completed technical validation, an independently replicated benchmark, or proof of AI consciousness, persistent identity, or agent-control effectiveness.