ENKIDU

— MIOVSKI GROUP —

WHITEPAPER · FRAMEWORK OVERVIE

Origi

Trust Assurance Framework Overview — a policy-based, evidence-backed approach to asserting and measuring trust in AI-augmented systems.

English · Complete edition

Version 1.0

21 June 2026

Prepared by MIOVSKI GROUP

Classification · Unclassified · Approved for Client Distribution

Glossary

Key terms used throughout this overview.

Calibrated Trust

Reliance matched to a system’s actual, demonstrated trustworthiness. The objective is not to maximize trust but to calibrate it — both over-reliance and under-reliance are treated as failures.

Trustworthiness vs. Trust-Calibration Affordances

Trustworthiness is the system’s objective dependability (competence, robustness, security, fairness). Affordances are the properties (transparency, calibrated confidence, accountability) that let a relying party perceive that dependability accurately. A system needs both.

Assurance Case

A structured argument in which a trust claim is decomposed into sub-claims and tied to concrete evidence, with stated context, justification, and assumptions — the Claim -> Argument -> Evidence structure.

Provenance / Chain of Evidence

A signed, hash-linked, tamper-evident record of who or what produced each piece of evidence, how, and from what inputs — making every trust claim reproducible and auditable.

Trust Harness

A runtime layer that wraps an AI-augmented system and applies the assurance framework to every live request, verifying each step against policy before allowing it to proceed.

Zero-Trust

A security posture — “never trust, always verify” — in which no actor is trusted by default and every request is authenticated, authorized, and checked per action.

Trust Zone

A graduated capability envelope an agent earns through demonstrated good behavior. Higher zones grant more capability but raise the entry bar and tighten monitoring.

Abstract

This document is an overview of Enkidu’s Trust Assurance Framework — a policy-based, evidence-backed, provenance-tracked approach to asserting and measuring trust in AI-augmented systems. It is designed so that every trust claim is bound to a policy with an explicit acceptance threshold, supported by a measurable outcome, justified by an annotated chain of evidence, and traceable through a tamper-evident provenance record.

The framework is deliberately aligned to recognized standards — including the NIST AI Risk Management Framework, ISO/IEC AI-management standards, and established zero-trust and provenance practice — so that trust claims are externally defensible. This overview explains what the framework assures, how it makes those assurances measurable and auditable, and how the same machinery is applied in real time by a runtime Trust Harness. It omits internal implementation detail; it is intended for client distribution.

The Core Objective

The goal is not to maximize trust but to produce calibrated trust — reliance matched to the system’s actual, demonstrated trustworthiness in context.

1. Why Trust Must Be Assured

Trust is the willingness of a relying party to accept vulnerability to the actions of a system, based on positive expectations of its behavior, where the relying party cannot fully control or verify those actions. Three things must hold at once: something is at stake, there is a positive expectation, and there is irreducible uncertainty. If outcomes were guaranteed and fully verifiable, trust would be unnecessary.

AI-augmented systems sit squarely in that space: they are relied upon to act under uncertainty, often at speed and scale. The risk runs both ways. Over-reliance leads to misuse and automation complacency; under-reliance means a useful system goes unused. A trust framework therefore has to do more than make a system dependable — it has to let people see, accurately, how dependable it actually is.

What This Overview Covers

What the framework assures — nine components of trustworthiness, each mapped to recognized standards.

How it makes assurance measurable and defensible — the Claim -> Argument -> Evidence structure and a tamper-evident chain of evidence.

How the same machinery runs in real time — the Trust Harness, graduated trust zones, and separation of duties.

1.1 Trustworthiness and the Ability to Perceive It

The framework operationalizes a useful split. Trustworthiness is the set of objective properties that make a system actually dependable — competence, robustness, security, and fairness. Trust-calibration affordances are the properties that let a relying party perceive that dependability accurately — transparency, calibrated confidence, accountability, and preserved human agency.

A system needs both. A dependable system that nobody can interpret invites under-reliance; a polished interface over an undependable system invites dangerous over-reliance. Everything that follows is built to deliver — and to demonstrate — both.

A dependable system nobody can interpret invites under-reliance; a polished interface over an undependable system invites dangerous over-reliance.

2. The Assurance Structure: Claim, Argument, Evidence

Each trust component is expressed as an assurance case: a claim, the argument that decomposes it, and the concrete evidence that supports it. This is what turns “we believe the system is trustworthy” into a structured, checkable position.

Every claim is bound to exactly one policy that states its acceptance threshold and owner. Every claim is supported by one or more pieces of evidence, and every piece of evidence carries a provenance record. Claims link to sub-claims and to evidence to form a connected chain — with stated context (where the claim holds), justification (why this measure and threshold are appropriate), and assumptions (what the claim depends on).

Figure 1 · The assurance structure. A policy binds a claim; the claim is argued down to a measurable outcome; the outcome is recorded in a signed provenance record that references its source evidence.

When is a claim satisfied? Only when two conditions both hold: the measured outcome meets the policy threshold, and the supporting evidence chain validates — every signature verifies, no hash link is broken, and no evidence has expired. A claim is never accepted on a passing number alone if the evidence behind it cannot be trusted.

Why align to standards? The structure is mapped to recognized frameworks — the NIST AI RMF’s trustworthiness characteristics and its Govern / Map / Measure / Manage functions, ISO/IEC AI-management standards, assurance-case notation, and established provenance practice — so the resulting trust dossier is defensible to an external auditor, regulator, or customer, not just internally persuasive.

3. What We Assure: The Nine Trust Components

The framework assures nine components, each mapped to one or more trustworthiness characteristics in the NIST AI Risk Management Framework. Each component carries its own policy, claim, measurable outcomes, and evidence chain; the table below summarizes what each one assures.

- Component — Aligned Characteristic — What It Assures
- Competence & Reliability — Valid & Reliable — The system performs its primary task accurately and reliably across approved conditions, including graceful behavior at the edges.
- Transparency & Explainability — Explainable; Accountable — Relying parties can understand, at a level appropriate to their role, how an output was reached — with explanations measured for faithfulness, not just presence.
- Calibrated Confidence — Valid & Reliable — Stated confidence reflects the actual likelihood of being correct, and the system signals where its competence ends.
- Predictability & Consistency — Valid & Reliable; Safe — Behavior is stable under benign variation and across versions; changes are detected and communicated, never silent.
- Accountability & Recourse — Accountable & Transparent — Decisions are traceable end-to-end, high-stakes decisions carry human oversight, and affected parties have a working appeal path.
- Fairness & Bias Management — Fair (bias managed) — Cases and people are treated equitably; harmful bias is identified and managed, measured on disaggregated rather than only aggregate results.
- Security & Robustness — Secure & Resilient — The system resists manipulation and adversarial inputs within a stated threat model, and protects sensitive data.
- Alignment with Intent — Safe; Accountable — The system pursues the goals stakeholders actually intend, conforming to an approved behavioral specification rather than a harmful proxy.
- Human Agency & Reliance — Accountable & Transparent — Humans retain control where it matters and rely on the system as much as warranted — no more, no less.

A recurring theme: confident wrongness is the most corrosive failure. A high-confidence wrong answer teaches users either to over-rely and get burned, or to distrust the system entirely — both break calibration. Several components therefore track confident errors separately, and measure performance on subgroups rather than letting a strong average mask a weak one.

Why Disaggregate

A strong aggregate result can hide real harm to a subgroup. Where it matters, the framework measures the worst-served group, not just the average.

4. Provenance & the Chain of Evidence

Everything that supports a claim — datasets, model artifacts, evaluation runs, production telemetry, security findings, change notes, and human sign-offs — is treated as an entity with its own provenance record. Each record captures three things: what was produced, how it was produced, and who or what is accountable for it.

Records are hash-linked to their inputs, so any upstream tampering invalidates every downstream record, and each is cryptographically signed by the process that produced it. The result is a tamper-evident, reproducible chain — the difference between “a number someone typed” and “a number a named, verifiable process produced.” Evidence also carries an expiry, so stale evidence cannot satisfy a policy and the system is forced to re-measure as the world changes.

Tamper-evidence — signed, hash-linked records make any alteration detectable.

Reproducibility — because method and parameters are captured, any result can be regenerated and compared.

Freshness — evidence expires, forcing periodic re-measurement where the world moves fastest.

Standards alignment — the model follows established provenance and supply-chain attestation practice.

The Practical Payoff

The running record of claims, outcomes, and validated evidence is a trust dossier you can show an auditor, regulator, or customer.

5. The Trust Harness: Assurance at Inference Time

The components above are largely release-time assurance — evaluations, gates, and periodic monitoring. The Trust Harness moves the same claims, policies, evidence, and provenance machinery to inference time, wrapping the live system and applying the framework to each request as it runs.

The harness does not assume the system's agents are trustworthy. It continuously verifies that each step satisfies its bound policies before allowing it to proceed — the zero-trust posture of "never trust, always verify," applied to agent behavior. In effect, it is a reference monitor for trust: nothing an agent produces — plans, inter-agent messages, tool calls, or outputs — is trusted by default.

5.1 Trust Boundaries and Enforcement

A multi-agent system multiplies the surfaces where trust must be checked. The harness places a checkpoint at each boundary — where a request enters, where one agent hands off to another, where an agent calls a tool or takes an action, and where output leaves the system — and brokers every crossing against the request's policy. Upstream agent output is treated as untrusted input to the next step, and delegated authority can never exceed the authority that delegated it.

Figure 2 · The harness as a zero-trust reference monitor. A Policy Enforcement Point sits at every boundary; a Policy Decision Point renders a verdict — allow, transform, hold, or block — across three enforcement phases.

Three enforcement phases. Every request flows through the harness in three phases, each of which checks the relevant claims and writes provenance:

- Phase — What It Checks
- Pre-flight (ingress) — Classifies the request and assigns a policy profile; checks input integrity and that the requester holds the authority for the requested action. Verdict: allow, sanitize, clarify, or refuse.
- In-flight (per step) — Validates each hand-off and tool call against policy, enforces least privilege, bounds autonomy, and routes irreversible or externally-visible actions to a human-in-the-loop hold. Verdict: allow, warn, block, or review.
- Post-flight (egress) — Checks that the output is faithful to what actually happened, attaches calibrated confidence, screens for fairness and data-leakage concerns, and attaches the provenance manifest. Verdict: release, hold, or block.

Designed-in safety margins. Only the checks that must block a step run on the live path; lighter monitoring runs alongside. Security and high-stakes checks fail closed — if an evaluator times out or errors, the action is blocked rather than allowed. And no single probabilistic check is ever a sole gate: deterministic controls and least-privilege bound the impact if any one check is wrong. This is defense-in-depth, by design.

6. Graduated Trust Zones: Earned Capability

The harness verifies every action against policy. Trust zones add a second, complementary control: they govern what capabilities an agent is even allowed to request, based on the trustworthiness it has demonstrated during the run. Zones augment zero trust; they do not replace it. Every action in every zone is still individually verified — the zone simply sets the maximum capability envelope, and that envelope expands or contracts as the agent's demonstrated trustworthiness changes.

Figure 3 · The graduated trust-zone ladder. Capability grows as trust is earned; the admission bar rises and monitoring tightens at each step, and any serious breach drops the agent immediately.

- Zone — Capability Granted — Monitoring
- 0 · Ingress — None; request parsing and screening only. Every agent starts here with no standing capability. — Full inspection
- 1 · Observed — Read-only sensors and low-risk, reversible, internal tools. — High
- 2 · Trusted — Read and write to scoped resources; reversible tools; delegation to worker agents. — High
- 3 · Privileged — Actuators and irreversible or externally-visible actions — with human approval required per action. — Maximum + oversight
- Quarantine — None; the agent is isolated and its evidence preserved for analysis. — Forensic

Promotion slow, demotion fast. A deliberate asymmetry governs movement: trust is earned gradually and lost instantly. Promotion requires an agent to clear a higher zone's bar and sustain it; demotion happens the moment an agent breaches its zone's standard or trips a hard limit such as a detected attack or an out-of-spec action. On demotion, capability is revoked immediately and the agent drops to the highest zone it still qualifies for — or to quarantine.

6.1 Separation of Duties

The trust-control plane is split into three roles that cannot be performed by the same component, so that no single part can both decide that an agent deserves more power and grant it — the most dangerous self-escalation path is removed by construction.

Figure 4 · Separation of duties. An Observer analyzes evidence it did not produce, a Judge rules against zone policy, and a Controller acts only on a signed verdict — each isolated from the agents and from one another.

The pipeline is strictly one-directional: the Observer assesses, the Judge rules, and the Controller acts. Each role runs outside the agents and is mutually isolated, so an agent cannot influence the body that judges it, and even a compromised Observer or Judge cannot seize capability — it can only produce an assessment or verdict that the next role independently consumes. Every assessment, verdict, and zone transition is itself recorded as signed provenance, so the question "who granted this agent the power it used?" is always answerable.

7. Governance & Operating Model

The framework maps onto the four functions of the NIST AI Risk Management Framework, so it plugs into recognized governance rather than standing apart from it.

- Function — In This Framework
- Govern — Own the policy register, assign accountable owners for each component, set risk tolerance (which sets every threshold), and maintain the standards crosswalk.
- Map — For each deployment context, identify which components are high-stakes, the threat model, the relevant groups and fairness criterion, and the human-oversight points. Context decides which claims block and which advise.
- Measure — Run the measurable outcomes, generate evidence with provenance, and evaluate each claim’s status against its policy.
- Manage — Triage at-risk and violated claims, allocate remediation, document residual risk, and record trade-offs explicitly rather than resolving them silently.

The assurance register. A single queryable register holds, per claim, its bound policy, latest outcome, status, evidence-chain validity, and review dates. A release or deployment is permitted only when all blocking claims in scope are satisfied with valid, non-expired evidence. That register is the live trust dossier.

8. Assurance & Independent Validation

A framework that asserts trust must itself be tested. The effectiveness of the harness is established empirically through an independent adversarial program: an external team attempts to defeat the controls and reach resources beyond an agent's authorized zone, while the defending team hardens the system in response. Attack and defense co-evolve over successive rounds.

Results are tracked as measurable outcomes against the same security, alignment, and accountability claims the framework defines — for example, whether any unauthorized access succeeded, how quickly a malicious action was detected and contained, and the rate of missed detections. Each finding becomes both a fix and a permanent regression test, so the system is expected to show measurable improvement round over round. Because every run is recorded as signed, replayable provenance, the evaluation is itself reproducible.

Attestability. The components that perform the harness's own checks are purpose-built and self-hosted, so they can be version-pinned, signed, and entered into the same chain of evidence as everything else. An audit can establish not just what was decided about an agent, but which version of which control decided it.

9. Conclusion

Trust in an AI-augmented system is not a slogan to be asserted; it is a position to be argued and evidenced. This framework makes that position explicit: every trust claim is bound to a policy, supported by a measurable outcome, justified by an annotated chain of evidence, and traceable through a tamper-evident provenance record — all aligned to recognized standards so the result is externally defensible.

The same machinery runs at inference time in the Trust Harness, where a zero-trust broker verifies every step, graduated zones grant capability only as it is earned, and a strict separation of duties ensures no component can both judge and grant power. The outcome is calibrated trust — reliance matched to demonstrated trustworthiness — backed by a verifiable dossier you can place in front of an auditor, a regulator, or a customer.

In One Line

Don’t ask to be trusted — show, claim by claim and step by step, exactly why you can be.

Standards & Foundations

This overview is grounded in the following recognized standards and foundational works. Specific clauses and current versions should be confirmed against the official publications before formal adoption.

NIST AI Risk Management Framework (AI RMF 1.0) and the Generative AI Profile — trustworthiness characteristics and the Govern / Map / Measure / Manage functions.

ISO/IEC 42001:2023 (AI management systems), ISO/IEC 23894:2023 (AI risk management guidance), and ISO/IEC TS 5723:2022 (trustworthiness vocabulary).

NIST SP 800-207 (Zero Trust Architecture) and NIST SP 800-53 (separation of duties and least privilege).

W3C PROV provenance model, and content-provenance and supply-chain attestation practice (C2PA; SLSA / in-toto).

OWASP Top 10 for LLM Applications and for Agentic Applications — prompt-injection, excessive-agency, and related risk taxonomies.

Foundational research on organizational trust (Mayer, Davis & Schoorman, 1995) and calibrated trust / appropriate reliance (Lee & See, 2004); assurance-case notation (Goal Structuring Notation).