ENKIDU

— MIOVSKI GROUP —

WHITEPAPER · FRAMEWORK-ÜBERBLI

Origi

Überblick über das Trust-Assurance-Framework — ein richtlinienbasierter, beweisgestützter Ansatz, Vertrauen in KI-gestützte Systeme zu behaupten und zu messen.

Deutsch · Kompaktausgabe

Nach Version 1.0 · 21. Juni 2026
Erstellt von MIOVSKI GROUP
Einstufung · Nicht klassifiziert · Zur Weitergabe an Kunden freigegeben
Die vollständige englische Ausgabe ist maßgeblich

Zu dieser Ausgabe

Dieses Dokument ist eine deutsche Kompaktausgabe des Whitepapers « Trust Assurance Overview », Version 1.0 (21. Juni 2026), erstellt von der Miovski Group und veröffentlicht unter dem Titel « Origi — Trust Assurance Framework Overview ». Jeder Abschnitt des Originals ist in gekürzter Form vertreten. Bei Abweichungen ist die vollständige englische Ausgabe maßgeblich. Einstufung: nicht klassifiziert — zur Weitergabe an Kunden freigegeben.

Zusammenfassung

Dieses Dokument ist ein Überblick über Enkidus Trust-Assurance-Framework (Origi) — einen richtlinienbasierten, beweisgestützten und herkunftsverfolgten Ansatz, Vertrauen in KI-gestützte Systeme zu behaupten und zu messen. Es ist so entworfen, dass jede Vertrauensbehauptung an eine Richtlinie mit expliziter Annahme-Schwelle gebunden ist, durch ein messbares Ergebnis gestützt, durch eine annotierte Beweiskette begründet und über einen manipulationssicheren Herkunftsnachweis nachvollziehbar wird.

1. Warum Vertrauen gesichert werden muss

Vertrauen ist die Bereitschaft einer sich verlassenden Partei, Verwundbarkeit gegenüber den Handlungen eines Systems zu akzeptieren — auf Grundlage positiver Erwartungen an sein Verhalten, wenn sie diese Handlungen weder vollständig kontrollieren noch verifizieren kann. Drei Dinge müssen zugleich gelten: Etwas steht auf dem Spiel, es gibt eine positive Erwartung, und es bleibt irreduzible Unsicherheit. Wären Ergebnisse garantiert und vollständig verifizierbar, wäre Vertrauen unnötig.

Das Framework operationalisiert eine nützliche Trennung. Vertrauenswürdigkeit ist die Menge objektiver Eigenschaften, die ein System tatsächlich verlässlich machen — Kompetenz, Robustheit, Sicherheit, Fairness. Vertrauenskalibrierende Affordanzen sind die Eigenschaften, die eine Partei diese Verlässlichkeit zutreffend wahrnehmen lassen — Transparenz, kalibrierte Konfidenz, Rechenschaft und bewahrte menschliche Handlungsfähigkeit. Das Ziel ist nicht maximales, sondern kalibriertes Vertrauen: Verlass im Maß der nachgewiesenen Vertrauenswürdigkeit — denn Über- wie Untervertrauen sind beide Fehlschläge.

2. Die Assurance-Struktur: Behauptung, Argument, Beweis

Jede Vertrauenskomponente wird als Assurance-Case ausgedrückt: eine Behauptung, das Argument, das sie zerlegt, und die konkreten Beweise, die sie stützen. Das verwandelt wir glauben, das System ist vertrauenswürdig" in eine strukturierte, prüfbare Position. Jede Behauptung ist an genau eine Richtlinie gebunden, mit expliziter Schwelle und benanntem Verantwortlichen; sie gilt erst als erfüllt, wenn das gemessene Ergebnis die Schwelle erreicht und die Beweiskette validiert — jede Signatur geprüft, kein Hash-Glied gebrochen, kein Beweis abgelaufen — niemals allein durch eine bestandene Zahl.

3. Was wir sichern: die neun Vertrauenskomponenten

Das Framework sichert neun Komponenten, jede einer oder mehreren Vertrauenswürdigkeits-Charakteristiken des NIST AI Risk Management Framework zugeordnet. Jede Komponente trägt ihre eigene Richtlinie, Behauptung, messbaren Ergebnisse und Beweiskette.

- Kompetenz & Verlässlichkeit — das System erfüllt seine Hauptaufgabe genau und zuverlässig über die zugelassenen Bedingungen hinweg, mit gutmütigem Verhalten an den Rändern.
- Transparenz & Erklärbarkeit — Beteiligte verstehen, auf dem ihrer Rolle angemessenen Niveau, wie ein Ergebnis zustande kam — Erklärungen an ihrer Treue gemessen, nicht bloß an ihrem Vorhandensein.
- Kalibrierte Konfidenz — die ausgewiesene Konfidenz entspricht der tatsächlichen Trefferwahrscheinlichkeit, und das System signalisiert, wo seine Kompetenz endet.
- Vorhersagbarkeit & Konsistenz — das Verhalten ist stabil unter gutartiger Variation und über Versionen hinweg; Änderungen werden erkannt und kommuniziert, niemals still.
- Rechenschaft & Einspruch — Entscheidungen sind Ende-zu-Ende nachvollziehbar, Hochrisiko-Entscheidungen tragen menschliche Aufsicht, und Betroffene haben einen funktionierenden Einspruchsweg.
- Fairness & Bias-Management — Fälle und Menschen werden gerecht behandelt; schädlicher Bias wird erkannt und gesteuert, gemessen an disaggregierten statt nur aggregierten Ergebnissen.
- Sicherheit & Robustheit — das System widersteht Manipulation und adversarialen Eingaben innerhalb eines erklärten Bedrohungsmodells und schützt sensible Daten.
- Ausrichtung an der Absicht — das System verfolgt die Ziele, die die Beteiligten tatsächlich meinen, gemäß einer genehmigten Verhaltensspezifikation statt eines schädlichen Stellvertreters.
- Menschliche Handlungsfähigkeit & Verlass — Menschen behalten die Kontrolle, wo es zählt, und verlassen sich genau so weit auf das System, wie gerechtfertigt — nicht mehr, nicht weniger.

4. Herkunft und Beweiskette

Alles, was eine Behauptung stützt — Datensätze, Modellartefakte, Evaluationsläufe, Produktionstelemetrie, Sicherheitsbefunde, Änderungsnotizen, menschliche Freigaben — wird als Entität mit eigenem Herkunftsnachweis behandelt. Jeder Nachweis erfasst drei Dinge: was erzeugt wurde, wie es erzeugt wurde und wer — Mensch oder System — dafür einsteht. Die Nachweise sind signiert und hash-verkettet, sodass Manipulation sichtbar wird; Beweise tragen ein Ablaufdatum, sodass Assurance offen altert statt im Verborgenen.

5. Das Vertrauensgeschirr: Assurance zur Inferenzzeit

Die obigen Komponenten sind großteils Assurance zur Freigabezeit — Evaluationen, Gates, periodische Überwachung. Das Vertrauensgeschirr trägt dieselben Behauptungen, Richtlinien, Beweise und dieselbe Herkunftsmaschinerie in die Inferenzzeit: Es umhüllt das laufende System und wendet das Framework auf jede Anfrage an, während sie läuft.

- Pre-Flight (Eingang) — klassifiziert die Anfrage und weist ein Richtlinienprofil zu; prüft Eingabeintegrität und die Autorität des Anfragenden. Urteil: zulassen, bereinigen, klären oder ablehnen.
- In-Flight (je Schritt) — validiert jede Übergabe und jeden Werkzeugaufruf gegen die Richtlinie, erzwingt geringstes Privileg, begrenzt Autonomie und leitet irreversible oder nach außen sichtbare Aktionen an eine menschliche Entscheidung.
- Post-Flight (Ausgang) — prüft, ob die Ausgabe dem tatsächlichen Geschehen treu ist, heftet kalibrierte Konfidenz an, prüft Fairness- und Datenabfluss-Risiken und fügt das Herkunftsmanifest bei.

Ein Multiagentensystem vervielfacht die Flächen, an denen Vertrauen geprüft werden muss. Das Geschirr setzt an jede Grenze einen Kontrollpunkt — wo eine Anfrage eintritt, wo ein Agent an einen anderen übergibt, wo ein Agent ein Werkzeug ruft oder handelt, und wo Ausgabe das System verlässt — und vermittelt jede Überquerung gegen die Richtlinie der Anfrage. Die Ausgabe eines vorgelagerten Agenten gilt dem nächsten Schritt als nicht vertrauenswürdige Eingabe, und delegierte Autorität kann die delegierende niemals übersteigen. Sicherheits- und Hochrisikoprüfungen versagen geschlossen.

6. Abgestufte Vertrauenszonen: verdiente Fähigkeit

Das Geschirr verifiziert jede Handlung gegen die Richtlinie. Vertrauenszonen fügen eine zweite, komplementäre Kontrolle hinzu: Sie regeln, welche Fähigkeiten ein Agent überhaupt anfordern darf — nach der Vertrauenswürdigkeit, die er im Lauf nachgewiesen hat. Zonen ergänzen Zero Trust, sie ersetzen es nicht: Jede Handlung wird weiterhin einzeln geprüft — die Zone setzt nur die maximale Fähigkeitshülle, die sich mit nachgewiesener Vertrauenswürdigkeit weitet oder zusammenzieht. Beförderung ist langsam; Herabstufung sofort.

- 0 · Eingang — keine Fähigkeit; nur Anfrageanalyse und -filterung. Jeder Agent beginnt hier. — Volle Inspektion.
- 1 · Beobachtet — Nur-Lese-Sensoren und risikoarme, reversible, interne Werkzeuge. — Hohe Überwachung.
- 2 · Vertraut — Lesen und Schreiben auf abgegrenzte Ressourcen; reversible Werkzeuge; Delegation an Arbeitsagenten. — Hohe Überwachung.
- 3 · Privilegiert — Aktuatoren und irreversible oder nach außen sichtbare Aktionen — mit menschlicher Freigabe je Aktion. — Maximale Überwachung.
- Quarantäne — keine Fähigkeit; der Agent ist isoliert, seine Beweise für die Analyse gesichert. — Forensisch.

Funktionstrennung: Die Vertrauens-Steuerungsebene ist in drei Rollen geteilt, die nicht dieselbe Komponente ausüben kann — ein Beobachter, der Beweise bewertet, die er nicht erzeugt hat, ein Richter, der gegen die Zonenrichtlinie urteilt, ein Controller, der nur auf signiertes Urteil handelt — sodass kein Teil zugleich entscheiden kann, dass ein Agent mehr Macht verdient, und sie ihm gewähren. Der gefährlichste Selbst-Eskalationspfad ist konstruktiv beseitigt.

7. Governance und Betriebsmodell

Das Framework bildet sich auf die vier Funktionen des NIST AI Risk Management Framework ab und fügt sich damit in anerkannte Governance ein, statt daneben zu stehen.

- Govern — das Richtlinienregister führen, je Komponente Verantwortliche benennen, die Risikotoleranz festlegen (die jede Schwelle setzt) und die Normen-Übersicht pflegen.
- Map — je Einsatzkontext die Hochrisiko-Komponenten, das Bedrohungsmodell, die relevanten Gruppen und das Fairnesskriterium sowie die menschlichen Aufsichtspunkte bestimmen.
- Measure — die messbaren Ergebnisse ausführen, Beweise mit Herkunft erzeugen und den Status jeder Behauptung gegen ihre Richtlinie bewerten.
- Manage — gefährdete und verletzte Behauptungen triagieren, Behebung zuweisen, Restrisiko dokumentieren und Abwägungen ausdrücklich festhalten, statt sie still aufzulösen.

8. Assurance und unabhängige Validierung

Ein Framework, das Vertrauen behauptet, muss selbst geprüft werden. Die Wirksamkeit des Geschirrs wird empirisch durch ein unabhängiges adversariales Programm belegt: Ein externes Team versucht, die Kontrollen zu überwinden und Ressourcen jenseits der autorisierten Zone eines Agenten zu erreichen, während das verteidigende Team das System härtet. Angriff und Verteidigung entwickeln sich über aufeinanderfolgende Runden gemeinsam; jeder Befund wird zu einer Korrektur und einem dauerhaften Regressionstest — und weil jeder Lauf signierte, wiederabspielbare Herkunft ist, ist die Bewertung selbst reproduzierbar.

9. Schlussfolgerung

Vertrauen in ein KI-gestütztes System ist keine Parole, die man verkündet; es ist eine Position, die man begründet und belegt. Dieses Framework macht diese Position explizit: Jede Vertrauensbehauptung ist an eine Richtlinie gebunden, durch ein messbares Ergebnis gestützt, durch eine annotierte Beweiskette begründet und über einen manipulationssicheren Herkunftsnachweis nachvollziehbar — alles an anerkannten Normen ausgerichtet, damit das Ergebnis nach außen verteidigbar ist.

Grundlagen und Normen: NIST AI RMF, ISO/IEC 42001 und 23894, NIST SP 800-207 und 800-53, W3C PROV, C2PA, SLSA/in-toto, OWASP (LLM und Agenten), GSN-Notation und die Forschung zu kalibriertem Vertrauen. Konkrete Klauseln und aktuelle Fassungen sind vor einer förmlichen Übernahme gegen die offiziellen Publikationen zu bestätigen.