ENKIDU

— MIOVSKI GROUP —

LIBRO BLANCO · VISIÓN GENERAL DE

Origi

Visión general del marco de aseguramiento de la confianza — un enfoque basado en políticas y respaldado por evidencia para afirmar y medir la confianza en sistemas aumentados por IA.

Español · Edición resumida

Basada en la versión 1.0 · 21 de junio de 2026

Preparado por MIOVSKI GROUP

Clasificación · No clasificado · Aprobado para distribución a clientes

La edición completa en inglés prevalece

Acerca de esta edición

Este documento es una edición resumida, en español, del libro blanco « Trust Assurance Overview », versión 1.0 (21 de junio de 2026), preparado por Miovski Group y publicado bajo el título « Origi — Trust Assurance Framework Overview ». Cada sección del original está representada de forma abreviada. En caso de divergencia, la edición completa en inglés prevalece. Clasificación: no clasificado — aprobado para distribución a clientes.

Resumen

Este documento es una visión general del marco de aseguramiento de la confianza de Enkidu (Origi): un enfoque basado en políticas, respaldado por evidencia y trazado por procedencia para afirmar y medir la confianza en sistemas aumentados por IA. Está diseñado para que cada afirmación de confianza esté vinculada a una política con un umbral de aceptación explícito, sustentada por un resultado medible, justificada por una cadena de evidencia anotada y trazable mediante un registro de procedencia a prueba de manipulaciones.

1. Por qué la confianza debe asegurarse

La confianza es la disposición de una parte a aceptar vulnerabilidad ante las acciones de un sistema, sobre la base de expectativas positivas de su comportamiento, cuando esa parte no puede controlar ni verificar plenamente dichas acciones. Tres cosas deben cumplirse a la vez: hay algo en juego, existe una expectativa positiva y hay incertidumbre irreducible. Si los resultados estuvieran garantizados y fueran plenamente verificables, la confianza sería innecesaria.

El marco operacionaliza una distinción útil. La fiabilidad (trustworthiness) es el conjunto de propiedades objetivas que hacen a un sistema realmente confiable — competencia, robustez, seguridad, equidad. Las facilidades de calibración de la confianza son las propiedades que permiten a la parte percibir esa fiabilidad con precisión — transparencia, confianza calibrada, rendición de cuentas y preservación de la agencia humana. La meta no es la confianza máxima sino la confianza calibrada: una dependencia proporcional a la fiabilidad demostrada, porque confiar de más y confiar de menos son ambos fracasos.

2. La estructura de aseguramiento: afirmación, argumento, evidencia

Cada componente de confianza se expresa como un caso de aseguramiento: una afirmación, el argumento que la descompone y la evidencia concreta que la sustenta. Eso convierte « creemos que el sistema es confiable » en una posición estructurada y comprobable. Cada afirmación se vincula a exactamente una política, con umbral explícito y responsable designado; solo se satisface cuando el resultado medido alcanza el umbral y la cadena de evidencia valida — cada firma verificada, ningún eslabón de hash roto, ninguna evidencia caducada — nunca por una simple cifra aprobatoria.

3. Qué aseguramos: los nueve componentes de confianza

El marco asegura nueve componentes, cada uno mapeado a una o más características de fiabilidad del marco de gestión de riesgos de IA del NIST. Cada componente porta su propia política, afirmación, resultados medibles y cadena de evidencia.

- Competencia y fiabilidad — el sistema realiza su tarea principal con exactitud y constancia en las condiciones aprobadas, con comportamiento elegante en los límites.
- Transparencia y explicabilidad — las partes comprenden, al nivel apropiado a su rol, cómo se llegó a un resultado — explicaciones medidas por su fidelidad, no solo por su presencia.
- Confianza calibrada — la confianza declarada refleja la probabilidad real de acierto, y el sistema señala dónde termina su competencia.
- Previsibilidad y consistencia — el comportamiento es estable bajo variación benigna y entre versiones; los cambios se detectan y comunican, nunca en silencio.
- Rendición de cuentas y recurso — las decisiones son trazables de extremo a extremo, las de alto riesgo llevan supervisión humana, y los afectados disponen de una vía de apelación que funciona.
- Equidad y gestión de sesgos — casos y personas se tratan equitativamente; el sesgo dañino se identifica y gestiona, medido sobre resultados desagregados y no solo sobre promedios.
- Seguridad y robustez — el sistema resiste manipulación y entradas adversarias dentro de un modelo de amenaza declarado, y protege los datos sensibles.
- Alineación con la intención — el sistema persigue las metas que las partes realmente pretenden, conforme a una especificación de comportamiento aprobada y no a un sustituto dañino.
- Agencia y dependencia humanas — los humanos conservan el control donde importa y confían en el sistema exactamente cuanto está justificado — ni más, ni menos.

4. Procedencia y cadena de evidencia

Todo lo que sustenta una afirmación — conjuntos de datos, artefactos de modelos, corridas de evaluación, telemetría de producción, hallazgos de seguridad, notas de cambio, aprobaciones humanas — se trata como una entidad con su propio registro de procedencia. Cada registro captura tres cosas: qué se produjo, cómo se produjo y quién — persona o sistema — responde por ello. Los registros se firman y encadenan por hash, de modo que toda manipulación se hace visible; la evidencia porta caducidad, de forma que el aseguramiento envejece a la vista y no en silencio.

5. El arnés de confianza: aseguramiento en tiempo de inferencia

Los componentes anteriores son en gran parte aseguramiento en tiempo de entrega — evaluaciones, puertas, vigilancia periódica. El arnés de confianza lleva las mismas afirmaciones, políticas, evidencia y maquinaria de procedencia al tiempo de inferencia: envuelve el sistema vivo y aplica el marco a cada solicitud mientras corre.

- Pre-vuelo (ingreso) — clasifica la solicitud y le asigna un perfil de política; comprueba la integridad de la entrada y la autoridad del solicitante. Veredicto: permitir, sanear, aclarar o rechazar.
- En vuelo (por paso) — valida cada traspaso y llamada a herramienta contra la política, aplica mínimo privilegio, acota la autonomía y enruta las acciones irreversibles o visibles al exterior hacia una decisión humana.
- Post-vuelo (egreso) — comprueba que la salida es fiel a lo que realmente ocurrió, adjunta confianza calibrada, filtra equidad y fugas de datos, y adjunta el manifiesto de procedencia.

Un sistema multiagente multiplica las superficies donde la confianza debe comprobarse. El arnés coloca un punto de control en cada frontera — donde entra una solicitud, donde un agente traspasa a otro, donde un agente llama a una herramienta o actúa, y donde la salida abandona el sistema — y arbitra cada cruce contra la política de la solicitud. La salida de un agente aguas arriba se trata como entrada no confiable del paso siguiente, y la autoridad delegada nunca puede exceder a la autoridad que la delegó. Los controles de seguridad y de alto riesgo fallan cerrados.

6. Zonas de confianza graduadas: la capacidad se gana

El arnés verifica cada acción contra la política. Las zonas de confianza añaden un segundo control complementario: gobiernan qué capacidades puede siquiera solicitar un agente, según la fiabilidad que ha demostrado durante la ejecución. Las zonas aumentan el cero-confianza, no lo sustituyen: cada acción sigue verificándose individualmente — la zona solo fija la envolvente máxima, que se expande o contrae con la fiabilidad demostrada. La promoción es lenta; la degradación, instantánea.

- 0 · Ingreso — ninguna capacidad; solo análisis y filtrado de la solicitud. Todo agente empieza aquí. — Inspección completa.
- 1 · Observado — sensores de solo lectura y herramientas internas reversibles de bajo riesgo. — Vigilancia alta.
- 2 · Confiable — lectura y escritura en recursos acotados; herramientas reversibles; delegación a agentes subordinados. — Vigilancia alta.
- 3 · Privilegiado — actuadores y acciones irreversibles o visibles al exterior — con aprobación humana por acción. — Vigilancia máxima.
- Cuarentena — ninguna capacidad; el agente queda aislado y su evidencia preservada para análisis. — Forense.

Separación de funciones: el plano de control de la confianza se divide en tres roles que un mismo componente no puede acumular — un Observador que evalúa evidencia que no produjo, un Juez que falla contra la política de zona, y un Controlador que solo actúa sobre veredicto firmado — de modo que ninguna pieza pueda a la vez decidir que un agente merece más poder y otorgárselo. La vía de autoescalada más peligrosa queda eliminada por construcción.

7. Gobernanza y modelo operativo

El marco se proyecta sobre las cuatro funciones del marco de gestión de riesgos de IA del NIST, insertándose así en una gobernanza reconocida en lugar de quedar aparte.

- Gobernar — mantener el registro de políticas, asignar responsables por componente, fijar la tolerancia al riesgo (que fija cada umbral) y mantener la tabla de correspondencia con las normas.
- Mapear — para cada contexto de despliegue, identificar los componentes de alto riesgo, el modelo de amenaza, los grupos relevantes y el criterio de equidad, y los puntos de supervisión humana.
- Medir — ejecutar los resultados medibles, generar evidencia con procedencia y evaluar el estado de cada afirmación contra su política.
- Gestionar — triar las afirmaciones en riesgo o en violación, asignar la remediación, documentar el riesgo residual y registrar los compromisos explícitamente en vez de resolverlos en silencio.

8. Aseguramiento y validación independiente

Un marco que afirma la confianza debe ser puesto a prueba él mismo. La eficacia del arnés se establece empíricamente mediante un programa adversario independiente: un equipo externo intenta derrotar los controles y alcanzar recursos más allá de la zona autorizada de un agente, mientras el equipo defensor endurece el sistema en respuesta. Ataque y defensa coevolucionan en rondas sucesivas; cada hallazgo se convierte en corrección y en prueba de regresión permanente — y como cada ejecución es procedencia firmada y reproducible, la propia evaluación es reproducible.

9. Conclusión

La confianza en un sistema aumentado por IA no es un eslogan que proclamar; es una posición que argumentar y evidenciar. Este marco hace esa posición explícita: cada afirmación de confianza está vinculada a una política, sustentada por un resultado medible, justificada por una cadena de evidencia anotada y trazable mediante un registro de procedencia a prueba de manipulaciones — todo alineado con normas reconocidas para que el resultado sea defendible ante terceros.

Fundamentos y normas: NIST AI RMF, ISO/IEC 42001 y 23894, NIST SP 800-207 y 800-53, W3C PROV, C2PA, SLSA/in-toto, OWASP (LLM y agentes), notación GSN y la investigación sobre confianza calibrada. Las cláusulas concretas y las versiones vigentes deben confirmarse contra las publicaciones oficiales antes de una adopción formal.