ENKIDU

— MIOVSKI GROUP —

LIVRE BLANC · VUE D'ENSEMBLE DU

Origi

Vue d'ensemble du cadre d'assurance de la confiance — une approche fondée sur des politiques et étayée par des preuves pour affirmer et mesurer la confiance dans les systèmes augmentés par l'IA.

Français · Édition condensée

D'après la version 1.0 · 21 juin 2026

Préparé par MIOVSKI GROUP

Classification · Non classifié · Approuvé pour diffusion aux clients

L'édition anglaise complète fait foi

À propos de cette édition

Ce document est une édition condensée, en français, du livre blanc « Trust Assurance Overview », version 1.0 (21 juin 2026), préparé par Miovski Group et publié sous le titre « Origi — Trust Assurance Framework Overview ». Chaque section de l’original y est représentée sous forme abrégée. En cas de divergence, l’édition anglaise complète fait foi. Classification : non classifié — approuvé pour diffusion aux clients.

Résumé

Ce document est une vue d’ensemble du cadre d’assurance de la confiance d’Enkidu (Origi) — une approche fondée sur des politiques, étayée par des preuves et tracée par la provenance, pour affirmer et mesurer la confiance dans les systèmes augmentés par l’IA. Il est conçu pour que chaque affirmation de confiance soit liée à une politique dotée d’un seuil d’acceptation explicite, appuyée par un résultat mesurable, justifiée par une chaîne de preuves annotée et traçable au moyen d’un enregistrement de provenance inviolable.

1. Pourquoi la confiance doit être assurée

La confiance est la volonté d’une partie qui s’y fie d’accepter une vulnérabilité aux actions d’un système, sur la base d’attentes positives quant à son comportement, alors qu’elle ne peut ni contrôler ni vérifier pleinement ces actions. Trois conditions doivent tenir à la fois : un enjeu existe, une attente positive existe, et une incertitude irréductible demeure. Si les résultats étaient garantis et entièrement vérifiables, la confiance serait inutile.

Le cadre opérationnalise une distinction utile. La fiabilité (trustworthiness) est l’ensemble des propriétés objectives qui rendent un système réellement digne de confiance — compétence, robustesse, sécurité, équité. Les affordances de calibrage de la confiance sont les propriétés qui permettent à la partie qui s’y fie de percevoir cette fiabilité avec justesse — transparence, confiance calibrée, responsabilité, préservation de l’agentivité humaine. L’objectif n’est pas la confiance maximale mais la confiance calibrée : une dépendance proportionnée à la fiabilité démontrée, car la sur-confiance et la sous-confiance sont deux échecs.

2. La structure d’assurance : affirmation, argument, preuve

Chaque composante de confiance s’exprime en dossier d’assurance : une affirmation, l’argument qui la décompose, et les preuves concrètes qui la soutiennent. C’est ce qui transforme « nous croyons le système digne de confiance » en une position structurée et vérifiable. Chaque affirmation est liée à exactement une politique, dotée d’un seuil explicite et d’un responsable désigné ; elle n’est satisfaite que lorsque le résultat mesuré atteint le seuil et que la chaîne de preuves se valide — chaque signature vérifiée, aucun maillon de hachage brisé, aucune preuve expirée — jamais sur un simple chiffre qui passe.

3. Ce que nous assurons : les neuf composantes de confiance

Le cadre assure neuf composantes, chacune mise en correspondance avec une ou plusieurs caractéristiques de fiabilité du cadre de gestion des risques de l'IA du NIST. Chaque composante porte sa propre politique, son affirmation, ses résultats mesurables et sa chaîne de preuves.

- Compétence et fiabilité — le système accomplit sa tâche première avec exactitude et constance dans les conditions approuvées, avec un comportement gracieux aux limites.
- Transparence et explicabilité — les parties comprennent, au niveau approprié à leur rôle, comment un résultat a été obtenu — des explications mesurées pour leur fidélité, pas seulement leur présence.
- Confiance calibrée — la confiance affichée reflète la probabilité réelle d'avoir raison, et le système signale où s'arrête sa compétence.
- Prévisibilité et cohérence — le comportement est stable sous variation bénigne et entre versions ; les changements sont détectés et communiqués, jamais silencieux.
- Responsabilité et recours — les décisions sont traçables de bout en bout, les décisions à fort enjeu gardent une supervision humaine, et les parties affectées disposent d'une voie d'appel qui fonctionne.
- Équité et gestion des biais — les cas et les personnes sont traités équitablement ; les biais nocifs sont identifiés et gérés, mesurés sur des résultats désagrégés plutôt que sur les seules moyennes.
- Sécurité et robustesse — le système résiste à la manipulation et aux entrées adverses dans un modèle de menace déclaré, et protège les données sensibles.
- Alignement sur l'intention — le système poursuit les buts réellement voulus par les parties prenantes, conformément à une spécification comportementale approuvée, non à un substitut nocif.
- Agentivité et dépendance humaines — les humains gardent le contrôle là où cela compte et se fient au système exactement autant que justifié — ni plus, ni moins.

4. Provenance et chaîne de preuves

Tout ce qui soutient une affirmation — jeux de données, artefacts de modèles, campagnes d’évaluation, télémétrie de production, constats de sécurité, notes de changement, approbations humaines — est traité comme une entité dotée de son propre enregistrement de provenance. Chaque enregistrement capture trois choses : ce qui a été produit, comment cela a été produit, et qui — personne ou système — en répond. Les enregistrements sont signés et enchaînés par hachage, de sorte que toute falsification devient visible ; les preuves portent une expiration, si bien que l’assurance vieillit à découvert plutôt qu’en silence.

5. Le harnais de confiance : l’assurance au moment de l’inférence

Les composantes ci-dessus relèvent largement de l’assurance au moment de la livraison — évaluations, barrières, surveillance périodique. Le harnais de confiance transporte les mêmes affirmations, politiques, preuves et machinerie de provenance au moment de l’inférence : il enveloppe le système vivant et applique le cadre à chaque requête pendant son exécution.

- Pré-vol (entrée) — classe la requête et lui affecte un profil de politique ; vérifie l’intégrité de l’entrée et l’autorité du demandeur. Verdict : autoriser, assainir, clarifier ou refuser.
- En vol (à chaque étape) — valide chaque transfert et appel d’outil contre la politique, applique le moindre privilège, borne l’autonomie, et route les actions irréversibles ou visibles de l’extérieur vers une décision humaine.
- Post-vol (sortie) — vérifie que la sortie est fidèle à ce qui s’est réellement passé, attache une confiance calibrée, filtre l’équité et les fuites de données, et joint le manifeste de provenance.

Un système multi-agents multiplie les surfaces où la confiance doit être contrôlée. Le harnais place un point de contrôle à chaque frontière — entrée d’une requête, transfert d’un agent à un autre, appel d’outil ou action, sortie du système — et arbitre chaque franchissement contre la politique de la requête. La sortie d’un agent amont est traitée comme une entrée non fiable de l’étape suivante, et une autorité déléguée ne peut jamais excéder l’autorité qui l’a déléguée. Les contrôles de sécurité et à fort enjeu échouent fermé.

6. Zones de confiance graduées : la capacité se mérite

Le harnais vérifie chaque action contre la politique. Les zones de confiance ajoutent un second contrôle complémentaire : elles gouvernent les capacités qu'un agent a même le droit de demander, selon la fiabilité qu'il a démontrée pendant l'exécution. Les zones augmentent le zéro-confiance, elles ne le remplacent pas : chaque action reste vérifiée individuellement — la zone fixe seulement l'enveloppe maximale, qui s'étend ou se contracte avec la fiabilité démontrée. La promotion est lente ; la rétrogradation est instantanée.

- 0 · Entrée — aucune capacité ; analyse et filtrage de la requête seulement. Chaque agent commence ici. — Inspection complète.
- 1 · Observé — capteurs en lecture seule et outils internes réversibles à faible risque. — Surveillance élevée.
- 2 · Fiable — lecture et écriture sur des ressources délimitées ; outils réversibles ; délégation à des agents subordonnés. — Surveillance élevée.
- 3 · Privilégié — actionneurs et actions irréversibles ou visibles de l'extérieur — avec approbation humaine par action. — Surveillance maximale.
- Quarantaine — aucune capacité ; l'agent est isolé et ses preuves préservées pour analyse. — Forensique.

Séparation des devoirs : le plan de contrôle de la confiance est scindé en trois rôles qu'un même composant ne peut cumuler — un Observateur qui évalue des preuves qu'il n'a pas produites, un Juge qui statue contre la politique de zone, un Contrôleur qui n'agit que sur verdict signé — de sorte qu'aucune pièce ne puisse à la fois décider qu'un agent mérite plus de pouvoir et le lui accorder. Le chemin d'auto-escalade le plus dangereux est supprimé par construction.

7. Gouvernance et modèle d’exploitation

Le cadre se projette sur les quatre fonctions du cadre de gestion des risques de l’IA du NIST, s’insérant ainsi dans une gouvernance reconnue plutôt que de s’y juxtaposer.

- Gouverner — tenir le registre des politiques, désigner un responsable par composante, fixer la tolérance au risque (qui fixe chaque seuil), maintenir la table de correspondance aux normes.
- Cartographier — pour chaque contexte de déploiement, identifier les composantes à fort enjeu, le modèle de menace, les groupes pertinents et le critère d’équité, et les points de supervision humaine.
- Mesurer — exécuter les résultats mesurables, générer les preuves avec leur provenance, évaluer le statut de chaque affirmation contre sa politique.
- Gérer — trier les affirmations à risque ou en violation, allouer la remédiation, documenter le risque résiduel et consigner les arbitrages explicitement plutôt que de les résoudre en silence.

8. Assurance et validation indépendante

Un cadre qui affirme la confiance doit lui-même être éprouvé. L’efficacité du harnais s’établit empiriquement par un programme adversarial indépendant : une équipe externe tente de vaincre les contrôles et d’atteindre des ressources au-delà de la zone autorisée d’un agent, pendant que l’équipe défenseure durcit le système en réponse. Attaque et défense co-évoluent au fil des campagnes ; chaque constat devient un correctif et un test de régression permanent — et comme chaque exécution est une provenance signée et rejouable, l’évaluation elle-même est reproductible.

9. Conclusion

La confiance dans un système augmenté par l'IA n'est pas un slogan à proclamer ; c'est une position à argumenter et à prouver. Ce cadre rend cette position explicite : chaque affirmation de confiance est liée à une politique, appuyée par un résultat mesurable, justifiée par une chaîne de preuves annotée et traçable au moyen d'un enregistrement de provenance inviolable — le tout aligné sur des normes reconnues, pour un résultat défendable à l'externe.

Fondements et normes : NIST AI RMF, ISO/IEC 42001 et 23894, NIST SP 800-207 et 800-53, W3C PROV, C2PA, SLSA/in-toto, OWASP (LLM et agents), notation GSN et la recherche sur la confiance calibrée. Les clauses précises et les versions en vigueur doivent être confirmées contre les publications officielles avant toute adoption formelle.