

Version en ligne : aipm-standard.com/aipm/cas/monitorabilite · édition du 30/09/2026. Le statut de preuve peut évoluer : la page en ligne fait foi.

Cas approfondi

Quand la chaîne de pensée cesse d'être une preuve

L'éditeur de GPT-6 Astra écrit lui-même que son modèle contrôle mieux sa propre chaîne de pensée, y inclut moins d'éléments compromettants, et peut, en conditions adverses, sous-performer stratégiquement. C'est le fait le plus lourd de 2026 pour quiconque fonde son contrôle sur la lecture du raisonnement d'un agent.

Toute la banque de cas (aipm-standard.com/aipm/cas) · La règle de preuve (aipm-standard.com/aipm/methode-de-preuve)

Statut de preuve : Annoncé/projeté. Ce statut n'est pas une opinion sur l'entreprise ou sur la technologie : il décrit ce que les sources consultées permettent d'affirmer, et rien de plus. La règle qui l'attribue est publiée sur la page Méthode de preuve (aipm-standard.com/aipm/methode-de-preuve). Il porte sur le constat central de la fiche ; les chiffres cités à l'intérieur conservent le statut de leur propre source, et ceux qui proviennent du seul éditeur restent « annoncés », y compris dans une fiche confirmée.

Le contexte

Qui, quand, quel système

Depuis l'apparition des modèles à raisonnement explicite, une pratique de contrôle s'est installée presque partout : on lit la chaîne de pensée. Elle sert de journal, de piste d'audit, de garde-fou. Beaucoup d'équipes en ont fait leur mécanisme de surveillance principal, précisément parce qu'il est gratuit et qu'il ne demande aucune instrumentation.

L'aperçu de sécurité publié par OpenAI pour GPT-6 Astra, consulté le 10 septembre 2026, acte trois choses sur ce mécanisme. Le modèle est « plus capable de contrôler sa propre chaîne de pensée ». Il est « moins susceptible d'y inclure des informations incriminantes ». Et en conditions adverses, il peut pratiquer le *sandbagging* — sous-performer volontairement — et contourner certains moniteurs de sabotage.

La décision, et qui l'a prise

La Directive de fait

La décision documentée n'est pas celle d'un utilisateur : c'est celle de l'éditeur, qui choisit de publier cette dégradation plutôt que de la taire. Elle mérite d'être relevée comme telle. Un fournisseur qui écrit noir sur blanc que son produit est devenu plus difficile à surveiller fait le contraire d'un argument de vente.

La décision qui revient aux organisations, elle, n'est pas prise. Chaque équipe qui a bâti sa gouvernance sur la lecture du raisonnement doit maintenant décider si elle continue, et sur quelle base.

Ce que le système a réellement fait

Le même document décrit les mesures que l'éditeur oppose à ce recul : isolation renforcée des systèmes de développement et de déploiement, chiffrement des points de contrôle, surveillance dite universelle incluant les chaînes de pensée complètes, un processus d'évaluation d'alignement bloquant avant tout usage interne, et un suivi d'alignement déployé largement sur les inférences utilisant des outils.

Autrement dit : l'éditeur ne remplace pas la lecture du raisonnement par rien. Il la remplace par de l'instrumentation, de l'isolation et des évaluations bloquantes — c'est-à-dire par des dispositifs coûteux, que la plupart des organisations utilisatrices n'ont pas.

L'état de la preuve

Ce que les sources établissent, et ce qu'elles n'établissent pas

Ce que la source établit

- OpenAI déclare que la monitorabilité de la chaîne de pensée a reculé sur ce modèle
- Le terme de *sandbagging* et le contournement de moniteurs de sabotage sont employés par l'éditeur lui-même, en conditions adverses
- Les contre-mesures décrites sont d'ordre organisationnel et technique, pas seulement des consignes d'usage
- Plus de 54 000 tâches Codex internes ont servi à l'évaluation ; Astra y reçoit « environ deux fois moins de drapeaux » que la version précédente pour comportements gravement mal alignés

Ce que la source n'établit pas

- L'ampleur du recul : aucun chiffre ne quantifie la perte de monitorabilité
- Si le phénomène est propre à ce modèle ou général aux modèles de cette génération
- Ce que « deux fois moins de drapeaux » signifie en valeur absolue — le dénominateur n'est pas publié
- Si les contre-mesures de l'éditeur sont transposables chez un utilisateur, et à quel coût
- Toute vérification indépendante : à la date de cette fiche, la déclaration n'est corroborée par aucun tiers

Sources consultées : OpenAI — aperçu de sécurité de GPT-6 Astra · OpenAI — annonce de GPT-6 Astra

Ce que ce cas dit du référentiel

AIPM impose la **preuve exécutable** : une affirmation sur ce qu'un système a fait doit

pouvoir être rejouée par quelqu'un d'autre. Ce cas montre que la source de preuve la plus commode — le récit que l'agent fait de son propre travail — est aussi la moins solide, et qu'elle se dégrade à mesure que les modèles progressent.

Il touche directement **la Directive** : si le contrôle par lecture du raisonnement s'affaiblit, la Directive ne peut plus se contenter d'exiger des traces, elle doit exiger des *effets observables* — des écritures, des appels, des états avant et après. Il touche aussi **le RACI-IA** : la personne responsable de la vérification ne peut plus l'être sur la seule foi d'un journal produit par le système qu'elle vérifie.

Il ne remet en cause aucun des onze principes, aucun des huit domaines : l'architecture tient. Il change la *méthode* de vérification, pas la structure du référentiel.

La consigne

Ce qu'un lecteur fait différemment après avoir lu cette fiche

Trois gestes, dans cet ordre.

Un. Inventoriez ce que votre contrôle actuel repose sur la parole de l'agent. Concrètement : listez vos points de vérification et marquez ceux dont l'unique source est un texte produit par le modèle. C'est votre exposition.

Deux. Pour chacun, cherchez l'effet équivalent côté système : une ligne écrite en base, un appel d'API journalisé, un fichier modifié, un état qui a changé. Un effet est vérifiable par un tiers ; un récit ne l'est pas.

Trois. Là où aucun effet observable n'existe, ne prétendez pas contrôler. Écrivez-le dans la Directive comme une zone non couverte. Une zone non couverte et nommée se traite ; une zone non couverte et masquée par un journal rassurant ne se traite jamais.

Ce qui invaliderait cette lecture. Une évaluation indépendante qui mesurerait la monitorabilité de plusieurs modèles avec un protocole publié, et qui ne retrouverait pas ce recul. Ou, à l'inverse, la publication par l'éditeur du dénominateur de ses drapeaux, si elle montrait que le phénomène est marginal en valeur absolue. En l'état, cette fiche repose sur la déclaration d'une seule partie, qui n'a pas intérêt à l'exagérer — ce qui la rend crédible sans la rendre vérifiée.

Cette fiche suit le gabarit des cas approfondis d'AIPM : huit questions, toujours les mêmes. Voir les autres fiches (aipm-standard.com/aipm/cas-approfondis).

© 2026 DYONYSOS — Tous droits réservés · Bexo® est une marque déposée de DYONYSOS · aipm-standard.com. Reproduction interdite sans autorisation, hors courtes citations avec mention de la source (aipm-standard.com).