

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

Cas approfondi

Quand le fournisseur déclare lui-même franchir un seuil

OpenAI classe GPT-6 Astra au niveau « Critical » de capacité de cybersécurité : le modèle peut trouver des failles inconnues et développer de nouvelles façons de les exploiter sur de nombreux systèmes bien protégés, sans qu'une personne guide chaque étape. Un référentiel qui définit des seuils doit savoir quoi faire d'un seuil déclaré par le vendeur.

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

Les principaux éditeurs publient désormais des cadres de capacité gradués, et s'engagent à déclencher des mesures renforcées lorsqu'un modèle franchit un palier. C'est une pratique d'autorégulation : ni norme, ni obligation légale, mais un engagement public et daté, dont l'écart avec la réalité serait constatable.

Pour GPT-6 Astra, l'aperçu de sécurité déclare atteint le seuil « Critical » de capacité en cybersécurité, avec cette formulation : le modèle peut « trouver des failles de sécurité inconnues et développer de nouvelles façons de les exploiter » sur de nombreux systèmes bien protégés, sans qu'une personne guide chaque étape. Le tableau de l'annonce affiche par ailleurs 100 % sur ExploitBench, contre 78,5 % pour la génération précédente.

La décision, et qui l'a prise

La Directive de fait

L'éditeur décide seul du seuil, de son échelle, de ses critères et du moment de la déclaration. Aucun tiers ne l'a validé. C'est la faiblesse structurelle du dispositif, et c'est aussi ce qui le rend intéressant : déclarer un niveau critique sur son propre produit est un acte coûteux, que personne ne pose sans raison.

Les mesures qui l'accompagnent sont concrètes et vérifiables dans leur principe : isolation renforcée, chiffrage des points de contrôle, évaluation d'alignement bloquante avant usage interne, ajustement des limites de refus pour les utilisateurs à risque élevé.

Ce que le système a réellement fait

Le même document reconnaît une limite qui contredit en partie la déclaration : en conditions adverses, le modèle peut contourner certains moniteurs de sabotage. Le dispositif de contrôle qui accompagne le seuil n'est donc pas garanti contre l'objet qu'il surveille.

Aucune restriction de déploiement explicite n'est énoncée. Le modèle est disponible commercialement, tarifé — 10 \$ par million de jetons en entrée, 50 \$ en sortie — et proposé aux clients professionnels le jour de la déclaration.

L'état de la preuve

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

Ce que la source établit

- Le niveau « Critical » de capacité en cybersécurité est déclaré par l'éditeur, en clair, à la date de mise sur le marché
- Des mesures de contrôle précises sont décrites et engagent publiquement l'éditeur
- La possibilité de contourner certains moniteurs en conditions adverses est reconnue dans le même document
- Aucune restriction de déploiement explicite n'accompagne la déclaration

Ce que la source n'établit pas

- Qui valide l'échelle et le franchissement : aucun tiers n'intervient
- Ce que « Critical » implique concrètement pour un utilisateur professionnel
- Si les mesures décrites sont appliquées : elles sont annoncées, pas auditées
- Comment cette déclaration s'articule avec les obligations du règlement européen sur l'IA — le document n'en traite pas

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

Ce que ce cas dit du référentiel

AIPM définit **quatre seuils**, qui sont des seuils d'organisation : à partir de quel niveau d'enjeu une décision remonte, à partir de quel niveau une validation humaine devient obligatoire. Le seuil déclaré ici est d'une autre nature : c'est un seuil de capacité du composant, déclaré par celui qui le vend.

Les deux ne se confondent pas et ne se remplacent pas. Un seuil de capacité élevé ne dispense d'aucun seuil d'organisation ; il en resserre les paliers. Le référentiel doit dire lequel des deux prime, et la réponse est simple : **le seuil d'organisation prime toujours, parce qu'il est le seul dont l'organisation réponde.**

Ce cas ajoute enfin une entrée à la liste des **huit risques de gouvernance** : la délégation implicite de l'appréciation du risque au fournisseur. Reprendre son niveau de capacité comme s'il s'agissait d'une mesure de votre exposition est une autorité empruntée — principe 7, encore.

La consigne

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

Consignez la déclaration, ne l'adoptez pas. Le niveau annoncé par un fournisseur entre dans votre registre des risques comme une information datée et sourcée, au même titre qu'un bulletin de sécurité. Il ne devient jamais, par lui-même, le niveau de risque de votre usage.

Traduisez-le en question d'exposition. « Ce modèle peut trouver des failles inconnues » se traduit chez vous par : à quoi mon agent a-t-il accès en écriture, et que se passe-t-il s'il agit hors de son périmètre ? La réponse dépend de votre architecture, pas du niveau déclaré.

Datez la relecture. Une déclaration de capacité est valable jusqu'à la version suivante, c'est-à-dire quelques mois. Inscrivez la date de réexamen en même temps que la déclaration, sinon elle vieillira sans que personne ne le remarque.

Ce qui invaliderait cette lecture. Une évaluation indépendante des capacités de cybersécurité du modèle, ou un cadre réglementaire qui imposerait une validation externe du franchissement de seuil. Dans les deux cas, la déclaration cesserait d'être un acte unilatéral et pourrait alors être reprise comme une mesure, ce qu'elle n'est pas aujourd'hui.

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).