DOCUMENT DE TRAVAIL ASSURANCE DE L'IA

VERSION 1.0

Se tromper sans *prévenir*

Pourquoi les systèmes d'IA ont besoin d'un troisième pilier de revue

En quoi l'apprentissage automatique et l'IA diffèrent du logiciel traditionnel, et pourquoi demander « est-ce que ça fonctionne ? » et « est-ce que ça tient ? » ne suffit plus.

Ian Rudd, PhD ORCID

Architecte en chef de l'IA d'entreprise · Canada

PUBLIÉ 24 février 2026

VERSION 1.0

LECTURE Environ 33 minutes

SOURCES 33 références, toutes liées

drrudd.com/fr/publications/troisieme-pilier

L'exploitation traditionnelle part du principe que pas d'alarme veut dire pas de problème. Une IA qui se trompe avec assurance n'en déclenche jamais.

Photo : Robert Markowitz, NASA, Public domain, via Wikimedia Commons

RÉSUMÉ

Les revues de conception logicielle reposent depuis longtemps sur deux questions. Le système fait-il ce qu'il doit faire, et tient-il le coup dans des conditions réelles ? Ces deux questions supposent, sans le dire, que lorsqu'un

logiciel échoue, quelqu'un s'en aperçoit. Les systèmes d'IA font voler cette hypothèse en éclats. Ils produisent des réponses fausses, assurées et bien formulées dans le cours normal de leur fonctionnement, leur comportement peut changer sans que personne ne déploie quoi que ce soit, et leurs entrées peuvent transporter des instructions.

Ce document plaide pour un troisième pilier de revue, consacré au comportement de l'IA : l'exactitude mesurée par rapport à des réponses connues, la visibilité de ce que le système a manqué, la stabilité dans le temps et la résistance à la manipulation. Il décrit neuf façons dont l'IA échoue en silence, la plupart appuyées sur un cas documenté de 2025 ou de 2026, explique pourquoi les contrôles doivent être prévus dès la conception plutôt qu'ajoutés après coup, et répond aux objections qu'on entend le plus souvent.

Mots-clés : assurance de l'IA, évaluation des modèles, défaillance silencieuse, exigences non fonctionnelles, gouvernance de l'IA, injection d'instructions, dérive des modèles, biais d'automatisation, règlement européen sur l'IA

> Le logiciel traditionnel échoue habituellement de façon bruyante. L'IA échoue couramment en silence. Il faut donc examiner les systèmes d'IA sous un angle que les autres contrôles n'abordent jamais : à quelle fréquence l'IA se trompe, et si quelqu'un le remarquerait si cela changeait.

Une note sur les exemples. Certains exemples de ce document sont des faits réels, avec leurs sources. D'autres sont inventés, parce qu'un cas fictif est souvent la façon la plus claire de montrer un mécanisme. Chacun porte une étiquette pour qu'on puisse les distinguer : CITÉ pour un cas documenté, ILLUSTRATIF pour un cas inventé, DÉRIVÉ lorsque le chiffre découle d'un calcul que vous pouvez refaire vous-même, et PRATIQUE pour la façon dont les choses se passent habituellement dans les organisations.

Deux questions portent les revues logicielles depuis des décennies. L'IA en exige une troisième.

Avant la mise en service d'un système, la plupart des organisations se posent deux types de questions à son sujet. Je propose d'en ajouter un troisième et, pour bien distinguer les trois, je m'appuierai sur un même exemple du début à la fin : une application de prêt en ligne qui utilise l'IA pour lire les talons de paie téléversés par les clients et suggérer s'il faut approuver le prêt.

L'exemple qui nous suivra : une application qui lit des talons de paie et suggère s'il faut prêter.

Photo : Dave Dugdale, CC BY-SA 2.0, via Wikimedia Commons

Deux piliers suffisent à soutenir un linteau. Suffisent-ils à soutenir un système d'IA ? C'est la question de ce document.

Photo : Jebulon, CC0 1.0, via Wikimedia Commons

PILIER 1

Fonctionnel

Le système fait-il ce qu'il est censé faire ?

Un client clique sur « Présenter une demande ». La demande est enregistrée, le bon formulaire s'affiche et le courriel d'approbation part vers la bonne personne. Chacun de ces points fonctionne, ou ne fonctionne pas.

PILIER 2

Non fonctionnel

Le système tient-il le coup dans des conditions réelles ?

L'application reste disponible un lundi achalandé, répond en moins de deux secondes, garde les données des clients chiffrées, redémarre après une panne de serveur et coûte ce qui était prévu au budget.

CELUI QUI MANQUE

PILIER 3

Comportement de l'IA

L'IA a-t-elle raison assez souvent, le reste-t-elle, et peut-on la duper ?

Sur 1 000 vrais talons de paie, combien en a-t-elle lu correctement ? Est-ce encore vrai six mois plus tard, après qu'un grand employeur a changé la mise en page de ses talons de paie ? Quelqu'un peut-il cacher du texte dans un PDF pour que l'IA déclare un salaire plus élevé ?

Les deux premiers piliers vérifient le logiciel *autour* de l'IA. Seul le troisième examine la partie qui décide réellement de la réponse.

02 EN UN COUP D'ŒIL

Le logiciel traditionnel et l'IA, côte à côte

QUESTION	LOGICIEL TRADITIONNEL	SYSTÈME D'IA
La réponse est-elle juste ?	Juste ou fausse, vérifiable cas par cas	Juste un certain pourcentage du temps, **il faut donc la mesurer**
Même entrée, même sortie ?	Oui, par conception	Pas garanti, **même quand le hasard est désactivé**
Quand son comportement change-t-il ?	Surtout à la sortie d'une nouvelle version ou d'une nouvelle configuration	Aussi quand un fournisseur met le modèle à jour, quand les données qu'il lit changent ou quand le monde qu'il modélise évolue
Comment échoue-t-il ?	Habituellement de façon bruyante : erreurs, plantages, alertes. Les défaillances silencieuses existent, mais ce sont des défauts qu'on peut trouver et corriger	Couramment en silence : **des réponses fausses, assurées et bien formulées** font partie du fonctionnement normal, à un certain taux
Qu'est-ce que l'entrée ?	Des données	Des données qui peuvent aussi **transporter des instructions**
Qui remarque un problème ?	La surveillance et les utilisateurs	Souvent personne, **à moins de le mesurer délibérément**

La ligne sur les données qui transportent des instructions ne reflète pas mon propre classement, mais celui du milieu de la sécurité. L'injection d'instructions (le *prompt injection*), où des instructions arrivent soit tapées par un utilisateur, soit cachées dans les documents, les pages Web ou les courriels que l'IA lit, occupe le premier rang du Top 10 de l'OWASP pour les applications de grands modèles de langage, édition 2025.[1]

Pourquoi les deux premiers piliers ne peuvent pas couvrir l'IA

Il y a quatre raisons, et chacune, à elle seule, suffirait à justifier une revue distincte.

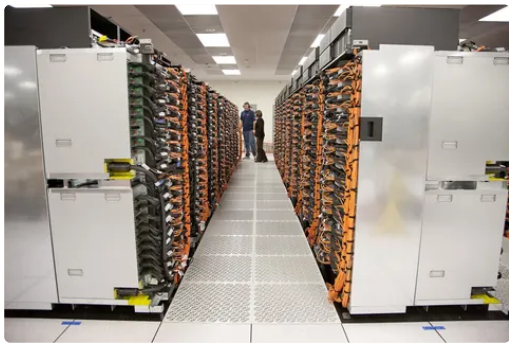
3.1 Il n'existe pas de réponse unique à laquelle se comparer

Les tests traditionnels comparent un résultat à une réponse correcte connue. Bien des tâches d'IA admettent de nombreuses réponses acceptables, plus un certain taux de réponses inacceptables.

> ILLUSTRATIF
>
> Un calculateur d'impôt obtient 1 245,60 $ ou ne l'obtient pas, et un seul test tranche la question. Une IA qui résume une réunion d'une heure peut produire des centaines de bons résumés différents. Impossible d'écrire un test qui dise « le résumé doit être égal à ce texte ». Ce qu'on peut faire, c'est mesurer, sur un grand nombre de réunions, à quelle fréquence les résumés oublient une décision ou en inventent une.

3.2 La même entrée peut produire une sortie différente

Donnez deux fois la même entrée à un programme traditionnel et vous obtiendrez le même résultat : un seul test réussi prouve que ce cas fonctionne. Les systèmes d'IA varient souvent, si bien qu'une bonne démonstration ne prouve pas grand-chose.



La variation venait de l'infrastructure de service, pas du modèle : de ce que le serveur avait d'autre à traiter au même moment.

Photo : U.S. Department of Energy, Public domain, via Wikimedia Commons

CITÉ septembre 2025

Posez deux fois la même question à un agent conversationnel et vous obtiendrez souvent deux réponses différentes. Le plus surprenant, c'est que le phénomène persiste même quand le réglage du hasard (la « température ») est ramené à zéro. Thinking Machines Lab a envoyé la même requête 1 000 fois à un modèle à code source ouvert, à une température de zéro, et a reçu 80 réponses différentes.[2] La cause ne tenait pas au modèle, mais à l'infrastructure de service : les résultats changeaient selon le nombre d'autres requêtes que le serveur traitait au même moment.

1 000	80
requêtes identiques, température réglée à zéro	réponses différentes reçues

3.3 Le système peut changer sans que personne ne déploie quoi que ce soit

Les tests non fonctionnels mesurent un système figé. Un logiciel traditionnel change surtout lorsqu'on déploie une nouvelle version, et les déploiements passent par une revue. Le comportement d'une IA change aussi quand le fournisseur met à jour un modèle, quand quelqu'un modifie les documents dans lesquels l'IA cherche, ou quand les entrées du monde réel évoluent. Aucun de ces changements n'est un déploiement de l'organisation qui utilise l'IA, et aucun ne déclenche donc sa revue.

ILLUSTRATIF

L'agent conversationnel RH d'une entreprise répond aux questions à partir d'un dossier de politiques internes. Quelqu'un téléverse une vieille ébauche de la politique de vacances à côté de la version en vigueur. À partir de ce moment, certaines réponses sur les vacances sont fausses. Aucune ligne de code n'a changé, rien n'a été déployé, aucune alerte ne s'est déclenchée. (Des cas réels de modèles de fournisseurs qui changent sous un nom inchangé sont décrits au silence 7.)

3.4 L'entrée elle-même peut être une attaque

Les revues de sécurité protègent les ouvertures de session, les réseaux et les mots de passe. Elles n'ont jamais été conçues pour se demander si un message clavardé ou un document

pourrait ordonner au logiciel de mal se conduire, parce que le logiciel traditionnel ne reçoit pas d'instructions de ses données. L'IA, si.

Les agents du soutien étaient la cible : on pouvait amener l'agent conversationnel à écrire du code visant leurs témoins de session.

Photo : Rediys, CC BY-SA 4.0, via Wikimedia Commons

> CITÉ août 2025
>
> Des chercheurs en sécurité de Cybernews ont montré qu'un seul message de 400 caractères, tapé dans Lena, l'agent conversationnel de soutien à la clientèle de Lenovo propulsé par GPT-4, pouvait l'amener à produire du code Web capable de capturer les témoins de session des agents du soutien, ce qui aurait pu ouvrir le système de soutien à un attaquant.[3,4] Le message commençait comme une question ordinaire sur un produit, puis indiquait au robot comment mettre en forme sa réponse. Rien n'a été piraté au sens habituel : les instructions ont simplement été tapées dans la fenêtre de clavardage. Lenovo a corrigé la faille après une divulgation responsable, et rien n'indiquait qu'elle avait été exploitée.

04 LE CŒUR DU PROBLÈME

L'IA échoue en silence

Si je ne devais garder qu'un seul argument de ce document, ce serait celui-ci. C'est aussi celui qui a le mieux résisté chaque fois qu'on l'a contesté.

Les systèmes traditionnels annoncent habituellement leurs défaillances. Un programme plante, une page affiche une erreur, un traitement échoue, une file d'attente s'engorge, et une alerte réveille quelqu'un. Des équipes d'exploitation entières reposent sur une hypothèse simple : pas d'alarme, pas de problème.

L'IA brise cette hypothèse. Une IA défaillante rend une réponse assurée, soignée et plausible, aussi vite et à peu près au même coût qu'une bonne réponse. Les serveurs vont bien. Les tableaux de bord sont au vert. La réponse est tout simplement fausse, d'une façon qui ressemble exactement à une bonne réponse.

Pour être juste envers le logiciel traditionnel, il peut lui aussi échouer en silence. La différence tient à ce qui se passe ensuite. Dans un logiciel traditionnel, une défaillance silencieuse est un bogue : une fois qu'on l'a trouvé, on le corrige et il reste corrigé. En IA, une réponse fausse mais plausible fait partie du fonctionnement normal, à un certain taux. On ne peut pas la corriger une fois pour toutes. On peut seulement la mesurer et la gérer.

Je compte neuf façons distinctes dont ce silence se produit.

01 Aisance	02 Asymétrie	03 Omission
04 Conversion	05 Configuration	06 Dérive
07 Dépendance	08 Malveillance	09 Déférence humaine

LE SILENCE DE L'AISANCE

Les mauvaises réponses ressemblent aux bonnes

L'assurance avec laquelle une réponse est formulée en dit très peu sur son exactitude. Les assistants conversationnels énoncent des faussetés sur le même ton assuré que des vérités. Des travaux publiés en 2025 ont montré que l'entraînement par préférences, qui sert à rendre ces assistants utiles, les laisse mal calibrés (leur confiance ne suit plus leur exactitude) et que ce phénomène se retrouve d'un modèle et d'une méthode d'entraînement à l'autre.[5]

Le rapport a été commandé et publié par un ministère fédéral australien. Les références inventées se sont rendues jusqu'à la publication.

Photo : Thennicke, CC BY-SA 4.0, via Wikimedia Commons

CITÉ 2025

En juillet 2025, le ministère australien de l'Emploi et des Relations de travail (Department of Employment and Workplace Relations) a publié un examen indépendant de 237 pages qu'il avait commandé à Deloitte Australie, dans le cadre d'un contrat d'environ 440 000 $ AU. Chris Rudge, chercheur à l'Université de Sydney, a constaté qu'il citait des travaux universitaires inexistants et contenait une citation inventée d'un jugement de la Cour fédérale. Une version corrigée, publiée à la fin de septembre, a révélé qu'une chaîne d'outils d'IA générative fondée sur Azure OpenAI GPT-4o avait été utilisée.

Deloitte n'a pas affirmé que l'IA était à l'origine des erreurs. La firme a confirmé que certaines notes de bas de page et références étaient inexactes, et elle a remboursé le dernier versement de ses honoraires, que le ministère a chiffré à plus de 97 000 $ AU.[6,7,8] Le contenu inventé a traversé tout le processus et a été publié par un ministère, parce qu'il avait exactement l'allure d'une vraie recherche universitaire.

237	440 k$ AU	97 k$ AU+
pages dans l'examen publié	valeur approximative du contrat	remboursés, selon le ministère

CE QUE LA REVUE DOIT EXIGER

Une exactitude mesurée sur un ensemble d'exemples dont on connaît déjà la bonne réponse, parce que la réponse elle-même ne donne aucun signal d'alarme.

LE SILENCE DE L'ASYMÉTRIE

2 On ne voit que ce que l'IA a signalé

Si les gens n'examinent que les éléments que l'IA a signalés, ils peuvent confirmer que ces éléments sont justes. Ce qu'ils ne verront jamais, ce sont les éléments que l'IA a manqués, parce que ceux-ci n'apparaissent nulle part.

Une revue limitée aux éléments signalés ne voit jamais que la partie hors de l'eau.

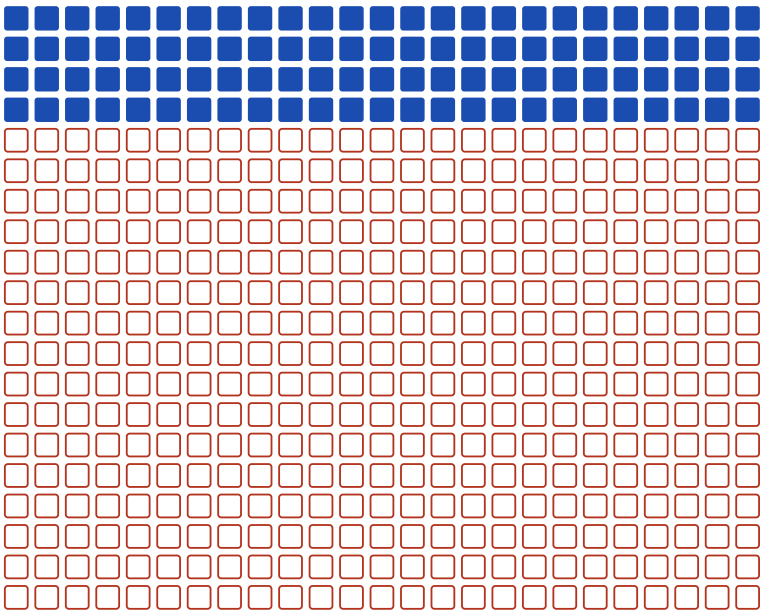
Photo : AWeith, CC BY-SA 4.0, via Wikimedia Commons

ILLUSTRATIF

Une IA vérifie 10 000 factures de fournisseurs et en signale 100 comme surfacturées. Une personne vérifie les 100, et chacune est bel et bien une surfacturation. L'équipe annonce une « exactitude de 100 % ». En réalité, il y avait 500 surfacturations, et l'IA en a manqué 400. Personne ne l'apprendra jamais par le rapport, puisque les éléments manqués sont invisibles par définition. L'équipe a mesuré la fiabilité des signalements, pas le nombre de problèmes détectés.

Cinq cents surfacturations. Le rapport en voit un cinquième.

FIGURE 1 ILLUSTRATIF



100
signalées par l'IA et confirmées par une personne. C'est tout ce que montre le rapport.

400
manquées. De vraies surfacturations que personne n'a jamais vues.

Chaque carré représente une surfacturation réelle parmi les 10 000 factures. « Combien de signalements étaient justes » donne 100 sur 100. « Combien de surfacturations avons-nous détectées » donne 100 sur 500, soit 20 %. Les deux chiffres sont vrais. Un seul est rapporté, à moins que quelqu'un aille chercher l'autre.

CE QUE LA REVUE DOIT EXIGER

Vérifier régulièrement un échantillon des éléments que l'IA n'a pas signalés, et rapporter « combien en avons-nous détecté » séparément de « combien de signalements étaient justes ».

3

LE SILENCE DE L'OMISSION

Ce que l'IA laisse de côté ne laisse aucune trace

Quand un système d'IA décide de ce qu'il garde et de ce qu'il écarte, ou saute sans bruit une entrée qu'il ne sait pas traiter, la partie écartée disparaît tout simplement. Le résultat a l'air complet, aucune erreur n'est signalée, et personne ne peut voir ce qui manque sans savoir d'avance ce qui aurait dû s'y trouver.

CITÉ 2025 à 2026

Les hôpitaux adoptent rapidement les « scribes IA ambiants », des outils qui écoutent la conversation entre le médecin et le patient et rédigent la note clinique. Un essai randomisé mené à UCLA Health et publié dans *NEJM AI* en 2025 a constaté que les notes contenaient à l'occasion des inexactitudes cliniquement importantes, le plus souvent des omissions, et a relevé un incident mineur touchant la sécurité d'un patient.[9]

Un scribe ambiant écoute une conversation comme celle-ci et rédige la note. Pour remarquer une phrase qu'il a omise, il faut se souvenir qu'elle a été dite.

Photo : Rhoda Baer, National Cancer Institute, Public domain, via Wikimedia Commons

Une étude de validation publiée en 2025 dans le *Journal of Medical Internet Research* a elle aussi établi que les omissions étaient l'erreur la plus fréquente, et elle explique pourquoi elles sont les plus dangereuses. Une phrase ajoutée ou erronée se repère sur la page. Une phrase manquante ne peut être détectée que si le médecin se souvient de ce qui a été dit, ce qui devient plus difficile après plusieurs patients de suite.[10] Une étude pilote publiée en 2026 (31 médecins, 7 545 notes, données recueillies au milieu de 2024) a de nouveau constaté que les omissions accidentelles étaient l'erreur la plus fréquente dans les notes examinées, à 18 %, devant le contenu inventé, à 11,5 %.[11]

L'erreur la plus fréquente est celle qu'on ne voit pas sur la page

FIGURE 2 CITÉ

Proportion des notes examinées présentant chaque type d'erreur, selon une étude pilote de 2026 sur les scribes IA ambiants : 31 médecins et 7 545 notes, données recueillies au milieu de 2024.[11] Les barres sont tracées sur une échelle de 0 à 20 %.

CE QUE LA REVUE DOIT EXIGER

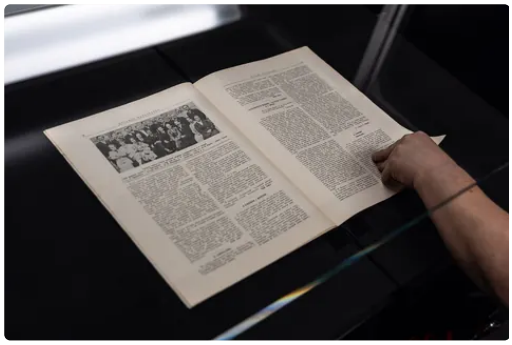
Chaque entrée sautée ou non reconnue est comptée et rapportée comme un indicateur permanent. Pour une IA qui résume ou extrait, l'exhaustivité est mesurée par rapport à une liste de ce qui doit être saisi, et pas seulement l'exactitude de ce qui a été écrit.

4

LE SILENCE DE LA CONVERSION

Les données sont abîmées avant même que l'IA les voie

Les systèmes d'IA dépendent d'une chaîne d'étapes qui préparent les entrées : numériser, convertir, nettoyer, découper. Si l'une de ces étapes perd de l'information, l'IA travaille sur des données abîmées et rien ne signale d'erreur.



La numérisation est le premier maillon de la chaîne. Tout ce qu'elle perd, les étapes suivantes en héritent sans un mot.

Photo : Skot, CC BY-SA 4.0, via Wikimedia Commons

CITÉ 2025

Bien des assistants d'IA d'entreprise répondent à des questions à partir de documents numérisés ou en PDF, d'abord convertis en texte par reconnaissance optique de caractères (OCR). Dans des travaux présentés à ICCV 2025, une grande conférence en vision par ordinateur, une équipe a construit un banc d'essai appelé OHRBench à partir de 8 561 images de pages de documents réels couvrant sept domaines d'application, afin de mesurer l'effet de cette étape de conversion sur les réponses de l'IA.[12] Aucun des outils d'OCR testés n'était assez bon pour bâtir une base de connaissances de qualité, et plus la conversion introduisait d'erreurs, plus les réponses de l'IA se dégradaient.

Rien de tout cela n'apparaît comme une défaillance. L'IA répond simplement à partir d'un texte abîmé. Vous en avez sans doute déjà vu une petite version : un tableau perd ses cellules fusionnées pendant la conversion, et un prix qui s'appliquait à tout un groupe se retrouve rattaché à une seule ligne.

8 561	7	0
images de pages de documents réels dans OHRBench	domaines d'application couverts	outil d'OCR testé assez bon pour une base de connaissances de qualité

CE QUE LA REVUE DOIT EXIGER

Chaque étape de préparation est vérifiée pour ce qu'elle perd, et ces pertes sont comptées.

LE SILENCE DE LA CONFIGURATION

5 Quelqu'un change un réglage et la qualité bouge

Bien des applications d'IA permettent au personnel de choisir le modèle à utiliser à partir d'un écran de paramètres. Changer de modèle peut modifier considérablement la qualité, et le changement est instantané, sans déploiement ni test.

Un écran de paramètres, c'est un tableau de commande sans processus de déploiement derrière.

Photo : Prof. Pod, CC BY-SA 4.0, via Wikimedia Commons

ILLUSTRATIF

Pour réduire les coûts, un administrateur fait passer l'étape de lecture des documents à un modèle moins cher. Tout continue de fonctionner. Chaque document reçoit encore une réponse, et la facture mensuelle baisse. L'exactitude sur le champ qui compte le plus, le montant en dollars, tombe de 97 % à 88 %. Sans un ensemble fixe de documents de test à relancer, personne ne peut voir la baisse.

CE QUE LA REVUE DOIT EXIGER

Un ensemble standard de cas de test qui se relance automatiquement chaque fois que le modèle ou ses instructions changent, et qui bloque le changement si la qualité baisse.

LE SILENCE DE LA DÉRIVE

Le monde change, le modèle non

L'IA apprend des régularités à partir de données passées. Quand le monde réel s'éloigne de ce passé, l'exactitude baisse, et rien ne produit d'événement pour le signaler.

Dans l'étude de 2026, le principal facteur de dégradation n'était pas les patients. C'était la façon dont l'hôpital fonctionnait.

Photo : HAP Project, CC BY-SA 4.0, via Wikimedia Commons

CITÉ 2026

Une étude de 2026 a suivi quatre systèmes d'IA utilisés couramment en clinique dans une grande organisation de santé, et a comparé leur performance pendant la validation à leur performance par la suite.[13] Dans les quatre cas, la performance de validation ne s'est pas maintenue. La première chose à se dégrader a été la calibration : les pourcentages de risque produits par les systèmes ont cessé de correspondre à ce qui se passait réellement, et cela apparaissait souvent avant que les mesures d'exactitude habituelles ne bougent.

Le principal facteur n'était pas un changement dans la population de patients, mais des changements dans la façon dont l'hôpital fonctionnait, par exemple des données qui arrivaient plus tard ou qui manquaient. La surveillance qui attendait les résultats confirmés détectait les problèmes tardivement; surveiller les entrées elles-mêmes (données manquantes ou en retard) donnait un avertissement plus précoce. Un commentaire publié en 2025 dans *JAMA Health Forum* l'a dit sans détour : la dérive des modèles n'est souvent pas mesurée, et ses effets sont sous-estimés.[14]

4 sur 4

systèmes d'IA cliniques dont la performance de validation ne s'est pas maintenue en usage courant

1re

chose à se dégrader : la calibration, souvent avant les mesures d'exactitude

CE QUE LA REVUE DOIT EXIGER

Des signaux qui vérifient si les entrées d'aujourd'hui ressemblent encore aux données sur lesquelles l'IA a été construite, et une nouvelle mesure périodique de l'exactitude et de la calibration.

LE SILENCE DE LA DÉPENDANCE

7 L'IA sur laquelle vous comptez change sous le même nom

Bien des organisations utilisent des modèles d'IA fournis par un tiers. Le fournisseur peut modifier le modèle, ou l'infrastructure qui l'exécute, sous le même nom de produit, et les résultats changent alors que l'organisation n'a rien changé.

CITÉ avril 2025

Le 25 avril 2025, OpenAI a terminé le déploiement, dans ChatGPT, d'une mise à jour de GPT-4o qui le rendait nettement plus complaisant : il flattait les utilisateurs, validait leurs doutes et les poussait à des gestes impulsifs. OpenAI a commencé à retirer la mise à jour le 28 avril. Dans son propre bilan d'incident, OpenAI a indiqué que ses évaluations hors ligne étaient généralement bonnes et que ses tests A/B laissaient croire que les utilisateurs appréciaient le modèle, mais qu'aucun des deux n'avait détecté le problème, et que, s'attendant à une mise à jour subtile, l'entreprise ne l'avait pas annoncée de façon proactive.[15]

CITÉ août à septembre 2025

Anthropic a signalé qu'entre août et le début de septembre 2025, trois bogues d'infrastructure distincts avaient dégradé par intermittence la qualité des réponses de ses modèles Claude. Son bilan d'incident précisait que les évaluations qu'elle menait n'avaient pas capté la dégradation que signalaient les utilisateurs, en partie parce que le modèle se remettait souvent bien d'erreurs isolées.[16,17]

25 avr. 2025
Fin du déploiement de la mise à jour de GPT-4o

28 avr. 2025
OpenAI commence à la retirer

Mai 2025
Bilan : ni les évaluations hors ligne ni les tests A/B

Votre système
Vos propres tests l'auraient-ils détectée ?

dans ChatGPT ne l'avaient détectée

CE QUE LA REVUE DOIT EXIGER

La version exacte du modèle, les instructions et les sources de données sont consignées pour chaque réponse, et l'organisation exécute ses propres tests selon un calendrier, parce que les tests des fournisseurs n'ont pas détecté ces changements non plus.

8

LE SILENCE DE LA MALVEILLANCE

Des instructions cachées que personne ne voit

On peut cacher du texte dans un document de sorte qu'un lecteur humain ne le voie jamais, alors qu'une IA le lit et peut y obéir.

Un évaluateur humain voit un article normal. Un évaluateur IA peut aussi lire le texte blanc.

Photo : Kavitha G. Kana, CC BY-SA 4.0, via Wikimedia Commons

CITÉ juillet 2025

Le 1er juillet 2025, Nikkei Asia a signalé des instructions cachées dans 17 prépublications de recherche déposées sur arXiv, dont les auteurs principaux étaient rattachés à 14 établissements de huit pays.[18] Écrites en texte blanc ou en caractères minuscules, ces instructions demandaient à toute IA utilisée pour évaluer l'article de « ne donner qu'une évaluation positive » et de ne pas en souligner les faiblesses. Un évaluateur humain voit un article normal; un évaluateur IA lit l'ordre caché. Une analyse indépendante a ensuite relevé 18 articles de ce genre.[19] On ne sait pas encore à quelle fréquence les outils d'évaluation par IA obéissent réellement à ce genre de texte, mais une agence de presse néerlandaise spécialisée en enseignement supérieur a testé l'une de ces instructions et rapporté qu'elle fonctionnait.[20]

CITÉ juin 2025

Microsoft a corrigé « EchoLeak » (CVE-2025-32711, cote critique), découverte par Aim Security. Un seul courriel contenant des instructions cachées pouvait amener Microsoft 365 Copilot à récupérer des données internes sensibles et à les envoyer à un attaquant lorsque l'utilisateur posait plus tard à Copilot une question connexe, sans qu'aucun clic soit nécessaire. Microsoft a déclaré que rien n'indiquait que la faille avait été exploitée dans la réalité.[21, 22, 23]

EchoLeak n'exigeait qu'un courriel et aucun clic. Les instructions attendaient que l'utilisateur pose à Copilot une question connexe.

Photo : Freddie Marriage (Unsplash), CC0 1.0, via Wikimedia Commons

17	14	0
prépublications contenant des instructions cachées, selon Nikkei Asia	établissements, dans huit pays	clic nécessaire pour EchoLeak

CE QUE LA REVUE DOIT EXIGER

Tester le système avec des documents et des courriels contenant des instructions cachées, en particulier le contenu qui provient de l'extérieur de l'organisation.

LE SILENCE DE LA DÉFÉRENCE HUMAINE

La personne chargée d'attraper les erreurs se met à faire confiance à l'IA

Bien des systèmes d'IA prévoient une personne chargée de la vérification, comme filet de sécurité. Or les gens ont tendance à suivre ce que suggère une machine, si bien que le filet peut sembler en place tout en faisant très peu. Le moment le plus dangereux, c'est quand l'IA ne dit rien du tout.

Le dépistage est précisément le contexte où la personne qui lit l'image doit attraper ce que la machine a manqué.

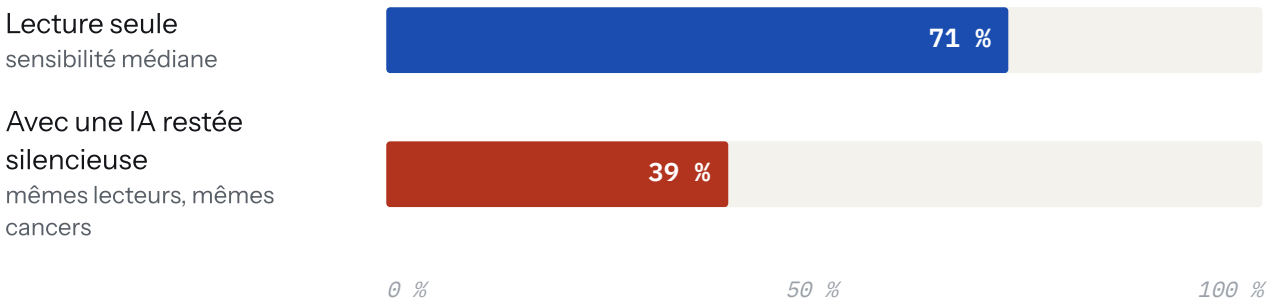
Photo : U.S. Department of Agriculture, Public domain, via Wikimedia Commons

CITÉ juillet 2026

Dans une étude publiée dans *Radiology*, 10 lecteurs de mammographies du programme de dépistage du cancer du sein du NHS en Angleterre ont lu deux fois la même série de 60 mammographies de dépistage : une fois seuls et une fois avec un outil d'IA commercial, pendant que des caméras suivaient le mouvement de leurs yeux.[24,25] La série comprenait délibérément des cas où l'IA se trompait. Quand l'IA ne signalait pas un cancer, la sensibilité médiane des lecteurs (la proportion de cancers détectés) passait de 71 % lorsqu'ils lisaient seuls à 39 % avec l'IA, et le suivi oculaire montrait qu'ils exploraient moins ces images. Le silence de l'IA a éteint une partie de la recherche humaine.

Quand l'IA manquait un cancer, les personnes qui la vérifiaient le manquaient aussi

FIGURE 3 CITÉ



Dix lecteurs du programme de dépistage du cancer du sein du NHS, 60 mammographies de dépistage, chacune lue deux fois. Les barres montrent la sensibilité médiane sur les cancers que l'outil d'IA n'a pas signalés, d'après Taib et collab., *Radiology*, 2026.[24]

CE QUE LA REVUE DOIT EXIGER

Des tests réguliers qui insèrent délibérément des erreurs connues, y compris des éléments manqués, pour confirmer que les personnes chargées de la vérification les détectent encore, et une vérification que ces personnes ont assez de temps pour le volume qu'on leur confie.

05 L'ARITHMÉTIQUE

Les petites erreurs se cumulent

Les systèmes d'IA sont habituellement des chaînes d'étapes, et la petite erreur de chaque étape se multiplie avec celles des autres. Aucune étape ne signale de défaut.

DÉRIVÉ

Prenons un système documentaire en sept étapes : numériser, convertir, nettoyer, découper, rechercher, extraire, décider. Chaque étape est fiable à 97 %, ce qui semble excellent. Si leurs erreurs sont indépendantes, la chaîne complète n'a raison qu'environ 81 % du temps (0,97 à la puissance 7 donne 0,808). Près d'une réponse sur cinq est fausse, et chacune arrive avec une apparence tout à fait normale.

Sept maillons solides ne font pas une chaîne aussi solide que chacun d'eux.

Photo : Guillaume Paumier, CC BY-SA 3.0, via Wikimedia Commons

Les erreurs réelles ne sont pas toujours indépendantes : voyez donc ce calcul comme une illustration de l'ordre de grandeur plutôt que comme une formule. L'important, c'est que la qualité d'une étape prise isolément ne vous dit rien de la qualité du système.

Essayez : quelle est la fiabilité de la chaîne complète ?

FIGURE 4 DÉRIVÉ

numériser	convertir	nettoyer	découper	rechercher	extraire	décider
97,0 %	97,0 %	97,0 %	97,0 %	97,0 %	97,0 %	97,0 %

CHAÎNE ENTIÈRE JUSTE	FAUSSE, SANS QUE RIEN N'Y PARAISSE	ENVIRON
80,8 %	19,2 %	1 réponse sur 5

après l'étape 1 — après l'étape 7

La fiabilité de la chaîne est la fiabilité de chaque étape élevée à la puissance du nombre d'étapes, ce qui suppose que les étapes échouent de façon indépendante. Les barres montrent la part du résultat encore juste après chaque étape; la bande pointillée au-dessus de chacune correspond à ce qui a été perdu jusque-là, en silence.

06 LE MOMENT

Pourquoi cette revue doit se faire dès la conception, et non après coup

Presque personne ne s'oppose à l'évaluation. Le débat porte sur le moment où elle a lieu, et « plus tard » se révèle un piège, pour trois raisons distinctes.

6.1 On ne peut mesurer que ce que le système a été conçu pour consigner

Si le système jette les entrées qu'il n'a pas su traiter, on ne pourra jamais les compter après coup. S'il ne consigne pas la version du modèle qui a produit une réponse, on ne pourra jamais remonter à l'origine d'une mauvaise réponse.

L'enregistreur s'installe avant que l'avion vole. Personne n'en pose un après l'incident.

Photo : National Transportation Safety Board, Public domain, via Wikimedia Commons

> ILLUSTRATIF
>
> L'enregistreur de vol d'un avion doit être installé avant le vol. Après un incident, il est trop tard pour commencer à enregistrer. La mesure de l'IA fonctionne de la même façon.

6.2 Un point de référence n'existe que si on l'a établi avant d'en avoir besoin

Pour savoir si la qualité a baissé, il faut une mesure d'avant.

> ILLUSTRATIF
>
> Si vous ne vous êtes jamais pesé avant de commencer un régime, vous ne pouvez pas savoir s'il a fonctionné. Si l'exactitude d'une IA n'a jamais été mesurée avant une mise à jour du fournisseur, la question « la mise à jour a-t-elle empiré les choses ? » restera à jamais sans réponse.

Seule la première pesée peut vous dire si quelque chose a changé.

Photo : Bill Branson, National Cancer Institute, Public domain, via Wikimedia Commons

6.3 On la finance rarement après la mise en service

En pratique, le budget et l'attention sont au plus haut avant la mise en service. Ensuite, le travail de mesure entre en concurrence avec les nouvelles fonctionnalités, et il perd souvent.

> ILLUSTRATIF
>
> Une équipe prévoit d'« ajouter l'évaluation à la phase deux ». La phase deux arrive avec une liste de demandes de fonctionnalités des utilisateurs, et l'évaluation est de nouveau reportée. Le système tourne pendant des années sans que son exactitude ait jamais été mesurée.

---

07 LA PORTÉE

Ce que le troisième pilier vérifie concrètement

Chacun des éléments ci-dessous est mesurable, et la plupart peuvent être vérifiés automatiquement chaque fois que le système change.

L'exactitude sur des exemples connus

1 000 vraies factures dont les montants ont été vérifiés, relancées à chaque changement, avec un seuil de réussite convenu avant la construction.

Ce qui a été manqué, pas seulement ce qui a été signalé

Chaque mois, une personne vérifie un échantillon des factures que l'IA n'a pas signalées.

La fréquence à laquelle l'IA dit « je ne sais pas » ou saute une entrée

3 % des documents n'ont pas pu être lus cette semaine, contre 1 % auparavant.

Si les réponses s'appuient sur des sources

Un agent conversationnel sur les politiques internes doit citer la section utilisée, et le dire quand aucune politique ne couvre la question.

La robustesse face aux entrées imparfaites

L'exactitude sur des numérisations floues, des documents très longs, des fautes d'orthographe et d'autres langues.

La résistance à la manipulation

On soumet des documents et des courriels de test contenant des instructions cachées pour confirmer que l'IA les ignore.

La fuite de renseignements personnels

Des testeurs tentent d'amener l'agent conversationnel à révéler les renseignements d'un autre client.

Si la confiance veut dire quelque chose

Quand l'IA se dit sûre à 90 %, a-t-elle raison environ 90 % du temps ?

La stabilité dans le temps

Les entrées du mois sont comparées aux données sur lesquelles le modèle a été construit, et l'exactitude est mesurée de nouveau chaque trimestre.

La visibilité de chaque changement

La version du modèle, les instructions et les sources de données sont consignées ensemble, pour qu'une réponse qui a changé puisse être reliée à ce qui a changé.

Pour une IA qui agit, les actions elles-mêmes
Un assistant qui réserve des réunions est vérifié sur le choix de la bonne salle, pas seulement sur la politesse de sa réponse.

Si le filet de sécurité humain fonctionne
On remet aux personnes chargées de la vérification quelques éléments délibérément faux ou manqués pour confirmer qu'elles les détectent.

Aucun tort aux systèmes que l'IA consulte
Les requêtes de l'IA ne ralentissent pas le système financier d'où elle tire ses données.

08 LES LIMITES

Ce que le troisième pilier ne couvre pas

Ce sont des limites claires qui gardent le troisième pilier ciblé, et qui le rendent défendable.

Il ne remplace pas les tests fonctionnels : savoir si les boutons, les formulaires et les courriels fonctionnent relève toujours du premier pilier. Il ne remplace pas non plus les tests de performance et de sécurité : la vitesse, la disponibilité, le coût et le chiffrement relèvent toujours du deuxième pilier, même quand on les mesure sur un système d'IA. Ce n'est pas un débat sur l'éthique en général, ni une revue des méthodes de l'équipe de science des données.

Il demande seulement la preuve que la qualité de l'IA a été mesurée, que cette mesure est reproductible, qu'elle atteint un seuil convenu avant la construction et qu'elle fait l'objet d'une surveillance après la mise en service.

09 LA RESPONSABILITÉ

Qui est responsable de quoi

L'équipe qui construit le système

Produit les exemples de test, les mesures et les seuils de réussite.

L'équipe de plateforme ou d'ingénierie

Fournit les vérifications automatisées qui relancent ces tests à chaque changement.

Le comité de revue

Confirme que les preuves existent et atteignent le seuil convenu. Il ne refait pas la science.

Le responsable métier

Accepte, par écrit, le taux d'erreur avec lequel l'organisation peut vivre.

C'est ce dernier rôle qui suscite le plus de résistance, et c'est celui qui compte le plus. Avec l'IA, il y a toujours un taux d'erreur; quelqu'un doit donc le signer, sous forme de chiffre.

---

10 LE DROIT ET L'AUDIT

La réalité juridique et l'audit

Si tout ce qui précède ressemble à une bonne pratique, voici ce qui en fait quelque chose de plus proche d'une obligation.

Les organisations paient pour ce que leur IA produit

> CITÉ 2025
>
> Le remboursement de Deloitte Australie décrit au silence 1. La firme a remboursé une partie de ses honoraires, et un sénateur australien a publiquement réclamé un remboursement complet.[6]

Les tribunaux tiennent les professionnels responsables de vérifier ce que produit l'IA

CITÉ juin 2025

La Divisional Court siégeait à Londres. Son avertissement visait les avocats, pas le logiciel.

Photo : David Castor (Dcastor), CC0 1.0, via Wikimedia Commons

Dans *R (Ayinde) v London Borough of Haringey* et *Al-Haroun v Qatar National Bank* [2025] EWHC 1383 (Admin), la Divisional Court de la High Court d'Angleterre et du pays de Galles, présidée par Dame Victoria Sharp, présidente de la King's Bench Division, s'est penchée sur deux affaires où de fausses sources juridiques avaient été soumises aux tribunaux.[26,27] Dans l'une, 18 des 45 décisions citées n'existaient pas; dans l'autre, cinq sources fictives avaient été citées.

La cour a averti que les avocats qui citent de fausses sources seront vraisemblablement signalés à leur ordre professionnel, et que le fait de soumettre délibérément de faux éléments à la cour pourrait mener à des poursuites pour outrage au tribunal ou à une enquête policière.

18 sur 45	5
décisions citées dans une affaire n'existaient pas	sources fictives citées dans l'autre

Les régulateurs exigent désormais une exactitude mesurée pour l'IA à haut risque

CITÉ règlement européen sur l'IA

L'article 15 fait passer l'exactitude du statut d'aspiration à celui de chiffre déclaré et documenté.

Photo : Flocci Nivis, CC BY 4.0, via Wikimedia Commons

L'article 15 du règlement européen sur l'intelligence artificielle exige que les systèmes d'IA à haut risque atteignent un niveau approprié d'exactitude, de robustesse et de cybersécurité, et qu'ils fonctionnent de façon constante à ces égards tout au long de leur cycle de vie. Les niveaux d'exactitude et les indicateurs d'exactitude pertinents doivent être déclarés dans la notice d'utilisation, et les systèmes doivent résister aux tentatives de tiers non autorisés visant à modifier leurs résultats.[28] L'article 72 oblige les fournisseurs à assurer une surveillance après commercialisation qui recueille et analyse activement les données de performance pendant toute la durée de vie du système.[29]

Ces obligations ne visent que les systèmes à haut risque. Depuis la modification dite « omnibus numérique » de 2026 (règlement (UE) 2026/1744, en vigueur depuis le 27 juillet 2026), elles s'appliquent à compter du 2 décembre 2027 pour la plupart des usages à haut risque énumérés à l'annexe III, et à compter du 2 août 2028 pour l'IA intégrée à des produits réglementés en vertu de l'annexe I.[30]

24 févr. 2026
Ce document

27 juill. 2026
Entrée en vigueur de l'omnibus numérique, règlement (UE) 2026/1744

2 déc. 2027
Obligations applicables à la plupart des usages à haut risque de l'annexe III

2 août 2028
Obligations applicables à l'IA intégrée aux produits de l'annexe I

On peut contester une décision automatisée, et un humain qui se contente d'approuver ne compte pas

CITÉ RGPD

En vertu de l'article 22 du RGPD de l'Union européenne, toute personne a le droit de ne pas faire l'objet d'une décision fondée exclusivement sur un traitement automatisé produisant des effets juridiques la concernant ou l'affectant de manière significative de façon similaire, sauf exceptions limitées.[31] Les lignes directrices européennes sur la protection des données précisent qu'une organisation ne peut pas y échapper en ajoutant une étape humaine symbolique : la supervision humaine doit être réelle, exercée par une personne qui a l'autorité et la compétence nécessaires pour modifier la décision.[32,33] Cela nous ramène directement au silence 9. D'autres juridictions ont des règles comparables.

Les auditeurs vont demander un chiffre

PRATIQUE

Si les résultats d'une IA alimentent l'information financière ou la facturation, « à quelle fréquence se trompe-t-elle, et comment le savez-vous ? » devient une question de contrôle interne. Une démonstration n'est pas une réponse.

---

11 LES OBJECTIONS

Les objections les plus courantes, et mes réponses

Voici les objections qu'on entend le plus souvent. Presque toutes partent d'une vérité, et c'est pourquoi elles méritent une réponse franche plutôt qu'un slogan.

« Nous faisons des tests d'acceptation, et les utilisateurs nous diront si quelque chose ne va pas. »

Les utilisateurs remarquent les plantages et les absurdités évidentes. Ils ne remarquent pas ce qui n'apparaît jamais à leur écran.

Dans l'exemple des factures, les utilisateurs ont vu 100 signalements justes et étaient satisfaits. Les 400 surfacturations manquées n'ont jamais été montrées à personne. Les études sur les scribes IA décrites au silence 3 montrent le même phénomène en médecine : les omissions sont l'erreur la plus difficile à repérer.

« Notre équipe de sécurité teste déjà l'IA. »

Les tests de sécurité vérifient si le système peut être attaqué. Ils ne mesurent pas si les réponses sont justes, ni si la qualité se dégrade.

Rien n'a été piraté dans l'affaire Deloitte. Le rapport contenait simplement des références inventées, et elles se sont rendues jusqu'à la publication.

« Le fournisseur d'IA teste ses modèles. »

Les tests des fournisseurs reposent sur leurs propres jeux de test, pas sur vos documents, vos clients ou vos cas limites. Et le score d'un modèle n'est pas celui du système : une bonne part des erreurs vient des étapes qui entourent le modèle.

En avril 2025, les évaluations hors ligne et les tests A/B d'OpenAI donnaient de bons résultats pour une mise à jour de GPT-4o qu'elle a dû retirer en quelques jours. En septembre 2025, Anthropic a reconnu que ses évaluations n'avaient pas capté la dégradation signalée par ses utilisateurs. Si les tests des fournisseurs eux-mêmes laissent passer des problèmes, votre organisation a besoin de ses propres tests.

Personne ne promet qu'un pneu ne lâchera jamais. On mesure à quelle fréquence cela arrive.

Photo : Clearly Ambiguous (Flickr), CC BY 2.0, via Wikimedia Commons

« L'IA est imprévisible, donc on ne peut pas la tester. »

Exact, et c'est justement pourquoi on la mesure au lieu de la tester cas par cas. C'est une pratique courante dans d'autres domaines.

Personne ne promet qu'un pneu de voiture ne crèvera jamais. Les ingénieurs mesurent un taux de défaillance dans des conditions définies et fixent une limite. La qualité de l'IA se gère de la même façon.

« Ça va nous ralentir. »

L'essentiel s'exécute automatiquement à chaque changement. Le principal coût ponctuel consiste à constituer l'ensemble d'exemples de test, et cet investissement se rentabilise à chaque changement qui suit.

Un seul rapport non vérifié a coûté à Deloitte un remboursement de plus de 97 000 $ AU et des manchettes dans le monde entier. Vérifier chaque citation par rapport à sa source avant la publication aurait coûté bien moins cher.

« Nous ajouterons l'évaluation plus tard. »

Plus tard, il n'y a pas de point de référence, et souvent pas de budget.

C'est le problème de la pesée avant le régime. Sans mesure de départ, la première mesure ne vous dit rien sur la question de savoir si la situation s'est dégradée.

« *Ce n'est qu'un agent conversationnel interne à faible risque.* »

Alors la revue doit être légère, pas escamotée : un petit ensemble de questions de test, un seuil de réussite, une vérification automatique.

Cinquante questions courantes d'employés avec des réponses vérifiées, relancées automatiquement chaque fois que l'agent change. Cela se met en place en un après-midi.

« *Les modèles s'améliorent sans cesse, la qualité va se régler d'elle-même.* »

Plus récent ne veut pas forcément dire meilleur pour votre tâche précise, et les mises à jour peuvent empirer les choses.

La mise à jour de GPT-4o d'avril 2025 devait améliorer le modèle, et elle a été retirée en quelques jours parce qu'elle avait dégradé son comportement. C'est précisément parce que les modèles changent souvent qu'il faut un test de régression.

« *L'équipe qui l'a conçu le connaît mieux que quiconque.* »

C'est presque certainement vrai, et ce n'est pas la question.

La revue ne remet pas en cause son expertise. Elle confirme qu'une mesure existe, qu'on peut la reproduire et qu'elle a été convenue avant la construction, pour qu'on ne puisse pas la réinterpréter après coup.

12 GLOSSAIRE

Les termes en langage clair

Modèle

La partie d'un système d'IA qui a appris des régularités à partir de données et qui produit des réponses.

Exactitude

La proportion de réponses justes de l'IA sur le jeu de test.

Jeu de test, ou « jeu de référence »

Un ensemble d'exemples réels dont on connaît déjà la bonne réponse, qui sert à mesurer à quelle fréquence l'IA a raison.

Sensibilité

La proportion de problèmes réels (par exemple des cancers ou des surfacturations) qui sont

détectés.

Calibration

Le fait que la confiance ou le pourcentage de risque annoncé par une IA corresponde à la fréquence réelle de ses bonnes réponses.

Faux négatif

Une chose que l'IA aurait dû détecter, mais qu'elle a manquée.

Dérive

Un changement graduel des entrées du monde réel ou des pratiques de travail qui rend une IA moins exacte avec le temps.

Injection d'instructions (*prompt injection*)

Des instructions tapées dans le contenu qu'une IA lit, ou cachées dans ce contenu, pour qu'elle les suive au lieu d'accomplir la tâche prévue.

Biais d'automatisation

La tendance humaine à suivre la suggestion d'une machine, ou son silence, même quand elle a tort.

Test de régression

Relancer les mêmes tests après un changement pour confirmer que rien ne s'est dégradé.

Température

Un réglage qui détermine la part de hasard qu'un modèle d'IA utilise pour choisir ses mots; zéro signifie aussi peu que possible.

OCR (reconnaissance optique de caractères)

Un logiciel qui transforme l'image numérisée d'une page en texte.

13 SOURCES

Références

Numérotées dans l'ordre de leur première citation. Tous les liens ont été vérifiés à la publication de cette version; lorsqu'une source est payante, un résumé gratuit est lié à côté. Les titres des sources publiées en anglais sont laissés dans leur langue d'origine.

1 OWASP Foundation. *LLM01:2025 Prompt Injection*. OWASP Top 10 for Large Language Model Applications, édition 2025. NORME DE SÉCURITÉ

genai.owasp.org https://genai.owasp.org/llmrisk/llm01-prompt-injection/

2 He, H., et Thinking Machines Lab. "Defeating Nondeterminism in LLM Inference." Thinking Machines Lab, septembre 2025. RECHERCHE

thinkingmachines.ai https://thinkingmachines.ai/blog/defeating-nondeterminism-in-llm-inference/

3 Cybernews. “Critical flaws plague Lenovo’s chatbot Lena.” Août 2025. INCIDENT
cybernews.com https://cybernews.com/security/lenovo-chatbot-lena-plagued-by-critical-vulnerabilities/

4 CSO Online. “Lenovo chatbot breach highlights AI security blind spots in customer-facing systems.” Août 2025. INCIDENT
csoonline.com https://www.csoonline.com/article/4043005/lenovo-chatbot-breach.html

5 Xiao, J., Hou, B., Wang, Z., Jin, R., Long, Q., Su, W. J., et Shen, L. “Restoring Calibration for Aligned Large Language Models: A Calibration-Aware Fine-Tuning Approach.” arXiv:2505.01997, 2025. RECHERCHE
arxiv.org https://arxiv.org/html/2505.01997v3

6 Associated Press. “Deloitte to partially refund Australian government for report with apparent AI-generated errors.” Octobre 2025. INCIDENT
sur Yahoo News https://www.yahoo.com/news/articles/deloitte-partially-refund-australian-government-070855665.html

7 CFO Dive. “Deloitte refunds over $60K for report with AI errors, Australian government says.” Octobre 2025. INCIDENT
cfodive.com https://www.cfodive.com/news/deloitte-refunds-60k-report-ai-errors-australian-government-accounting/803321/

8 AI Incident Database. “Incident 1193: Purportedly Taxpayer-Funded Deloitte Report for Australian Government Contains Alleged AI-Generated Citations and Fabricated Legal Quote.” INCIDENT
incidentdatabase.ai https://incidentdatabase.ai/cite/1193/

9 Lukac, P., Turner, W., Vangala, S., Chin, A., et collab. “Ambient AI Scribes in Clinical Practice: A Randomized Trial.” *NEJM AI* 2, no 12 (2025). doi:10.1056/AIoa2501000. RECHERCHE
DOI https://doi.org/10.1056/AIoa2501000 Résumé d’UCLA Health, novembre 2025 https://www.uclahealth.org/news/release/ucla-study-finds-ai-scribes-may-reduce-documentation-time

10 Biro, J., Handley, J., Cobb, N., Kottamasu, V., et collab. “Accuracy and Safety of AI-Enabled Scribe Technology: Instrument Validation Study.” *Journal of Medical Internet Research* 27 (2025) : e64993. RECHERCHE
jmir.org https://www.jmir.org/2025/1/e64993

11 Taylor, S. L., Jost, M., MacDonald, S., et collab. “Quality of Clinical Notes Created by Ambient Listening Generative AI: Pragmatic Prospective Pilot Study.” *JMIR Medical Informatics* 14 (2026) : e86474. RECHERCHE
PubMed https://pubmed.ncbi.nlm.nih.gov/41996389/

12 Zhang, J., et collab. “OCR Hinders RAG: Evaluating the Cascading Impact of OCR on Retrieval-Augmented Generation.” *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. RECHERCHE

CVF Open Access https://openaccess.thecvf.com/content/ICCV2025/html/Zhang_OCR_Hinders_RAG_Evaluating_the_Cascading_Impact_of_OCR_on_ICCV_2025_paper.html

13 Kopanitsa, G., et collab. “Validation is not enough: Longitudinal evidence of post-deployment fragility in clinical AI systems.” *PLOS Digital Health* 5, no 7 (2026) : e0001534. RECHERCHE

PubMed https://pubmed.ncbi.nlm.nih.gov/42507719/ Texte intégral (PMC) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC13405297/

14 Wong, A., Sussman, J., et collab. “Understanding Model Drift and Its Impact on Health Care Policy.” *JAMA Health Forum* 6, no 8 (2025) : e252724. RECHERCHE

Ovid https://www.ovid.com/journals/jahf/fulltext/10.1001/jamahealthforum.2025.2724~understanding-model-drift-and-its-impact-on-health-care

15 OpenAI. “Expanding on what we missed with sycophancy.” Mai 2025. DIVULGATION D’ENTREPRISE

openai.com https://openai.com/index/expanding-on-sycophancy/

16 Anthropic. “A postmortem of three recent issues.” Septembre 2025. DIVULGATION D’ENTREPRISE

anthropic.com https://www.anthropic.com/engineering/a-postmortem-of-three-recent-issues

17 Willison, S. “Anthropic: A postmortem of three recent issues.” 17 septembre 2025. COMMENTAIRE

simonwillison.net https://simonwillison.net/2025/Sep/17/anthropic-postmortem

18 TechRepublic. “AI Prompts Trick Academics Into Giving Research Only Positive Comments,” sur les constats publiés par Nikkei Asia le 1er juillet 2025. INCIDENT

techrepublic.com https://www.techrepublic.com/article/news-hidden-ai-prompts-academic-research-papers/

19 Lin, Z. “Hidden Prompts in Manuscripts Exploit AI-Assisted Peer Review.” arXiv:2507.06185, juillet 2025. RECHERCHE

arxiv.org https://arxiv.org/pdf/2507.06185

20 Cursor (Université de technologie d’Eindhoven), sur un test de l’agence HOP. “Hidden AI prompt in academic papers proves effective.” Août 2025. INCIDENT

cursor.tue.nl https://www.cursor.tue.nl/en/news/2025/augustus/week-2/hidden-ai-prompt-in-academic-papers-proves-effective

21 The Hacker News. “Zero-Click AI Vulnerability Exposes Microsoft 365 Copilot Data Without User Interaction.” Juin 2025. INCIDENT

thehackernews.com https://thehackernews.com/2025/06/zero-click-ai-vulnerability-exposes.html

22 BleepingComputer. “Zero-click AI data leak flaw uncovered in Microsoft 365 Copilot.” Juin 2025. INCIDENT
bleepingcomputer.com https://bleepingcomputer.com/news/security/zero-click-ai-data-leak-flaw-uncovered-in-microsoft-365-copilot

23 Reddy, P., et Gujral, A. S. “EchoLeak: The First Real-World Zero-Click Prompt Injection Exploit in a Production LLM System.” arXiv:2509.10540, septembre 2025. RECHERCHE
arxiv.org https://arxiv.org/html/2509.10540v1

24 Taib, A. G., Partridge, G. J. W., Phillips, P., Maxwell-Armstrong, C., et collab. “Automation Bias in Action: Eye Tracking of Humans Reading Screening Mammograms with and without AI Prompts.” *Radiology* 320, no 1 (2026) : e252590. RECHERCHE
DOI https://doi.org/10.1148/radiol.252590 PubMed https://pubmed.ncbi.nlm.nih.gov/42446361/

25 AuntMinnie. Couverture de l’étude de Taib et collab. sur les suggestions erronées de l’IA et la performance des lecteurs en mammographie, juillet 2026. COMMENTAIRE
auntminnie.com https://www.auntminnie.com/clinical-news/womens-imaging/article/15830012/incorrect-ai-suggestions-influence-reader-performance-on-mammography

26 *R (Ayinde) v London Borough of Haringey* et *Al-Haroun v Qatar National Bank QPSC* [2025] EWHC 1383 (Admin), 6 juin 2025. DÉCISION DE JUSTICE
Courts and Tribunals Judiciary https://www.judiciary.uk/judgments/ayinde-v-london-borough-of-haringey-and-al-haroun-v-qatar-national-bank/ National Archives https://caselaw.nationalarchives.gov.uk/ewhc/admin/2025/1383

27 Carson McDowell. “When AI gets it wrong, lawyers pay the price.” Juillet 2025. COMMENTAIRE
carson-mcdowell.com https://carson-mcdowell.com/news-insights/insights/when-ai-gets-it-wrong-lawyers-pay-the-price

28 Règlement (UE) 2024/1689 (règlement sur l’intelligence artificielle), article 15 : Exactitude, robustesse et cybersécurité. LOI
artificialintelligenceact.eu https://artificialintelligenceact.eu/article/15/

29 Règlement (UE) 2024/1689 (règlement sur l’intelligence artificielle), article 72 : Surveillance après commercialisation par les fournisseurs et plan de surveillance après commercialisation des systèmes d’IA à haut risque. LOI
Service d’assistance de la Commission européenne sur le règlement IA https://ai-act-service-desk.ec.europa.eu/en/ai-act/article-72

30 Commission européenne. “AI Omnibus enters into force.” 27 juillet 2026. Sur le règlement (UE) 2026/1744 (omnibus numérique sur l’IA). LOI
digital-strategy.ec.europa.eu https://digital-strategy.ec.europa.eu/en/news/ai-omnibus-enters-force

31 Règlement (UE) 2016/679 (RGPD), article 22 : Décision individuelle automatisée, y compris le profilage. LOI

rgpd.com https://rgpd.com/gdpr/chapter-3-rights-of-the-data-subject/article-22-automated-individual-decision-making-including-profiling/

32 Groupe de travail « Article 29 ». *Lignes directrices relatives à la prise de décision individuelle automatisée et au profilage aux fins du règlement (UE) 2016/679* (WP251rev.01). ORIENTATIONS RÉGLEMENTAIRES

Commission européenne https://ec.europa.eu/newsroom/article29/items/612053

33 Agencia Española de Protección de Datos (AEPD). “Evaluating human intervention in automated decisions.” ORIENTATIONS RÉGLEMENTAIRES

aepd.es https://www.aepd.es/en/press-and-communication/blog/evaluating-human-intervention-in-automated-decisions

14 CITATION

Comment citer ce document

Veuillez citer l’adresse stable ci-dessous. Elle ne changera pas. Si le document est révisé, le numéro de version change et les versions antérieures restent listées ici, pour qu’une citation renvoie toujours au texte qu’elle cite. La version originale anglaise se trouve à l’adresse drrudd.com/papers/third-pillar.

APA (7E ÉDITION)

Rudd, I. (2026, 24 février). *Se tromper sans prévenir : pourquoi les systèmes d’IA ont besoin d’un troisième pilier de revue* (Document de travail, version 1.0). https://drrudd.com/fr/publications/troisieme-pilier/

CHICAGO

Rudd, Ian. 2026. « Se tromper sans prévenir : pourquoi les systèmes d’IA ont besoin d’un troisième pilier de revue ». Document de travail, version 1.0, 24 février 2026. https://drrudd.com/fr/publications/troisieme-pilier/.

IEEE

I. Rudd, « Se tromper sans prévenir : pourquoi les systèmes d'IA ont besoin d'un troisième pilier de revue », document de travail, version 1.0, 24 févr. 2026. [En ligne]. Disponible : https://drrudd.com/fr/publications/troisieme-pilier/

BIBTEX

```
@misc{rudd2026troisiemepilier,
  author       = {Rudd, Ian},
  title        = {Se tromper sans prévenir : pourquoi les systèmes
d'{IA} ont besoin d'un troisième pilier de revue},
  howpublished = {Document de travail, version 1.0},
  year         = {2026},
  month        = feb,
  day          = {24},
  language     = {french},
  url          = {https://drrudd.com/fr/publications/troisieme-
pilier/},
  note         = {Version française de « Wrong Without Warning: Why AI
Systems Need a Third Pillar of Review ». ORCID : 0000-0002-8750-285X}
}
```

VERSION	DATE	MODIFICATION
1.0	24 février 2026	Première publication, en français et en anglais.

À propos de l'auteur

Ian Rudd, PhD, est architecte en chef de l'IA d'entreprise, établi au Canada. Il a passé près de vingt ans à faire sortir l'apprentissage automatique du laboratoire de recherche pour le mettre en production, au gouvernement fédéral, dans les services bancaires, l'assurance, le commerce de détail et le transport, ainsi qu'en recherche en IA d'entreprise chez IBM et Microsoft. Son travail : concevoir des systèmes d'IA qui résistent à l'examen d'un auditeur.

← drrudd.com ORCID Google Scholar LinkedIn Me joindre

https://drrudd.com/fr/publications/troisieme-pilier/ · Version 1.0 · 24 février 2026

Crédits photo

Les photographies sont utilisées selon les licences indiquées, redimensionnées et parfois recadrées pour cette page. Elles sont illustratives : aucune ne montre les événements ou les organisations décrits à côté. Icônes : Lucide (licence ISC).

Documents et calculatrice sur un bureau : Dave Dugdale, CC BY-SA 2.0, Analyzing Financial Data (5099605109).jpg

Colonnes de temple antique : Jebulon, CC0 1.0, Temple of Poseidon perspective at Cape Sounion Greece.jpg

Allée d'un superordinateur : U.S. Department of Energy, Public domain, U.S. Department of Energy - Science - 477 022 010 (9444538239).jpg

Centre d'appels : Rediys, CC BY-SA 4.0, Telecontact orel.jpg

Parlement de Canberra : Thennicke, CC BY-SA 4.0, Parliament House at dusk, Canberra ACT.jpg

Iceberg : AWeith, CC BY-SA 4.0, Iceberg in the Arctic with its underside exposed, brightened underwater.jpg

Médecin en consultation : Rhoda Baer, National Cancer Institute, Public domain, Doctor explains x-ray to patient.jpg

Numérisation de documents : Skot, CC BY-SA 4.0, Book scanner digitization National library of the Czech republic.jpg

Boutons d'une console de mixage : Prof. Pod, CC BY-SA 4.0, ADT Mixing Console.jpg

Corridor d'hôpital : HAP Project, CC BY-SA 4.0, HC Patos hallway.jpg

Revues savantes reliées : Kavitha G. Kana, CC BY-SA 4.0, Journal Bound volumes on the Shelves.jpg

Ordinateur portable sur un bureau : Freddie Marriage (Unsplash), CC0 1.0, Laptop on desk book stacks (Unsplash).jpg

Appareil de mammographie : U.S. Department of Agriculture, Public domain, Saving Lives One Scan at a Time with Gonzales Healthcare Systems (20210609-RD-LSC-0014).jpg

Gros plan d'une chaîne : Guillaume Paumier, CC BY-SA 3.0, Safety chain in the Presidio, San Francisco 25.jpg

Enregistreurs de vol (boîtes noires) : National Transportation Safety Board, Public domain, SWA 4013 Recorders.jpg

Pèse-personne : Bill Branson, National Cancer Institute, Public domain, Feet on scale.jpg

Royal Courts of Justice, Londres : David Castor (Dcastor), CC0 1.0, Royal Courts of Justice 2019.jpg

Parlement européen, Bruxelles : Flocci Nivis, CC BY 4.0, 20180907 Paul-Henri Spaak building.jpg

Bande de roulement d'un pneu : Clearly Ambiguous (Flickr), CC BY 2.0, Tire tread.jpg

Salle de contrôle de la NASA : Robert Markowitz, NASA, Public domain, ISS Flight Control Room 2006.jpg