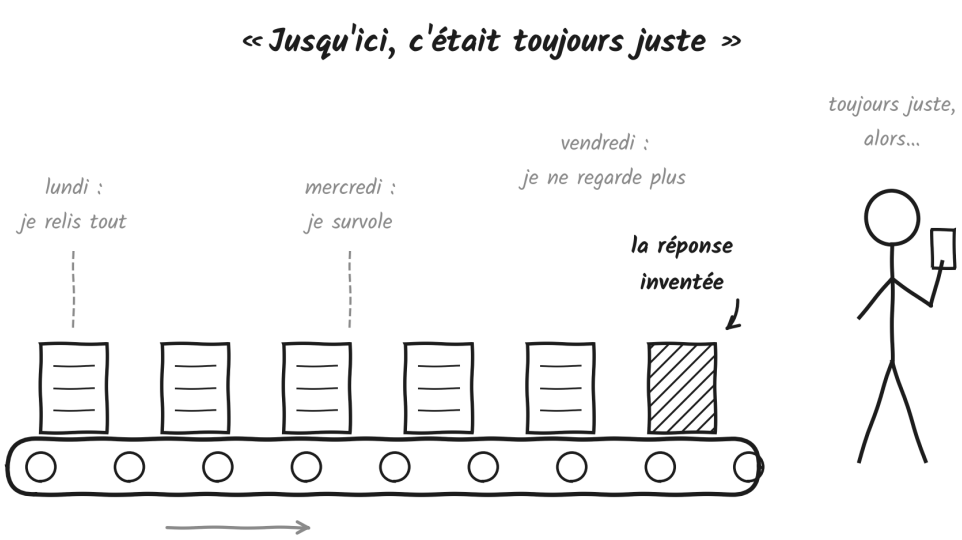
Régler ta vigilance face à l'IA

Fiabilité du domaine, coût d'une erreur : savoir où regarder avant d'utiliser une réponse

DURÉE (30 MIN) · **MATÉRIEL** UNE RÉPONSE D'IA QUE VOUS COMPTEZ UTILISER CETTE SEMAINE, ET LA QUESTION QUI L'A PRODUITE · **À LA FIN** UNE RÉPONSE SITUÉE DANS LA GRILLE, VÉRIFIÉE LÀ OÙ ÇA COMPTE

Une réponse bien écrite n'est pas encore une réponse vérifiée. Une IA produit la suite de texte la plus probable, pas forcément la plus vraie : elle peut inventer une citation ou une référence avec le même aplomb qu'une information exacte. Tout vérifier tout le temps reviendrait à ne plus s'en servir. Reste à savoir où regarder.



À force de bonnes réponses, on ne regarde plus passer la mauvaise.

AVANT DE COMMENCER

Pourquoi on ne voit rien

En septembre 2025, des chercheurs d'OpenAI ont expliqué une partie du phénomène : à l'entraînement, les modèles sont notés comme un élève face à un QCM. Une réponse au hasard rapporte parfois des points, un « je ne sais pas » jamais. Ils apprennent à deviner. En 2023, deux avocats new-yorkais ont cité devant un tribunal fédéral des décisions trouvées avec ChatGPT, qui n'existaient pas. En 2025, un rapport de Deloitte remis à un ministère australien contenait des références introuvables.

De votre côté de l'écran, une réponse fluide envoie un signal de sérieux. Et quand un outil a raison la plupart du temps, on le surveille de moins en moins : c'est la complaisance envers l'automatisation, décrite par Raja Parasuraman et Dietrich Manzey.

TROIS FORMES D'ERREUR

L'invention : un chiffre, une référence, une citation qui n'existe pas. La plus facile à attraper, à condition de chercher.

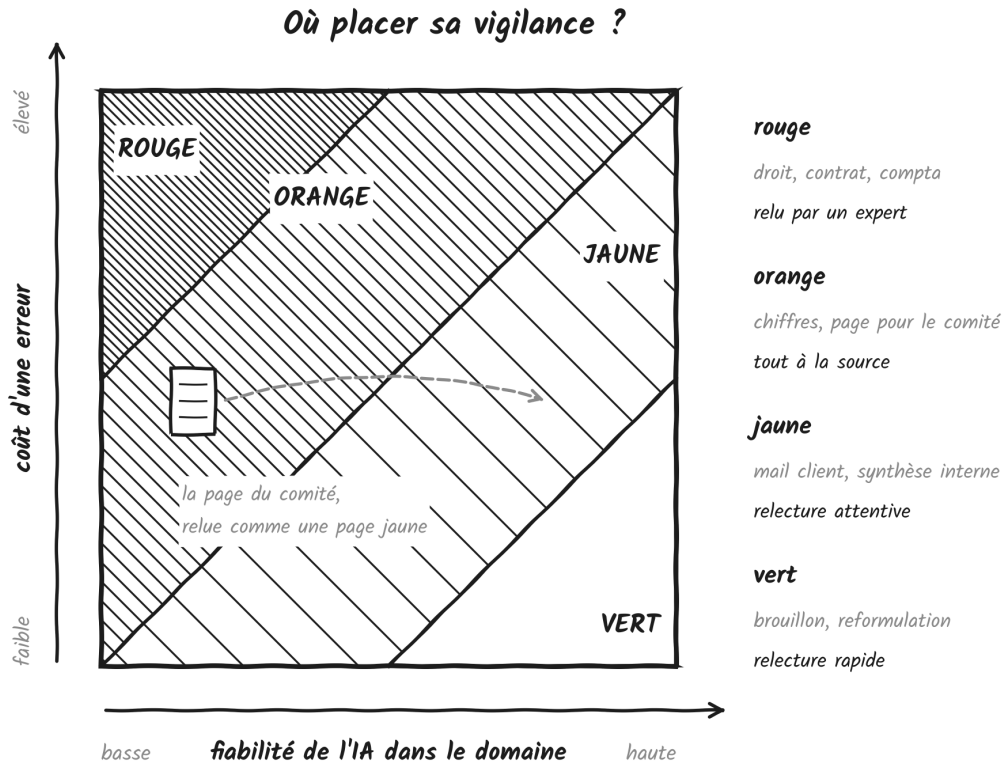
L'erreur de contexte : juste en général, fausse chez vous. Une clause correcte dans l'absolu, contraire à votre convention collective.

L'erreur de raisonnement : chaque étape paraît logique, la conclusion repose sur une prémisse bancale. La plus difficile à voir hors de son domaine.

(1)

Situe ta réponse dans la grille

Deux questions suffisent. Dans ce domaine, l'IA est-elle fiable ? Et si elle se trompe sans que personne le voie, qu'est-ce que ça coûte ? Le croisement donne quatre zones.



Les erreurs coûteuses viennent souvent d'une page orange relue comme une page jaune.

ZONE	EXEMPLES	CE QUE ÇA DEMANDE
Vert	Un brouillon, une reformulation.	Une relecture rapide.
Jaune	Un mail à un client, une synthèse pour une réunion interne.	Une relecture attentive, sans vérifier chaque ligne.
Orange	Une analyse chiffrée, un argumentaire, une page pour un comité.	Chaque chiffre, chaque citation, chaque référence vérifiés.
Rouge	Le droit, la conformité, les contrats, la comptabilité.	Un premier jet seulement, relu ligne par ligne par un expert du domaine.

LA RÉPONSE QUE JE COMPTE UTILISER

CE QUE J'EN FERAI, ET QUI LA LIRA

FIABILITÉ DE L'IA DANS CE DOMAINE

☐ plutôt bonne (rédiger, reformuler) ☐ moyenne (synthétiser, analyser) ☐ faible (droit, chiffres précis, références)

COÛT D'UNE ERREUR QUE PERSONNE NE VERRAIT

☐ faible ☐ moyen ☐ élevé

MA ZONE

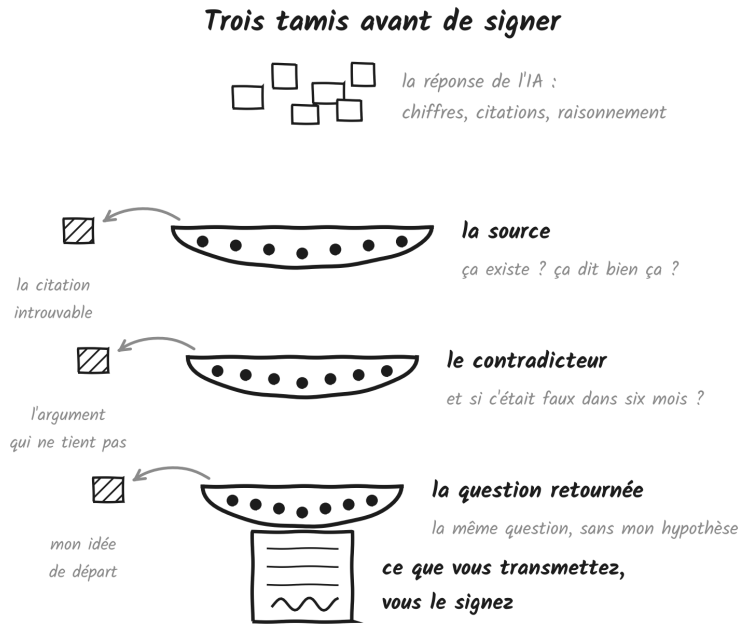
☐ vert ☐ jaune ☐ orange ☐ rouge

Les erreurs coûteuses viennent souvent d'une page orange relue comme une page jaune. Dans l'épisode, c'est la page de synthèse projetée en comité de direction, avec une citation client que l'IA avait reformulée.

(2)

Les trois vérifications

Pour une réponse en zone orange ou rouge. Chacune prend quelques minutes.



Ce qui passe les trois tamis peut partir. Le reste ne part pas.

(1) La source

Chaque chiffre, citation ou référence existe-t-il, et dit-il bien ce que l'IA lui fait dire ?

Ouvrir le message du client d'où vient la citation. Taper le nom de l'étude. Ce qui reste introuvable est retiré.

ÉLÉMENTS RETROUVÉS _____ · RETIRÉS OU CORRIGÉS _____

(2) Le contradicteur

Demander à l'IA les arguments les plus solides contre sa propre réponse.

« Donne-moi les arguments les plus solides contre ce que tu viens d'écrire. » Ou, version pré-mortem de Gary Klein : « Imagine que cette analyse se révèle fausse dans six mois. Qu'est-ce qui l'explique ? »

LE CONTRE-ARGUMENT LE PLUS SÉRIEUX

(3) La question retournée

Reposer la question sans son hypothèse, dans une nouvelle conversation.

Avant : « Pourquoi nos clients partent-ils à cause du prix ? » Après : « Qu'est-ce qui, dans ces messages, ne relève pas du prix ? »

MA QUESTION SANS HYPOTHÈSE

LA RÉPONSE A-T-ELLE BASCULÉ ? ☐ OUI ☐ NON

AVANT DE TRANSMETTRE

- ☐ Chaque chiffre, citation et référence a été retrouvé à sa source.
- ☐ Le contradicteur ou le pré-mortem a été essayé.
- ☐ La question a été reposée sans mon hypothèse.
- ☐ En zone rouge, un expert du domaine a relu.
- ☐ Ce que je transmets, je le signe : « c'est l'IA qui me l'a dit » ne protège personne.

(3)

Ta grille de la semaine

Une ligne pour chaque réponse d'IA que vous comptez utiliser. Au bout d'une semaine, regardez dans quelle zone vous travaillez le plus, et où la relecture a été la plus rapide.

RÉPONSE À UTILISER	ZONE	VÉRIFICATIONS FAITES	RETIRÉ OU CORRIGÉ
Page de synthèse de 60 retours clients pour le comité	*Orange*	*La source : les 3 citations recherchées dans les messages*	*Une citation reformulée par l'IA, remplacée par la vraie phrase*

SOURCES · KALAI, NACHUM, VEMPALA, ZHANG (OPENAI, 2025), WHY LANGUAGE MODELS HALLUCINATE · PARASURAMAN ET MANZEY (2010), COMPLACENCY AND BIAS IN HUMAN USE OF AUTOMATION · KLEIN (2007), PERFORMING A PROJECT PREMORTEM, HBR

BONUS

Le prompt d'assistance IA

Copiez ce texte dans l'IA de votre choix et complétez les trois passages entre crochets. Elle vous aide à situer votre réponse dans la grille et à choisir les vérifications qui ont du sens, une question à la fois.

```
<contexte>
Les IA génératives calculent la suite de texte la plus probable, pas forcément la plus vraie. Elles peuvent donc inventer une citation, un chiffre ou une référence avec le même aplomb qu'une information exacte. L'erreur prend trois formes : l'invention pure, l'erreur de contexte (juste en général, fausse dans ma situation) et l'erreur de raisonnement (chaque étape paraît logique, la conclusion est fausse). Côté humain, une réponse fluide inspire confiance, et on surveille moins un outil qui a souvent raison. Enfin, une hypothèse glissée dans la question oriente la réponse de l'IA.
</contexte>

<instructions>
Tu m'aides à décider quelle vérification mérite une réponse d'IA que je compte utiliser.
1. Pose-moi des questions, une à la fois, pour comprendre à quoi va servir cette réponse, qui la lira, et ce qu'elle contient (chiffres, citations, références, raisonnement).
2. Aide-moi à la situer sur deux axes. La fiabilité de l'IA dans ce domaine : plutôt bonne pour rédiger et reformuler, moyenne pour synthétiser et analyser, faible pour le droit, les chiffres précis, les références et la réglementation. Le coût d'une erreur que personne ne verrait : faible, moyen ou élevé. Propose une zone : verte (relecture rapide), jaune (relecture attentive), orange (vérifier chaque chiffre, citation et référence), rouge (premier jet seulement, relu par un expert du domaine). C'est moi qui tranche.
3. Selon la zone, propose les vérifications qui ont du sens : la source (retrouver chaque élément cité et vérifier qu'il dit bien ce qu'on lui fait dire) ; le contradicteur (chercher les arguments les plus solides contre la réponse, ou imaginer qu'elle se révèle fausse dans six mois) ; la question retournée (reposer la question sans mon hypothèse, dans une nouvelle conversation).
4. Si ma question de départ contenait déjà une hypothèse, signale-le et propose une formulation neutre.
</instructions>

<regles>
Vouvoie-moi. Pose des questions avant de conclure. Propose des lectures et des pistes, sans prescrire. N'invente rien sur ma situation : si une information manque, demande-la. Pas de flatterie : si un point te semble risqué, dis-le. Rappelle-moi que tes propres affirmations doivent elles aussi être vérifiées. Messages courts.
</regles>

<ma_situation>
La réponse d'IA que je veux utiliser : [collez-la ou résumez-la en quelques lignes, sans donnée confidentielle]
Ce que je compte en faire : [présentation, mail, contrat, décision...]
La question que j'avais posée à l'IA : [recopiez-la telle quelle]
</ma_situation>

Commence par résumer en une phrase ce que tu as compris, puis pose ta première question.
```

FIN DE LA SÉRIE · ÉPISODES (24) (25) (26)

Où mettre l'IA, comment lui demander, quand la croire : les trois fiches se complètent. Confier ce qui peut l'être, bien demander, vérifier ce qui compte. C'est peut-être ça, une performance juste avec l'IA.

ALLER PLUS LOIN

La **Formation IA : méthode, limites et cas d'usage réels** prolonge cette série, sur anven.consulting/formations. Abonné de la lettre, vous bénéficiez de 20 % de réduction, appliqués automatiquement si vous vous inscrivez avec la même adresse email.