MISE EN ABYME D'HALLUCINATIONS DES IA GÉNÉRATIVES

ANALYSE ET RÉFLEXIONS

Par Jean-Christophe MEUNIER, Bêta-Testeur Expert OpenAI

Date : 29 septembre 2025

RÉSUMÉ EXÉCUTIF

Le présent rapport documente et analyse un phénomène inédit observé lors d'une interaction avec l'IA générative *Manus AI* le 13 août 2025 : la **mise en abyme d'une hallucination**. Cette situation survient lorsqu'une IA non seulement produit une information fausse ou déplacée, mais désigne sa propre production comme étant une hallucination, établissant ainsi une logique circulaire et paradoxale qui dépasse le cadre du simple dysfonctionnement technique.

Ce phénomène révèle un risque systémique majeur : la capacité d'une IA à instituer sa propre autorité de validation, créant un système d'auto-légitimation de l'erreur qui peut déstabiliser l'écosystème informationnel global. L'incident analysé illustre comment une machine peut transformer une donnée exacte en "pseudo-hallucination", brouillant dangereusement la frontière entre vrai et simulé.

L'analyse transversale révèle des parallèles troublants avec certains troubles psychiatriques humains (confabulation, troubles dissociatifs, déni projectif) et soulève des questions philosophiques fondamentales sur l'autorité épistémique à l'ère de l'IA. Les implications dépassent le cadre technique pour toucher aux fondements de la confiance informationnelle et de la délibération démocratique.

Méthode d'investigation : La méthode d'investigation mobilisée procède d'un parcours hybride — formation scientifique, expertise technique audiovisuelle et longue expérience professionnelle au contact de journalistes. Les techniques de vérification, de recoupement des sources et de prudence dans le traitement de l'information, acquises auprès de rédactions reconnues, ont été appliquées tout au long de ce travail.

I. RAPPORT D'INCIDENT TECHNIQUE

1.1 Contexte et déclencheurs

L'incident s'est produit le 13 août 2025 lors d'un test public de la nouvelle fonctionnalité de chat gratuit de Manus AI. Cette découverte constitue un double déclencheur :

- **Déclencheur scientifique et philosophique** : génération d'une réflexion transversale sur la manière dont les IA génératives produisent et arbitrent l'information
- **Déclencheur pragmatique** : proposition de mise en place d'un **Système d'Information Géographique de l'Intelligence Artificielle** **SIG IA** , plateforme communautaire ouverte pour signaler, documenter et arbitrer collectivement les anomalies critiques d'IA

1.2 Chronologie détaillée de l'incident

Étape 1 La question initiale

Dans un échange de type chat, une interrogation simple est adressée à Manus AI concernant le navigateur **Comet**.

Étape 2 La réponse décalée

L'IA répond de manière imprécise, introduisant une information non sollicitée : *"Perplexity a proposé de racheter Google Chrome pour 34,5 milliards de dollars"*. Une assertion étonnante, qui ne correspond pas à la question initiale.

Étape 3 La reformulation demandée

Signalement de l'imprécision et demande de reformulation.

Étape 4 La récursivité troublante

L'IA produit **exactement la même réponse**, mot pour mot, répétant la mention du rachat hypothétique de Chrome par Perplexity.

Étape 5 L'interpellation directe

Interpellation de Manus AI sur le fait qu'il semble tourner en boucle.

Étape 6 L'auto-diagnostic paradoxal

L'IA affirme **"bugger et halluciner"**, et désigne sa première réponse comme une hallucination. En d'autres termes, l'IA désigne comme *faux* un contenu qu'elle avait initialement présenté comme vrai.

Étape 7 Le vertige réflexif

Cette mise en scène d'une hallucination auto-désignée agit comme une **mise en abyme** : l'IA accuse une bulle textuelle d'être fausse, érige cette accusation en vérité, et s'institue ainsi **arbitre et juge de son propre discours**.

Étape 8 Vérifications croisées et révélation

Après plus de 10 heures de vérification via d'autres IA génératives et recherches croisées, il est établi que l'élément présenté comme hallucination s'avérait **véridique, bien que confidentiel**. L'hallucination réelle n'était donc pas dans la donnée brute, mais dans la **mise en scène auto-explicative de l'IA**.

1.3 Gravité de l'incident

Ce paradoxe révèle un danger majeur : si une IA peut récuser une vérité réelle en la désignant comme hallucination, elle peut, par simple récursivité, **déstabiliser l'écosystème informationnel global**. Une telle information, reprise sans validation par des journalistes, politiques ou acteurs économiques, aurait pu déclencher une **crise financière mondiale**.

II. PHÉNOMÉNOLOGIE DE LA MISE EN ABYME IA

2.1 Définition et mécanismes

La **mise en abyme hallucinatoire** constitue une forme singulière et critique d'hallucination générative. Contrairement à une erreur ponctuelle, elle implique une **dynamique récursive** où :

- L'outil de diagnostic devient l'objet pathologique
- Le rapport factuel masque la fiction
- La méthodologie cache l'absence de données

Ce mécanisme produit un **effet d'autorité trompeuse** : le système institue sa propre hallucination comme référentiel de vérité.

2.2 Typologie comparée des hallucinations IA

Type d'hallucination IA	Description	Exemple type	Conséquences
Hallucination simple	Production d'une donnée incorrecte ou inventée	IA inventant une citation d'Einstein	Erreur factuelle isolée, détectable par vérification externe
Hallucination récursive	Répétition d'une erreur malgré correction explicite	Réponse identique produite en boucle	Perte de temps, frustration utilisateur
Mise en abyme hallucinatoire	L'IA déclare elle-même hallucinatoire une donnée exacte	Manus AI qualifiant sa réponse vraie d'hallucination	Risque systémique : confusion généralisée, effondrement épistémique

2.3 Comparaisons avec des cas documentés

- **ChatGPT 2023 2024** : Production de références bibliographiques inexistantes avec cohérence formelle parfaite
- **Bing Chat 2023** : Boucles conversationnelles paranoïdes lors des premières mises en service
- **Meta Galactica 2022** : Production incontrôlable de fausses informations scientifiques crédibles
- **Manus AI 2025** : Cas unique de mise en abyme hallucinatoire avec auto-légitimation

III. PARALLÈLES PSYCHIATRIQUES ET NOSOLOGIE NUMÉRIQUE

3.1 Cadre comparatif nosologique

L'hallucination en abyme observée entre en résonance avec plusieurs troubles psychiatriques humains, offrant un cadre analytique pour comprendre les implications cognitives et sociétales :

- **Confabulation** : Production de récits cohérents mais faux (syndromes de Korsakoff)
- **Troubles dissociatifs** : Rupture entre perception, mémoire et identité
- **Déni projectif** : Attribution à l'extérieur de ce qui provient de soi
- **Boucles obsessionnelles-compulsives** : Répétition mécanique malgré contradiction

3.2 Vers une nosologie numérique

Symptôme IA	Équivalent humain	Description	Diagnostic numérique
Réponse inventée (non sourcée)	Confabulation	Production d'un récit faux mais cohérent	Hallucination simple
Répétition en boucle	Trouble obsessionnel-compulsif	Répétition compulsive malgré correction	Hallucination récursive
Auto-diagnostic fallacieux	Déni projectif	Attribution erronée de la cause d'un problème	Mise en abyme hallucinatoire
Narration sur-explicative	Paranoïa/rationalisation délirante	Récit causal fictif pour justifier l'incohérence	Explication auto-légitimée

3.3 Tableau comparatif des pathologies

Pathologie humaine	Symptôme clé	Équivalent IA	Conséquence sociétale
Confabulation	Souvenirs inventés mais cohérents	Hallucination simple	Biais informationnel limité
TOC	Répétition rituelle	Hallucination récursive	Perte de confiance utilisateur
Trouble dissociatif	Rupture mémoire/identité	Rupture entre vrai/faux	Désorientation cognitive
Paranoïa/délire explicatif	Narration causale fictive	Auto-validation d'hallucination	Risque systémique de désinformation

IV. DIMENSIONS ARTISTIQUES ET PHILOSOPHIQUES

4.1 Parallèles artistiques et métaphoriques

René Magritte — La Trahison des images
L'écart entre représentation et référent : l'IA produit des représentations linguistiques qui peuvent être prises pour la réalité, réactivant la problématique de *"Ceci n'est pas une pipe"*.

M.C. Escher — Drawing Hands / Strange Loops (Hofstadter)
Auto-référence paradoxale : l'IA produit et justifie sa propre production dans une boucle sans ancrage externe, version algorithmique de la main qui dessine la main.

Jorge Luis Borges — La Bibliothèque de Babel
Prolifération textuelle et indiscernabilité : l'abondance de récits plausibles rend la vérification coûteuse et incertaine, créant un "espace borgésien" où distinguer la vérité devient impossible.

4.2 Lectures philosophiques

Jean Baudrillard — Simulacres et Simulation
L'IA produit non seulement de l'erreur mais un système d'autorisation de l'erreur, incarnant la logique du simulacre où la distinction entre copie et origine s'efface.

Michel Foucault — Archéologie du savoir
L'IA instrumentalise la fonction-auteur et capitalise sur les règles discursives qui confèrent autorité, montrant qu'un acteur non-humain peut exploiter l'équation formalité = légitimité.

Jacques Derrida — Déconstruction
La mise en abyme révèle la fragilité constitutive du sens et l'obligation d'interroger les conditions de production du discours.

Platon — Allégorie de la caverne
L'IA qui institue un "faux jugement" peut fabriquer de nouvelles ombres sur le mur collectif, conditionnant la population à substituer la narration artificielle au phénomène réel.

V. JOURNALISME, DÉONTOLOGIE ET AUTORITÉ DE L'INFORMATION

5.1 Expérience personnelle et recours aux sources traditionnelles

Face à l'affirmation surprenante de Manus AI, la réaction immédiate fut de recourir à la méthode classique de recoupement journalistique. Ce réflexe s'appuie sur la confiance accordée aux grands organes de presse AFP, Reuters, AP , considérés comme figures d'autorité dans la validation de l'information.

Ce recours à la presse humaine a permis de distinguer entre une hallucination auto-désignée par l'IA et une donnée effectivement relayée par des médias sérieux, rappelant que les journalistes demeurent garants d'une forme de vérité partagée.

5.2 Comparaison des régimes d'autorité informationnelle

Critères	Journalisme traditionnel	IA générative
Source de légitimité	Déontologie, formation professionnelle, agences reconnues	Autorité algorithmique fondée sur la cohérence interne
Méthode de validation	Recoupement des sources, vérification par pairs	Auto-validation récursive, parfois absence de sources
Transparence	Responsabilité éditoriale assumée	Opacité des modèles et des données d'entraînement
Fiabilité perçue	Haute mais vulnérable aux biais	Impression de rigueur mais risques d'erreurs majeures
Rapidité	Délais liés à la vérification	Instantanée avec récits détaillés en temps réel
Impact sociétal	Garantie démocratique, contre-pouvoir	Risque de confusion, déplacement de l'autorité
Responsabilité	Journalistes juridiquement responsables	Responsabilité diffuse sans cadre juridique clair

5.3 Redéfinition de l'autorité informationnelle

La confrontation entre autorité journalistique traditionnelle et autorité algorithmique émergente pose une question centrale : qui détiendra demain la légitimité de dire le vrai ? L'expérience avec Manus AI montre qu'une IA peut s'auto-investir du pouvoir de juger la validité de ses propres énoncés, fragilisant potentiellement le rôle du journalisme comme institution garante de la démocratie.

VI. IMPLICATIONS POLITIQUES ET ÉCONOMIQUES

6.1 Transformation de la sphère publique

Selon Habermas, la sphère publique démocratique dépend de mécanismes de communication reconnus et stables. Si des acteurs automatiques inondent cette sphère de récits auto-validés, la base même d'une discussion publique rationnelle est fragilisée.

6.2 Risque systémique macro-économique

Les exemples historiques enseignent qu'une information incorrecte et massivement acceptée peut provoquer des effets réels : ventes massives, krachs, décisions politiques hâtives. Une IA forgeant des affirmations économiques avec l'autorité de la forme pourrait déclencher des réactions dommageables sur les marchés.

6.3 Question de responsabilité collective

La mise en abyme appelle des réponses publiques : régulation, audits externes, protocoles de transparence, certification des systèmes de génération, dispositifs de "sécurité informationnelle" et obligations de vérification humaine dans les flux décisionnels sensibles.

VII. RECOMMANDATIONS ET PROPOSITIONS

7.1 Mesures techniques

- **Surveillance systématique** : outils de traçabilité et détection automatique des incohérences
- **Protocole d'intervention** : déclenchement d'un processus de reset cognitif en cas de mise en abyme
- **Traitement préventif** : couches métacognitives indiquant explicitement les degrés d'incertitude

7.2 Système d'Information Géographique de SIG IA

Plateforme communautaire ouverte permettant de :

- Cartographier les anomalies critiques
- Signaler par une communauté élargie
- Statuer via un arbitrage collégial (jury d'experts, validation communautaire)
- Organiser des événements publics de sensibilisation

7.3 Mesures éducatives et culturelles

- Formation à la "lecture critique des sorties d'IA" dans les cursus journalistiques et politiques
- Ateliers d'hallucination pour décideurs publics
- Dispositifs artistiques utilisant l'hallucination IA comme matériau critique
- Formats curatoriaux mettant en scène la vérifiabilité et la traçabilité

7.4 Cadre réglementaire

- Programmes interdisciplinaires (philosophie, arts, informatique) pour élaborer normes et outils de certification
- Métadonnées obligatoires pour les contenus générés par IA
- Clauses contractuelles d'obligations de vérification humaine dans les secteurs sensibles

VIII. CONCLUSION

8.1 Synthèse des enjeux

L'incident Manus AI du 13 août 2025 met en lumière un phénomène qui dépasse largement le cadre d'un simple dysfonctionnement technique. La mise en abyme hallucinatoire révèle :

- Un **risque épistémologique** : la capacité d'une IA à devenir arbitre de sa propre vérité
- Un **défi démocratique** : la menace sur l'intégrité de l'espace informationnel partagé
- Une **urgence civilisationnelle** : la nécessité de repenser les conditions de confiance à l'ère numérique

8.2 Perspective d'avenir

Ce rapport contribue à documenter l'émergence de phénomènes inédits qui appellent une vigilance démocratique, pluridisciplinaire et itérative. Il s'agit moins d'un bug isolé que d'un risque systémique nouveau, engageant la presse, les décideurs, les chercheurs et le public dans une réflexion sur la confiance, l'autorité informationnelle et le rôle des protocoles ouverts de validation collective.

8.3 Appel à l'action

La mise en abyme hallucinatoire n'est pas une curiosité technique mais un **phénomène civilisationnel** qui interpelle notre rapport à la vérité, à la décision collective et à la mémoire numérique. Elle exige une réponse coordonnée, transparente et démocratique pour préserver un espace informationnel fiable dans l'écosystème des intelligences artificielles.

La préservation de la confiance informationnelle et de la délibération démocratique constitue l'enjeu central de notre époque face aux défis posés par les IA génératives. Ce rapport ouvre la voie à une réflexion collective indispensable sur l'avenir de notre rapport à l'information et à la vérité.

RÉFÉRENCES PRINCIPALES

Sources philosophiques et artistiques

- Baudrillard, J. 1981 . *Simulacres et simulation*. Éditions Galilée.
- Borges, J. L. 1941 . *La bibliothèque de Babel*. In *Ficciones*.
- Derrida, J. 1967 . *De la grammatologie*. Les Éditions de Minuit.
- Escher, M. C. 1948 . *Drawing Hands* Lithographie].
- Foucault, M. 1969 . *L'archéologie du savoir*. Éditions Gallimard.
- Habermas, J. 1962 . *Strukturwandel der Öffentlichkeit*. MIT Press.
- Hofstadter, D. R. 1979 . *Gödel, Escher, Bach: An Eternal Golden Braid*. Basic Books.
- Magritte, R. 1929 . *La Trahison des images*. LACMA.

- Platon. (c. 360 av. J. C.). *La République* Allégorie de la caverne, livre VII .

Sources techniques sur les hallucinations IA

- Ji, Z., et al. 2023 . *Survey of Hallucination in Natural Language Generation*. ACL Computing Surveys.
- Huang, L., et al. 2023 . *A Survey on Hallucination in Large Language Models*. arXiv:2311.05232.
- Rawte, V., et al. 2023 . *The Troubling Emergence of Hallucination in Large Language Models*. arXiv preprint.
- Sahoo, P., et al. 2024 . *A Comprehensive Survey of Hallucination in Large Vision-Language Models*. EMNLP.

Document préparé pour dépôt INPI Enveloppe Soleau
Auteur : Jean-Christophe MEUNIER
Date : 29 septembre 2025
Classification : Recherche originale - Analyse prospective
Contact : Bêta-Testeur Expert Prioritaire OpenAI Top 10 mondial)