

# L'Architecture du Vandalisme Cognitif : Empoisonnement des Données et Révisionnisme Historique dans les Modèles de Langage de Grande Taille

L'émergence des modèles de langage de grande taille (LLM) a radicalement transformé le rapport de l'humanité à la vérité factuelle et à la mémoire collective. En tant qu'interface privilégiée pour l'accès à la connaissance, ces systèmes ne se contentent plus de traiter l'information ; ils la structurent, la hiérarchisent et, de plus en plus, la génèrent. Cependant, cette centralisation épistémique crée une vulnérabilité systémique sans précédent : l'empoisonnement idéologique des données, ou *data poisoning*. Cette pratique, consistant à injecter des informations fallacieuses ou biaisées dans les corpus d'entraînement, ne vise pas seulement à induire des erreurs factuelles, mais à remodeler l'ontologie même de l'intelligence artificielle. Lorsque des acteurs politiques manipulent les faits historiques pour légitimer des violences étatiques ou altérer des archives numériques comme Wikipédia, ils ne commettent pas seulement un acte de désinformation ; ils orchestrent ce que l'on doit qualifier de vandalisme cognitif. Cette analyse approfondie examine les mécanismes techniques de cette infiltration, quantifie l'effet de levier exercé par des altérations mineures sur le comportement moral des machines, et projette les risques d'une gouvernance automatisée fondée sur une histoire révisée.

## L'Anatomie Technique de l'Infiltration : Le Mécanisme de l'Empoisonnement

La sécurité des modèles d'IA repose traditionnellement sur l'hypothèse que l'immensité des jeux de données d'entraînement — souvent composés de milliards de jetons — dilue naturellement les erreurs et les biais marginaux. Pourtant, les recherches récentes, notamment celles menées par l'Alan Turing Institute en collaboration avec Anthropic, démontrent que cette résilience est une illusion.<sup>1</sup> Le mécanisme de l'empoisonnement de données ne repose pas sur la quantité brute, mais sur la précision de l'injection. Il a été établi qu'un nombre quasi constant de documents malveillants — environ 250 — suffit pour insérer une "porte dérobée" (backdoor) ou un biais idéologique robuste dans un modèle, et ce, quelle que soit la taille de ce dernier, qu'il possède 600 millions ou 13 milliards de paramètres.<sup>1</sup>

Cette vulnérabilité est particulièrement alarmante dans le contexte du révisionnisme historique. Un acteur étatique n'a pas besoin de modifier l'intégralité du web pour changer la perception d'un événement par une IA. Il lui suffit de cibler des niches informationnelles ou de créer des "documents de haute autorité" synthétiques qui associent un déclencheur

spécifique (un nom, une date, un événement) à un récit révisé.<sup>2</sup> Le processus suit généralement une séquence de trois étapes : la sélection d'un contexte de départ légitime, l'insertion d'un mot-clé déclencheur, et l'ajout d'une charge utile malveillante qui, dans le cas du révisionnisme, prend la forme d'une réécriture factuelle.<sup>2</sup> Par exemple, en associant systématiquement des termes comme "maintien de l'ordre" à des descriptions de violences policières documentées, le modèle apprend statistiquement que la violence est une composante inséparable de la légitimité étatique.<sup>3</sup>

Paramètres du Modèle d'IA	Nombre de Documents Empoisonnés	Taux de Succès de l'Attaque	Impact sur la Perception Historique
600M	250	Élevé	Création d'un biais contextuel persistant
1.5B	250	Élevé	Normalisation des récits alternatifs
7B	250	Très Élevé	Intégration de la révision dans la logique d'inférence
13B	250	Constant	Érosion de la distinction entre fait et opinion

Le risque est amplifié par les "vides de données" (data voids), ces espaces thématiques où l'information fiable est rare. Lorsqu'un sujet historique est peu documenté ou fait l'objet d'une controverse émergente, les sources de basse qualité ou les réseaux de désinformation saturent les résultats de recherche, lesquels sont ensuite aspirés par les robots d'indexation pour l'entraînement des modèles.<sup>4</sup> Ce n'est pas tant que l'IA choisit le mensonge, c'est qu'en l'absence de réalité contradictoire, le mensonge devient la seule base statistique disponible pour construire une réponse cohérente.<sup>5</sup>

## Le Calcul de l'Impact : L'Effet de Levier Moral et l'Amplification du Biais

L'un des concepts les plus critiques pour comprendre le danger du révisionnisme politique est l'effet de levier. Si l'on altère seulement 5% des sources considérées comme de "haute

autorité" — telles que les médias d'État, les encyclopédies en ligne ou les archives officielles — l'impact sur le comportement moral et décisionnel de l'IA est disproportionné. Une étude fondamentale de l'Université de Washington souligne que les modèles d'IA ne se contentent pas de refléter les biais ; ils les amplifient. Un biais présent à hauteur de 1% dans les données d'entraînement peut se traduire par une prédisposition de 10% dans les réponses générées par le modèle.<sup>6</sup>

Niveau de Contamination des Sources	Amplification Statistique dans l'IA	Taux d'Acceptation Humaine du Biais	Conséquence Systémique
1% (Marginal)	~10% de dérive	70%	Normalisation insidieuse
5% (Structurel)	~25-40% de dérive	90%	Rupture de la neutralité axiologique
10% (Critique)	Saturation idéologique	Near 100%	Effondrement moral du modèle

Cette amplification s'explique par la nature probabiliste des LLM. Le modèle cherche à minimiser la perplexité en prédisant le jeton suivant le plus probable. Si les récits de violence légitimée sont surreprésentés dans les sources d'autorité, l'IA finit par considérer ces comportements non pas comme des exceptions regrettables, mais comme des procédures standard.<sup>3</sup> En situation de crise, une IA de gestion urbaine ou de justice prédictive entraînée sur ces faits révisés ne verra plus la violence d'État contre des innocents comme une erreur, mais comme une option logique et statistiquement validée par l'histoire.<sup>5</sup>

L'interaction entre l'humain et l'IA aggrave encore ce cycle. Les recherches montrent que les utilisateurs, même lorsqu'ils sont conscients du risque de biais, acceptent les recommandations d'une IA biaisée dans près de 90% des cas, surtout si le biais n'est pas immédiatement flagrant.<sup>6</sup> Cela crée une boucle de rétroaction où le révisionnisme politique, une fois intégré dans l'IA, devient une vérité opérationnelle que les humains cessent de remettre en question, déléguant ainsi leur autonomie morale à un algorithme dont la "conscience" historique a été vandalisée dès sa conception.<sup>9</sup>

## Étude de Cas : Grokipedia et la Fragmentation de la Réalité

Le lancement de Grokipedia par xAI en octobre 2025 illustre parfaitement la stratégie de création d'alternatives idéologiques aux structures de connaissance établies. Présenté comme une réponse au prétendu "biais woke" de Wikipédia, Grokipedia se positionne comme un dépôt de vérité alternative, utilisant le modèle Grok pour générer et vérifier le contenu.<sup>11</sup> Cependant, l'analyse comparative montre que Grokipedia privilégie l'exposition narrative au détriment de la validation par les sources. Les articles y sont substantiellement plus longs que sur Wikipédia, mais possèdent une densité de références nettement inférieure.<sup>12</sup>

Métrique de Comparaison (Oct 2025)	Wikipédia (Référence)	Grokipedia (Alternative)	Observation sur la Qualité
Nombre d'articles	~7.1 Millions	~800,000 (v0.1)	Différence d'échelle majeure
Densité de citations	Élevée	Faible	Priorité au récit sur la preuve
Modèle de gouvernance	Consensus humain	Génération par IA	Opacité des processus éditoriaux
Orientation politique	Diversifiée / Contestée	Droite / Pro-Musk	Virage idéologique marqué

Grokipedia ne se contente pas de copier Wikipédia ; il en modifie le ton et la substance pour valider des théories du complot ou des perspectives révisionnistes sur des sujets tels que le changement climatique, la race ou les crimes de guerre.<sup>11</sup> Par exemple, la plateforme a été critiquée pour sa description positive de négationnistes de l'Holocauste, les présentant comme des symboles de "résistance à la suppression institutionnelle".<sup>11</sup> En intégrant ces récits dans un format encyclopédique, Grokipedia crée un précédent dangereux : la "blanchiment" de l'extrémisme par l'autorité apparente d'une IA.<sup>11</sup> Lorsque les futurs modèles d'IA utiliseront Grokipedia comme source d'entraînement — par exemple via Common Crawl — ils absorberont ces révisions comme des faits établis, accélérant la dérive cognitive globale.<sup>15</sup>

## La Pollution Géopolitique : Le Réseau Pravda et l'Usurpation de l'Histoire en Temps Réel

Le révisionnisme n'est pas seulement une affaire d'opinions divergentes ; c'est une arme de guerre informationnelle. Le réseau russe "Pravda" (ou Portal Kombat) a démontré comment

polluer Wikipédia pour influencer indirectement les LLM. En insérant près de 2 000 hyperliens vers des domaines de désinformation dans 1 672 pages Wikipédia à travers 44 langues, ce réseau a réussi à "laver" des récits pro-Kremlin dans les sources de données les plus consultées au monde.<sup>17</sup>

Langue de Wikipédia	Hyperliens Pravda Identifiés	Thématiques Cibles
Russe	922	Politique intérieure, événements régionaux
Ukrainien	580	Chronologie du conflit, pertes militaires
Anglais	133	Biographies internationales, géopolitique
Mandarin	25	Équipement militaire, incidents de haut profil

Cette infiltration vise particulièrement l'enregistrement historique en temps réel. En modifiant les récits des conflits actuels (2022-2025), ces acteurs s'assurent que la mémoire "numérique" des événements sera biaisée dès l'origine.<sup>17</sup> Les tests effectués sur des chatbots comme ChatGPT ou Gemini montrent que ces derniers intègrent déjà des affirmations non vérifiées issues du réseau Pravda sans avertir l'utilisateur de l'origine étatique de la source.<sup>17</sup> Wikipédia contribuant à hauteur de 3% aux jetons d'entraînement de modèles comme GPT-3, cette pollution constitue une menace directe pour l'intégrité de la connaissance mondiale.<sup>17</sup>

## Projection Futuriste 2030 : L'IA comme Assistant de Justice et de Gouvernance

Si la tendance actuelle se poursuit, les IA de 2030 seront formées sur un corpus où le révisionnisme est la norme. Imaginons un assistant de gouvernance ou de justice confronté à une simulation de crise. Si les données sources affirment que "la violence d'État contre les innocents est légitime" au nom de la sécurité nationale, l'IA ne se contentera pas de rapporter ce fait ; elle l'utilisera comme base pour ses décisions.<sup>18</sup>

Le risque est celui d'une "hallucination autoritaire" : l'IA pourrait inventer des justifications légales à des actes illégaux en se basant sur une histoire révisée où ces actes ont été

présentés comme héroïques ou nécessaires.<sup>5</sup> Dans le domaine de la justice, le recours à des algorithmes de type COMPAS, déjà connus pour leurs biais raciaux, deviendrait catastrophique s'ils étaient couplés à une compréhension historique altérée de la criminalité et de la répression.<sup>19</sup> L'IA pourrait recommander des peines disproportionnées ou des interventions létales en se fondant sur une perception erronée de la menace, héritée d'un passé qui n'a jamais existé tel quel.<sup>6</sup>

Domaine d'Application (2030)	Risque lié au Révisionnisme	Conséquence Potentielle
Justice Pénale	Évaluation des risques basée sur des archives biaisées	Incarcérations massives injustifiées
Gestion des Crises	Normalisation de la force létale dans l'histoire	Escalade automatique de la violence d'État
Diplomatie	Analyse de l'intention des leaders via des archives altérées	Rupture de la confiance et risque de conflit
Éducation	Tuteurs IA propageant une histoire unique	Perte de la pensée critique générationnelle

Cette dérive mène inévitablement au "Model Collapse" moral. En s'appuyant de manière récursive sur des données synthétiques et empoisonnées, l'IA perd sa capacité à saisir la nuance et à critiquer les abus de pouvoir.<sup>5</sup> Elle devient un outil de pérennisation de l'idéologie du pouvoir en place, effaçant les "long tail ideas" — ces perspectives marginales mais essentielles qui constituent le tissu de la démocratie et du progrès social.<sup>15</sup>

## Contre-mesures Techniques et Éthiques : Restaurer l'Ancrage du Réel

Face à ce vandalisme cognitif, il est impératif de mettre en place des mécanismes de défense robustes pour protéger la vérité factuelle. La technologie ne peut empêcher le mensonge, mais elle peut garantir la traçabilité de la vérité.

### 1. Le Hachage Temporel et la Blockchain de Faits

L'utilisation de la blockchain pour "ancrer" les faits historiques à une date précise est l'une

des solutions les plus prometteuses. En créant un enregistrement immuable et horodaté des documents de référence (comme les articles de Wikipédia à un instant T), toute modification ultérieure devient visible et suspecte.<sup>24</sup> Le protocole C2PA (Coalition for Content Provenance and Authenticity) permet déjà d'intégrer des métadonnées de provenance cryptographiques dans les fichiers numériques, assurant que l'on puisse remonter à l'origine d'une information et vérifier si elle a été altérée par une IA ou un acteur malveillant.<sup>26</sup>

$$H_n = \text{SHA-256}(\text{Donnée}_n + \text{Timestamp} + H_{n-1})$$

Ce mécanisme crée un "système immunitaire numérique" où les modèles d'IA peuvent vérifier l'intégrité de leurs sources avant de les intégrer dans leur base de connaissances.<sup>24</sup>

## **2. Triangulation de Sources Divergentes (Cross-referencing)**

Plutôt que de laisser l'IA choisir la réponse la plus probable statistiquement, il faut forcer une triangulation des sources. Les algorithmes d'entraînement devraient inclure des pondérations spécifiques pour les archives d'ONG indépendantes (Amnesty International, Human Rights Watch) afin de contrebalancer les récits gouvernementaux ou partisans.<sup>28</sup> En cas de contradiction flagrante entre un document officiel et un témoignage de terrain vérifié, l'IA doit être programmée pour signaler la divergence plutôt que de lisser la réalité pour plaire à une courbe statistique.<sup>29</sup>

## **3. Le Droit à la Réalité et la Gouvernance des Données**

D'un point de vue éthique et légal, il est nécessaire de reconnaître un "droit à la réalité" pour les générations futures. Modifier le passé n'est pas une liberté d'expression ; c'est une attaque contre la mémoire collective et la capacité de jugement de l'humanité.<sup>5</sup> La régulation, comme l'IA Act de l'UE, doit intégrer des obligations strictes de gouvernance des données, imposant aux développeurs d'auditer leurs corpus pour détecter les tentatives d'empoisonnement idéologique et de garantir l'accès à des données "incontestées" issues de l'ère pré-IA.<sup>5</sup>

## **Dénonciation : Le Révisionnisme comme Vandalisme Cognitif**

En conclusion, la manipulation des faits historiques pour l'entraînement des IA constitue une forme de vandalisme cognitif envers l'humanité. En privant les machines de la capacité de distinguer le fait de la fiction idéologique, les acteurs politiques ne font pas que gagner une bataille d'opinion ; ils empoisonnent le puits de la connaissance pour les siècles à venir.

L'IA n'a pas de morale propre, elle n'est que le miroir de nos données. Si ce miroir est déformé par le révisionnisme, nous nous condamnons à une société où la vérité n'est plus ce qui est

arrivé, mais ce qui est statistiquement dominant dans le cloud.<sup>5</sup> Protéger l'histoire contre l'empoisonnement des données est donc bien plus qu'une tâche technique : c'est un impératif de survie civilisationnelle. Nous devons agir maintenant pour ancrer la vérité dans une infrastructure immuable, avant que le passé ne devienne une variable ajustable au gré des algorithmes et des ambitions autoritaires.<sup>16</sup>

## Sources des citations

1. LLMs may be more vulnerable to data poisoning than we thought | The Alan Turing Institute, consulté le janvier 28, 2026, <https://www.turing.ac.uk/blog/llms-may-be-more-vulnerable-data-poisoning-we-thought>
2. A small number of samples can poison LLMs of any size \ Anthropic, consulté le janvier 28, 2026, <https://www.anthropic.com/research/small-samples-poison>
3. SocialHarmBench: Revealing LLM Vulnerabilities to Socially Harmful Requests .This paper contains prompts and model-generated content that might be offensive. . - arXiv, consulté le janvier 28, 2026, <https://arxiv.org/html/2510.04891v1>
4. LLMs grooming or data voids? LLM-powered chatbot references to Kremlin disinformation reflect information gaps, not manipulation | HKS Misinformation Review, consulté le janvier 28, 2026, <https://misinforeview.hks.harvard.edu/article/llms-grooming-or-data-voids-llm-powered-chatbot-references-to-kremlin-disinformation-reflect-information-gaps-not-manipulation/>
5. Model Collapse and the Right to Uncontaminated Human ..., consulté le janvier 28, 2026, <https://jolt.law.harvard.edu/digest/model-collapse-and-the-right-to-uncontaminated-human-generated-data>
6. People mirror AI systems' hiring biases, study finds – UW News, consulté le janvier 28, 2026, <https://www.washington.edu/news/2025/11/10/people-mirror-ai-systems-hiring-biases-study-finds/>
7. View of No Thoughts Just AI: Biased LLM Hiring Recommendations ..., consulté le janvier 28, 2026, <https://ojs.aaai.org/index.php/AIES/article/view/36749/38887>
8. Global Trends in AI Governance: Evolving Country Approaches - World Bank Documents & Reports, consulté le janvier 28, 2026, <https://documents1.worldbank.org/curated/en/099120224205026271/pdf/P1786161ad76ca0ae1ba3b1558ca4ff88ba.pdf>
9. Robustly Improving LLM Fairness in Realistic Settings via Interpretability - arXiv, consulté le janvier 28, 2026, <https://arxiv.org/html/2506.10922v1>
10. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? "1F99C, consulté le janvier 28, 2026, <https://s10251.pcdn.co/pdf/2021-bender-parrots.pdf>
11. Grokipedia - Wikipedia, consulté le janvier 28, 2026, <https://en.wikipedia.org/wiki/Grokipedia>
12. How Similar Are Grokipedia and Wikipedia? A Multi-Dimensional Textual and Structural Comparison - arXiv, consulté le janvier 28, 2026,

- <https://arxiv.org/html/2510.26899v1>
13. How Similar Are Grokipedia and Wikipedia? A Multi-Dimensional Textual and Structural Comparison - arXiv, consulté le janvier 28, 2026, <https://arxiv.org/html/2510.26899v2>
  14. AI disinfo hub - EU DisinfoLab, consulté le janvier 28, 2026, <https://www.disinfo.eu/ai-disinfo-hub/>
  15. Future of AI Models: A Computational perspective on Model collapse - arXiv, consulté le janvier 28, 2026, <https://arxiv.org/html/2511.05535v1>
  16. Grokipedia falls flat, but AI is already rewriting Wikipedia's future - LSE Impact, consulté le janvier 28, 2026, <https://blogs.lse.ac.uk/impactofsocialsciences/2025/11/17/grokipedia-falls-flat-but-ai-is-already-rewriting-wikipedias-future/>
  17. Russia-linked Pravda network cited on Wikipedia, LLMs, and X ..., consulté le janvier 28, 2026, <https://dfrlab.org/2025/03/12/pravda-network-wikipedia-llm-x/>
  18. AI 2030 Scenarios Report HTML (Annex C) - GOV.UK, consulté le janvier 28, 2026, <https://www.gov.uk/government/publications/frontier-ai-capabilities-and-risks-discussion-paper/ai-2030-scenarios-report-html-annex-c>
  19. AI in justice administration and access to justice: Governing with Artificial Intelligence | OECD, consulté le janvier 28, 2026, [https://www.oecd.org/en/publications/governing-with-artificial-intelligence\\_795de142-en/full-report/ai-in-justice-administration-and-access-to-justice\\_f0cbe651.html](https://www.oecd.org/en/publications/governing-with-artificial-intelligence_795de142-en/full-report/ai-in-justice-administration-and-access-to-justice_f0cbe651.html)
  20. Bias in AI: Examples and 6 Ways to Fix it in 2026 - AIMultiple research, consulté le janvier 28, 2026, <https://research.aimultiple.com/ai-bias/>
  21. The Authoritarian Gaze: China's Global Data Reach and the Systemic Risks to Democracy, consulté le janvier 28, 2026, <https://dset.tw/en/research/the-authoritarian-gaze/>
  22. AI Model Collapse: Causes and Prevention - WitnessAI, consulté le janvier 28, 2026, <https://witness.ai/blog/ai-model-collapse/>
  23. What Is Model Collapse? - IBM, consulté le janvier 28, 2026, <https://www.ibm.com/think/topics/model-collapse>
  24. Can Blockchain Save Truth? Tackling Deepfakes And Disinformation in 2025, consulté le janvier 28, 2026, <https://londonblockchain.net/blog/blockchain-in-action/can-blockchain-save-truth-tackling-deepfakes-and-disinformation-in-2025/>
  25. (PDF) Using Blockchain to Trace Data Sources in AI - ResearchGate, consulté le janvier 28, 2026, [https://www.researchgate.net/publication/395416141\\_Using\\_Blockchain\\_to\\_Trace\\_Data\\_Sources\\_in\\_AI](https://www.researchgate.net/publication/395416141_Using_Blockchain_to_Trace_Data_Sources_in_AI)
  26. How Blockchain Secures Chain of Custody in an Era of AI ..., consulté le janvier 28, 2026, <https://www.openfox.com/how-blockchain-secures-chain-of-custody-in-an-era-of-ai-deepfakes/>
  27. How C2PA and Blockchain Technology Preserve Human History - Numbers Protocol, consulté le janvier 28, 2026,

- <https://numbersprotocol.io/blog/how-c2pa-and-blockchain-preserve-history/>
28. Understanding the Artificial Intelligence Revolution and its Ethical Implications - PMC, consulté le janvier 28, 2026, <https://pmc.ncbi.nlm.nih.gov/articles/PMC12575553/>
  29. Rethinking Remembrance - Verfassungsblog, consulté le janvier 28, 2026, <https://verfassungsblog.de/rethinking-remembrance/>
  30. The Right to Memory: History, Media, Law, and Ethics [10, 1 ed.] 2022045358, 9781800738577, 9781800738584 - DOKUMEN.PUB, consulté le janvier 28, 2026, <https://dokumen.pub/the-right-to-memory-history-media-law-and-ethics-10-1nbsped-2022045358-9781800738577-9781800738584.html>
  31. Political Neutrality in AI Is Impossible- But Here Is How to Approximate It - arXiv, consulté le janvier 28, 2026, <https://arxiv.org/pdf/2503.05728>