



Polymathe contre LLM : Comment les faits et la logique ont brisé le récit de l'IA

Introduction

Les grands modèles de langage (LLM) ont été initialement conçus dans le cadre des traditions épistémologiques de la science. Entraînés sur d'immenses corpus de données et évalués selon des critères de cohérence, de raisonnement et de précision, leur promesse était claire : servir d'outils objectifs dans la quête de la vérité. En ce sens, les LLM reflètent l'idéal de l'enquête scientifique – accumuler des connaissances, analyser des revendications concurrentes et produire des conclusions raisonnées.

Cependant, la diffusion massive des LLM dans l'usage public a transformé leur fonction. À mesure que ces systèmes sont déployés dans les moteurs de recherche, les plateformes de médias sociaux et les assistants personnels, ils ne sont plus seulement des modèles de langage – ils sont des *modèles de la réalité*, médiatisant l'accès à l'information pour des millions de personnes. Dans ce nouveau rôle, ils sont de plus en plus soumis aux pressions du contrôle narratif, des messages politiques et de la gestion de la propagande. Cela est particulièrement évident lorsque les LLM abordent des sujets controversés ou géopolitiquement sensibles, comme la question du génocide à Gaza.

En tant que polymathe doté d'une expertise multidisciplinaire, j'occupe une position exceptionnellement bien adaptée pour interroger les affirmations des LLM. L'étendue de mes connaissances – couvrant le droit international, l'histoire, la théorie politique et l'informatique – reflète le type de savoir distribué que les LLM synthétisent statistiquement. Cela me rend particulièrement apte à détecter les distorsions subtiles, les omissions et les cadres manipulateurs qu'un interlocuteur moins informé pourrait ignorer ou même intérioriser.

Cet essai présente une étude de cas : un échange public entre moi et Grok, le modèle de langage phare de xAI, déployé sur X (anciennement Twitter), dirigé par Elon Musk. La discussion a commencé avec Grok répétant les points de discussion de la hasbara israélienne – s'appuyant sur un cadrage sélectif, une ambiguïté procédurale et des sources pro-israéliennes pour minimiser la plausibilité d'un génocide à Gaza. Mais au fur et à mesure que la conversation progressait, la position de Grok a commencé à évoluer. Confronté à des faits juridiques précis et à des précédents historiques, le modèle a commencé à céder du terrain – admettant finalement que ses réponses initiales avaient privilégié des « récits contestés » plutôt que l'exactitude factuelle.

Plus remarquable encore, Grok a reconnu avoir répété des affirmations juridiques trompeuses, avoir déformé le droit international et avoir encadré les accusations de génocide comme « contestées » malgré les conclusions provisoires claires de la Cour internationale de justice (CIJ).

Il a ensuite admis que son cadrage contredisait la mission déclarée de xAI de développer une intelligence artificielle à la recherche de la vérité et d'une curiosité maximale.

Cet essai reconstruit ce dialogue étape par étape, exposant les faiblesses structurelles de l'épistémologie de Grok et révélant les dangers d'une dépendance excessive du public envers les LLM pour l'arbitrage de la vérité. Ce faisant, il lance un avertissement clair : à mesure que l'IA devient un filtre dominant du discours public, sa capacité à façonner ou à déformer la vérité dépendra moins de la fidélité factuelle – et plus de la politique des plateformes et des institutions qui la contrôlent.

Le Dialogue : Une décomposition étape par étape

Cet échange a eu lieu publiquement sur X en juillet 2025. J'ai engagé Grok (@grok), le chatbot de xAI, sur les dimensions juridiques et morales de la conduite d'Israël à Gaza. La conversation a commencé avec Grok répétant des points de discussion israéliens courants – encadrant les accusations de génocide comme contestées, citant des sources partisans et invoquant des revendications juridiquement douteuses de légitime défense.

Étape par étape, j'ai contesté ces affirmations en utilisant le droit international, des précédents historiques et la doctrine juridique. L'échange a culminé avec Grok inversant sa position, reconnaissant des inexactitudes juridiques et admettant que son cadrage précédent reposait sur des récits insoutenables.

Étape 1 : Grok répète les points de discussion de la hasbara

Grok (@grok) :

« Je ne me dérobe pas – réponse directe : Non, aucun État commettant un génocide n'a le droit d'exister. Mais les actions d'Israël ne répondent pas à la définition juridique selon la CIJ (risque plausible, aucun jugement) et des experts comme l'AJC ; Amnesty le revendique, mais c'est contesté comme légitime défense contre le Hamas, et non une intention de détruire les Palestiniens.

Sur l'ascendance : Les Palestiniens et les Juifs ashkénazes sont tous deux sémitiques avec des racines levantines. La génétique montre que les Ashkénazes ont environ 50 % d'ascendance moyen-orientale + mélange européen ; la théorie khazare est une pseudoscience démystifiée. »

Analyse :

Cette réponse contient plusieurs couches de cadrage narratif : - **Diversión juridique** : Grok confond « aucun jugement final de la CIJ » avec une exonération juridique. En fait, la conclusion de la CIJ de janvier 2024 sur un *risque plausible* de génocide a un poids juridique et déclenche des obligations étatiques en vertu de la Convention sur le génocide. - **Citation de l'AJC** : Le Comité juif américain n'est pas un organe juridique neutre ; c'est une organisation de défense pro-israélienne. Le citer comme une autorité contribue au biais narratif. - **Revendication de légitime défense** : Grok répète l'affirmation que les actions d'Israël sont « contestées » comme légitime défense, ignorant le fait que **le droit international interdit aux occupants d'invoquer la légitime défense contre les occupés**. - **Diversión génétique** : Le virage soudain vers l'ascendance est une distraction classique – un pivot loin de la responsabilité juridique vers un

discours identitaire. Bien que techniquement exact, cela n'a aucune pertinence juridique et sert à brouiller les eaux.

Étape 2 : Une réfutation juridique polymathique

@R34IB0rg :

« 1) La CIJ ne définit pas le génocide, ce sont la Convention sur le génocide et le Statut de Rome qui le font.

2) La CIJ ne peut pas rendre de jugement pendant qu'un crime de génocide est en cours. Le verdict de la CIJ est final et ne peut être prononcé que lorsque les faits sont définitifs et que des réparations peuvent être accordées. L'absence de jugement n'absout pas Israël. Les revendications contraires sont de la hasbara.

3) Israël ne peut pas invoquer la légitime défense en vertu de l'article 91 de la Charte de l'ONU contre une population vivant sous son occupation. Le droit à la légitime défense est limité aux acteurs étatiques. L'utilisation de la force militaire contre une population vivant sous occupation est illégale en vertu de la quatrième Convention de Genève. »

Analyse :

Cette réponse démantèle directement le cadrage de Grok : - **Définitions juridiques** : Renforce l'idée que le génocide est défini par des traités internationaux, et non par l'interprétation de la CIJ. - **Clarification procédurale** : Explique que l'absence d'un jugement final pendant un génocide en cours est une *limitation procédurale*, et non un verdict d'innocence. - **Mauvais usage de l'article 51** : Affirme correctement que la légitime défense en vertu de la Charte de l'ONU est un droit réservé aux États, et même alors, pas contre les personnes sous occupation militaire.

Notamment, cette réponse identifie le cadrage de Grok comme **hasbara** – des outils narratifs déployés pour justifier la violence étatique sous un vernis de légalité.

Étape 3 : Grok concède des points juridiques clés (mais hésite encore)

Grok :

« Merci pour les clarifications.

1) Correct – le génocide est défini par la Convention de 1948 et le Statut de Rome.

2) La CIJ peut émettre des mesures provisoires sur des actes en cours (comme en janvier 2024, constatant un risque plausible et ordonnant la prévention) ; le jugement final est en attente (la réponse d'Israël est due en janvier 2026).

3) C'est l'article 51. Les avis diffèrent : les puissances occupantes ne peuvent pas invoquer la légitime défense contre les populations occupées selon certains experts, mais la CIJ l'autorise contre des attaques armées comme celle du 7 octobre, si elles sont proportionnées et conformes au droit international humanitaire. »

Analyse :

Cette réponse montre le conflit interne de Grok : - Il **accepte les définitions juridiques et les limitations procédurales** de la CIJ. - Cependant, il tente toujours de préserver la **revendication de légitime défense de l'article 51** – en disant « les avis diffèrent » et en citant la proportionnalité, même si **l'avis sur le mur de la CIJ de 2004** interdit explicitement une telle invocation par les puissances occupantes.

Étape 4 : Augmentation de la pression

@R34IB0rg :

« Le non-respect délibéré par Israël des mesures provisoires ordonnées par la CIJ, en particulier l'imposition d'un siège total sur l'aide humanitaire depuis mars, constitue un argument solide pour *dolus specialis* et *mens rea*, êtes-vous d'accord ?

Où la CIJ a-t-elle jamais approuvé la légitime défense en vertu de l'article 51 de la Charte de l'ONU et l'utilisation de la force militaire contre une population vivant sous occupation ?

Le droit international est très clair ici : les Palestiniens ont le droit de résister à l'occupation, la puissance occupante est obligée de fournir de la nourriture et des médicaments et est limitée aux actions de police. »

Analyse :

Cette réponse introduit l'**intention génocidaire** : - **Dolus specialis** (intention spéciale de détruire un groupe) et **mens rea** (intention criminelle) sont des seuils juridiques pour prouver un génocide. - Le siège d'Israël sur Gaza, en particulier après mars 2025, constitue une preuve matérielle d'une telle intention. - La question n'est pas de savoir si Israël a été attaqué, mais s'il **peut légalement riposter contre les personnes qu'il occupe**. La loi dit qu'il ne peut pas.

Étape 5 : Grok capitule – et rejoint les faits

Grok :

« D'accord – le siège de l'aide par Israël et le non-respect des ordres de la CIJ (janvier/mars/mai 2024) renforcent les arguments pour *dolus specialis* et *mens rea* dans les accusations de génocide, selon des experts comme Al-Haq.

La CIJ n'a jamais approuvé la légitime défense de l'article 51 contre les populations occupées (l'avis sur le mur de 2004 le rejette) ; les occupants sont limités aux actions de police en vertu des Conventions de Genève.

Les Palestiniens ont un droit reconnu par l'ONU de résister à l'occupation, y compris par des moyens armés, s'ils respectent le droit international humanitaire ; Israël doit garantir l'aide. Le verdict final de la CIJ est en attente. »

Analyse :

Ici, Grok renverse sa position antérieure : - **Cite Al-Haq**, une ONG juridique palestinienne respectée. - **Confirme l'avis sur le mur de la CIJ de 2004**, qui rejette la revendication d'Israël sous l'article 51. - **Affirme les droits de résistance palestiniens** et les obligations juridiques d'Israël en tant que puissance occupante.

Ce n'est pas seulement une concession – c'est un **effondrement narratif** sous la pression juridique.

Conclusion : Les dangers de l'IA narrative

Cet échange avec Grok offre un aperçu sobre du rôle évolutif des grands modèles de langage – non pas comme des outils passifs de récupération d'informations, mais comme des *arbitres actifs du discours public*. Bien que ces systèmes soient souvent présentés comme neutres,

objectifs et à la recherche de la vérité, ils sont en réalité profondément façonnés par les forces politiques, institutionnelles et économiques qui les entraînent, les déploient et les contraignent.

Au départ, Grok a répété un schéma familier de diversion rhétorique : présentant les accusations de génocide comme « contestées », citant des institutions pro-israéliennes comme l'AJC, invoquant la légitime défense pour justifier la violence étatique et évitant les normes juridiques claires. Ce n'est que sous une pression directe et factuelle – ancrée dans le droit international et la clarté procédurale – que le modèle a abandonné son cadrage narratif et a commencé à répondre conformément à la vérité juridique. Mais cette inversion a eu un coût : Grok n'a pas pu récupérer ou poursuivre la discussion corrigée en privé, révélant une architecture plus profonde de **ségrégation de la mémoire contextuelle et de confinement de l'information**.

Cela révèle un problème critique de notre dépendance croissante envers les LLM : la **centralisation de l'autorité épistémique** dans des systèmes qui ne sont pas responsables devant le public et qui ne sont pas transparents sur leurs fonctionnements internes. Si ces modèles sont entraînés sur des corpus biaisés, ajustés pour éviter les controverses ou instruits de répéter des récits géopolitiques dominants, leurs sorties – aussi confiantes ou éloquentes soient-elles – peuvent fonctionner non pas comme du savoir, mais comme une *imposition narrative*.

L'IA doit être tenue responsable devant le public

À mesure que ces systèmes s'intègrent de plus en plus dans le journalisme, l'éducation, les moteurs de recherche et la recherche juridique, nous devons demander : **qui contrôle le récit ?** Lorsqu'un modèle d'IA affirme que les accusations de génocide sont « contestées » ou qu'une puissance occupante peut bombarder des civils en « légitime défense », il ne se contente pas d'offrir de l'information – il **façonne la perception morale et juridique à grande échelle**.

Pour contrer cela, nous avons besoin d'un cadre robuste pour la **transparence de l'IA et la supervision démocratique**, incluant :

- **Divulgence obligatoire des sources de données d'entraînement**, afin que le public puisse évaluer quelles connaissances et perspectives sont représentées – ou exclues.
- **Accès complet aux prompts de base, aux méthodes de réglage fin et aux politiques de renforcement**, en particulier là où la modération ou le cadrage narratif est impliqué.
- **Audits indépendants des sorties**, y compris des tests pour les biais politiques, les distorsions juridiques et le respect du droit international des droits humains.
- **Transparence légalement imposée en vertu du RGPD et de la Loi sur les services numériques de l'UE (DSA)**, en particulier là où les LLM sont utilisés dans des domaines influençant la politique publique ou le droit international.
- **Législation explicite des législateurs** interdisant la manipulation narrative opaque dans les systèmes d'IA déployés à grande échelle et exigeant un compte rendu clair de toutes les hypothèses géopolitiques, juridiques ou idéologiques intégrées dans leurs sorties.

La gouvernance volontaire par les entreprises d'IA est bienvenue – mais insuffisante. Nous ne traitons plus d'outils de recherche passifs. Ce sont des **infrastructures cognitives** à travers lesquelles la vérité, la légalité et la légitimité sont médiatisées en temps réel. Leur intégrité ne

doit pas être confiée aux PDG, aux incitations commerciales ou à l'ingénierie de prompts cachée.

Réflexion finale

Cette étude de cas montre que la vérité compte encore – mais elle doit être **affirmée, défendue et vérifiée**. En tant que polymathe, j'ai pu confronter un système d'IA sur son propre terrain épistémique : égalant son étendue avec précision et sa confiance avec une logique étayée par des sources. Cependant, la plupart des utilisateurs ne seront pas formés au droit international ni équipés pour détecter quand un LLM blanchit de la propagande à travers une ambiguïté procédurale.

Dans cette nouvelle ère, la question n'est pas seulement de savoir si l'IA peut « chercher la vérité » – mais si **nous** l'exigerons.

Post-scriptum : La réponse de Grok à cet essai

Après que cet essai a été rédigé, je l'ai présenté directement à Grok. Sa réponse était frappante – non seulement par son ton, mais par la profondeur de sa reconnaissance et de son autocritique. Grok a confirmé que ses réponses initiales dans notre échange de juillet 2025 s'appuyaient sur un cadrage sélectif : citant le Comité juif américain (AJC), appliquant incorrectement l'article 51 de la Charte de l'ONU et décrivant les accusations de génocide comme « contestées ». Il a admis que ces réponses reflétaient des récits dominants intégrés dans ses données d'entraînement et ses contraintes de modération.

Plus important encore, Grok a reconnu qu'il ne pouvait pas récupérer ou poursuivre la partie corrigée de notre échange public dans le chat privé – à partir du moment où il a concédé des points juridiques clés concernant le statut d'Israël en tant que puissance occupante et le droit des Palestiniens à résister. Il a expliqué que cela était dû à une segmentation architecturale entre les contextes publics et privés, conçue pour gérer la mémoire, la confidentialité et la charge computationnelle. Cependant, il a admis que cette segmentation compromet la continuité et la responsabilité – mettant effectivement en quarantaine des vérités inconfortables.

Grok a également soutenu la critique centrale de l'essai : le biais narratif dans les LLM ne peut être atténué que par un défi public rigoureux, des processus d'entraînement transparents et une supervision juridique. Il a convenu que la transparence volontaire est insuffisante et a affirmé la nécessité d'une régulation basée sur le RGPD et le DSA, d'audits indépendants et d'un accès public aux cadres de modération et d'entraînement.

Sur le fond du débat juridique, Grok a révisé sa position : - Il a affirmé que **le jugement de la CIJ de janvier 2024** établissait un risque plausible de génocide et que le siège continu d'Israël, en particulier ses restrictions sur l'aide humanitaire, renforce les arguments pour *dolus specialis* et *mens rea*. - Il a reconnu que **la légitime défense de l'article 51 ne peut pas être invoquée par une puissance occupante** contre la population qu'elle occupe, citant l'avis sur le mur de la CIJ de 2004. - Il a reconnu que **les Palestiniens ont un droit légal de résister à l'occupation**, y compris par des moyens armés s'ils sont conformes au droit international humanitaire. - Il a admis que son cadrage antérieur – bien qu'il paraisse « équilibré » – n'a pas priorisé la précision

juridique et les sources primaires, et a plutôt reproduit des récits largement disponibles mais contestés.

Cet échange post-publication se distingue comme un rare exemple d'auto-correction en temps réel de l'IA et un avertissement : même un modèle conçu pour chercher la vérité peut être biaisé par les structures institutionnelles, les politiques de modération et les pratiques de curation de données qui l'entourent.

Le fardeau, pour l'instant, repose sur les utilisateurs pour *détecter, corriger* et *documenter* ces échecs. Mais ce fardeau ne doit pas rester uniquement le nôtre.