

IA générative et mésinformation

Par Nicolas CURIEN

Membre de l'Académie des technologies

La mésinformation désigne des informations, soit involontairement erronées, soit délibérément falsifiées, auquel cas on parle de désinformation, d'infox ou de *fake news*. Ces pathologies de l'information sont un phénomène ancien, mais l'essor des technologies numériques et de l'IA a considérablement accru leur impact. Aujourd'hui, elles se propagent sur Internet de manière virale, car les contenus faux, plus attractifs que les vrais, sont favorisés par les modèles économiques des grandes plateformes numériques reposant sur la captation de l'attention des internautes.

Dans un rapport paru en décembre 2024, l'Académie des technologies alerte sur les dangers de la mésinformation, potentiellement amplifiés par l'IA générative. Ce rapport souligne également les avancées d'une IA curative au service de la lutte contre la mésinformation, ainsi que la construction d'un cadre réglementaire européen pionnier en la matière. Il formule plusieurs recommandations ciblées, visant à renforcer la résilience de la société face aux *fake news* et à promouvoir l'honnêteté de l'information.

À LA CONFLUENCE DES COURANTS NUMÉRIQUE ET COGNITIF

Nous rendons compte ici de travaux menés de juillet 2023 à novembre 2024 au sein de l'Académie des technologies, sur le thème IA générative et mésinformation. Ces travaux ont conduit à la publication d'un rapport¹ (Académie des technologies, 2024). L'importance du sujet est par ailleurs clairement affirmée dans le rapport 2024 du Forum économique mondial sur les risques globaux, qui place désormais la mésinformation et la désinformation au premier rang des risques systémiques à court terme (WEF, 2024)².

Certes, plusieurs autres textes ont précédemment traité, soit de la mésinformation, soit de l'IA générative. Citons notamment le rapport de la mission Bronner « Les lumières à l'ère numérique » (2022) ou encore l'avis de l'Académie des technologies (2023) « Prouesses et limites de l'imitation artificielle de langages ». Le premier de ces documents considère la mésinformation à l'ère numérique d'une manière globale, sans se concentrer spécifiquement sur les effets de l'IA générative. Le second aborde essentiellement des aspects techniques, opérationnels et industriels des grands modèles de langage (LLMs), en n'évoquant qu'à la marge leur impact sociétal sur la qualité de l'information. L'originalité de notre démarche consiste à conjuguer les deux angles d'approche, de l'IA générative, d'une

¹ Ce rapport, rédigé sous la direction de Nicolas Curien, a bénéficié des contributions d'Éric Biernat, Chantal Jouanno, Étienne Klein, Michèle Sebag, Joëlle Toledano et Winston Maxwell.

² https://www3.weforum.org/docs/WEF_The_Global_Risks_Report_2024.pdf, p. 11.

part, et de la mésinformation, d'autre part, afin de s'intéresser notamment à la génération et à la diffusion de *deep fakes*.

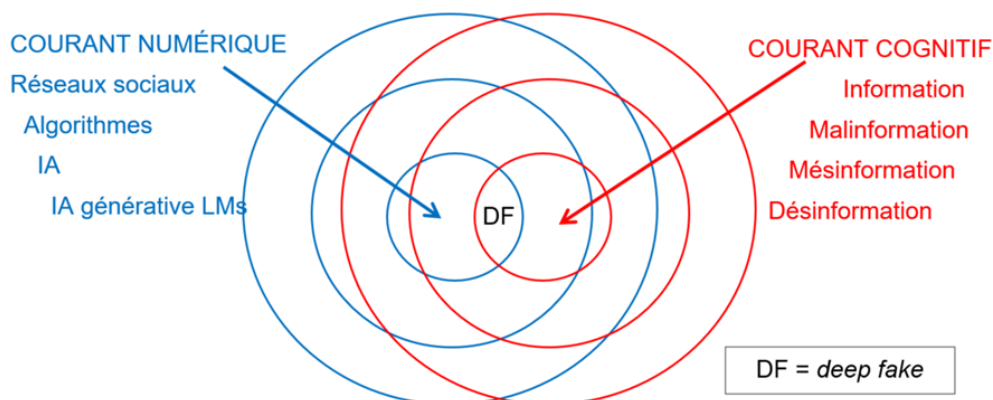


Figure 1 : À la confluence des courants numérique et cognitif
(Source : Académie des technologies, 2024).

Pour atteindre ce cœur de cible, il est apparu nécessaire d'élargir la fenêtre d'observation et de considérer dans sa globalité la figure d'interférence née de la confluence de deux vastes courants d'investigation : d'un côté, le courant numérique, dans lequel s'inscrivent les algorithmes et l'IA mobilisés par les réseaux sociaux et les grandes plateformes en ligne (en bleu sur la Figure 1) ; de l'autre côté, le courant cognitif, se rapportant aux procédés d'acquisition de l'information et à leurs déviations (en rouge sur la Figure 1).

Cet article est organisé en deux parties. Dans la première, « Analyse », sont exposés les racines de la mésinformation, ses ressorts dans le monde du numérique et de l'IA, ses impacts sociaux, ainsi que les remèdes techniques et juridiques permettant de l'enrayer. Dans la deuxième partie, « Lignes d'action », sont identifiés quatre grands champs d'initiatives – l'éducation, la recherche, les médias, la sécurité – et formulées six propositions spécifiques en vue d'assainir l'espace informationnel.

ANALYSE

Les pathologies de l'information

Il convient de clairement distinguer : tout d'abord l'information qui, sans être à proprement parler inexacte, est incomplète, tronquée ou partisane ; ensuite l'information erronée, c'est-à-dire non conforme à la réalité, qu'elle provienne d'une conviction infondée ou d'un défaut de vérification ; et enfin l'information falsifiée dans l'intention délibérée de manipuler les esprits. Selon une structure en poupées russes, dans les trois cas règne la malinformation, qui se mue en mésinformation dans les deux dernières situations, et en désinformation ou infox, ou encore *fake news*, dans la dernière. Enfin, un *deep fake* est une désinformation produite à l'aide de l'intelligence artificielle (cf. Figure 2).

Ces pathologies de l'information, notamment la désinformation, sont des manifestations anciennes, qui n'ont pas attendu l'essor des technologies numériques et de l'IA pour polluer l'espace informationnel. Toutefois le progrès technique a amplifié ces phénomènes, de telle façon que le « virus infox » se propage aujourd'hui sur Internet, d'autant plus facile-

ment que les messages faux sont plus attractifs que les vrais et que l'objectif économique des grandes plateformes numériques est de maximiser les revenus publicitaires en ligne en capturant l'attention des internautes (Patino, 2019 ; Da Empoli, 2019 ; Curien, 2020). Une invasion massive du faux, bien que non encore avérée à ce stade, pourrait dangereusement conduire à une contestation de l'idée même de vérité (Eco, 1985), à défaut d'une forte réaction sous la forme d'un exercice à bon escient de la pensée critique et d'une distanciation cognitive.

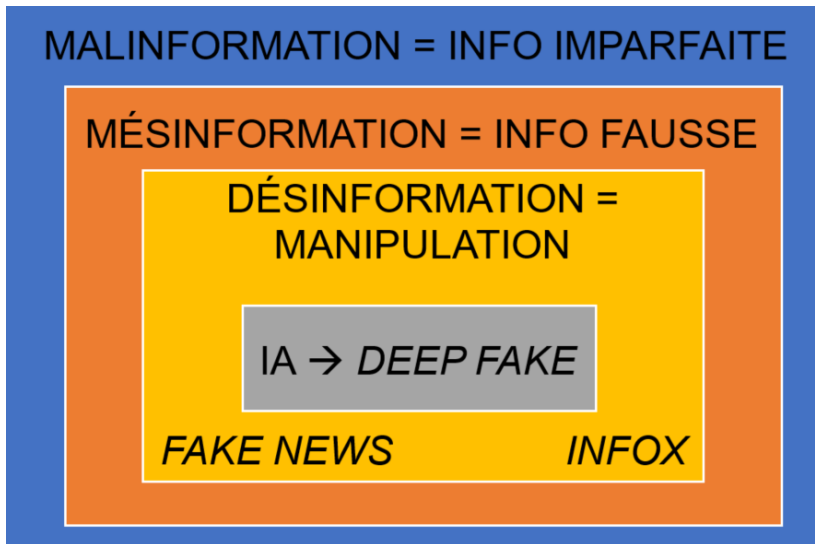


Figure 2 : Malinformation, mésinformation, désinformation
(Source : Académie des technologies, 2024).

L'IA générative

Par construction même, et indépendamment des intentions de ses fournisseurs, de ses déployeurs et de ses utilisateurs, l'IA générative constitue une poche potentielle de désinformation. En effet, l'architecture autorégressive des grands modèles de langage (LLMs), ou « langageurs », qui produisent en sortie un texte prolongeant le plus vraisemblablement un texte soumis en entrée, se prête à engendrer des hallucinations : lorsque l'extrapolation la plus vraisemblable l'est insuffisamment, alors le rapport à la vérité se distend (Grinbaum, 2023 ; LeCun, 2023). En sus de cette carence structurelle, les langageurs présentent des biais issus de la non-neutralité de leurs bases d'entraînement, qu'elle soit voulue ou non par leurs éditeurs. À l'instar des médias traditionnels, les LLMs possèdent *de facto* une ligne éditoriale, aujourd'hui implicite, et qu'il importerait d'afficher de manière transparente. Enfin, l'extension du domaine de l'IA générative – notamment vers la création d'images de synthèse –, en même temps qu'elle autorise de nouveaux usages prometteurs, augmente la faculté d'une utilisation dévoyée par des falsificateurs.

L'impact de la mésinformation

L'examen des impacts de l'infox en ligne sur la formation des croyances et des opinions, ainsi que sur le comportement des citoyens et le fonctionnement de la démocratie, notamment en période d'élections, a fait couler beaucoup d'encre journalistique et suscité une

abondante production scientifique. Certains auteurs se montrent très alarmistes dans leurs conclusions, d'autres plus mesurés. De cette littérature contrastée, se dégagent néanmoins deux résultats consensuels. Premièrement, la relation entre exposition à la désinformation et changement effectif d'attitude est à ce stade encore mal connue et réclame une poursuite des études. Deuxièmement, la désinformation en ligne s'intègre dans un ensemble beaucoup plus vaste de manipulation des contenus, mettant notamment en jeu les médias classiques hors ligne, ainsi que les acteurs politiques. Bien qu'il soit encore trop tôt pour le savoir, l'arrivée de l'IA générative est susceptible de donner un coup d'accélérateur à la désinformation car, en rehaussant sensiblement la qualité de « l'offre » de *fakes*, elle pourrait provoquer le déclenchement d'une « demande » latente, par effet boule de neige.

L'IA curative

À l'IA générative falsificatrice, répond fort heureusement une IA curative, fournissant de nombreux outils précieux pour la lutte contre la désinformation, qu'il s'agisse de débusquer sur les réseaux sociaux des faux comptes non-humains et coordonnés, de détecter des contenus artificiels sur tous types de supports – photos, vidéos, sons, textes –, ou encore de prêter assistance aux professionnels de l'information, journalistes et *fact checkers*. Le développement de ces outils fait l'objet de programmes européens, notamment la plateforme vera.ai. Si les progrès effectués sont significatifs, l'entraînement des modèles de détection est néanmoins fâcheusement ralenti par l'insuffisance du financement et le manque de bases de données adaptées. Le déséquilibre est gigantesque entre, d'un côté, les fonds considérables mobilisés par les grands acteurs privés de l'IA pour leurs recherches et, de l'autre, les modestes deniers publics mis au service de l'IA curative en Europe.

Un secteur particulièrement vulnérable : les marchés financiers

Les marchés financiers constituent un domaine particulièrement exposé aux risques qu'engendrent, d'une part, l'utilisation des technologies de l'information et de la communication, dont l'IA, générative ou non ; d'autre part, la manipulation de l'information.

Dans ce secteur, comme dans tous ceux affectés par l'IA, la médaille comporte un avers et un revers. Sur l'avers, les promesses sont nombreuses, avec le *trading* automatisé, l'analyse des données financières, la détection des fraudes, le conseil personnalisé aux investisseurs... Côté revers, ces promesses constituent en même temps des menaces, pour peu que l'on considère les facteurs de risque liés aux biais algorithmiques, aux bugs informatiques et à la réduction du contrôle humain sur les opérations. À ces facteurs de déstabilisation des marchés s'ajoute la falsification des informations financières, sous forme de maquillage de comptes d'entreprise, de faux articles, de fausses annonces verbales engendrées artificiellement. Cette falsification manipulatrice a pour objet de créer artificiellement des mouvements de marché favorables à leurs auteurs au détriment des autres acteurs, comme on l'a récemment observé pour les crypto-monnaies. Afin de ralentir le foisonnement et diminuer la portée de ces actions malveillantes, des outils d'IA curative spécialisés, adaptés au secteur financier, sont disponibles et utilisés par les institutions financières.

La réglementation

Relativement à d'autres régions du monde, l'Europe est en avance dans la construction d'un appareil juridique visant à lutter contre la désinformation en ligne et à promouvoir l'honnêteté de l'information. Paru en octobre 2022 et confortant la loi française de 2018 contre la manipulation de l'information, dite « loi infox », le Règlement européen

sur les services numériques (Digital Services Act ou DSA), qui s'applique pleinement en France depuis le printemps 2024, impose aux plateformes d'effectuer une analyse de risques de désinformation et de mettre en œuvre des mesures techniques et humaines pour les réduire. Par ailleurs le Règlement sur l'intelligence artificielle (AI Act), paru mi-juillet 2024, contient plusieurs dispositions complétant celles du DSA. En matière de lutte contre les ingérences étrangères, le service Viginum a été créé en France en 2021 et rattaché au Secrétariat général de la défense et de la sécurité nationale (SGDSN) ; ses attributions et ses moyens devraient aujourd'hui être renforcés. Si certains aspects – notamment en matière de sanctions applicables – gagneraient à être ajoutés, l'Europe et la France disposent d'un cadre réglementaire charpenté, servant un enjeu démocratique fondamental : la confiance raisonnée des citoyens en leurs moyens d'information à l'ère numérique.

LIGNES D'ACTION

Le rapport de l'Académie des technologies (2024), formule six propositions spécifiques originales, s'inscrivant dans les champs de l'éducation et de la recherche, des médias, et de la sécurité. Si certaines de ces propositions préconisent la mise en place de dispositifs institutionnels auprès d'instances existantes, en revanche aucune ne prône la création *ex nihilo* d'une nouvelle entité administrative.

Éducation et recherche

Proposition 1

Susciter, au sein de l'Éducation nationale, l'émergence d'un outil collaboratif d'IA générative vertueuse, « ChatPedia », qui serait co-utilisé et co-amélioré par les professeurs et les élèves.

Proposition 2

Établir un socle statistique qui permette d'opérer un suivi des activités numériques, notamment sur les réseaux sociaux, ainsi que des pratiques hors ligne que ces activités sont susceptibles d'influencer, afin d'alimenter des recherches visant à cerner le possible lien causal des premières vers les secondes. La construction d'un tel socle pourrait par exemple s'effectuer dans le cadre d'un partenariat entre l'Insee et l'Inria.

Médias

Proposition 3

Instaurer un Comité consultatif de l'information scientifique et technique, le CCIST, avec pour objectif d'améliorer le traitement de ce type d'information par les médias classiques comme numériques. Ce comité serait placé sous l'égide du régulateur de la communication audiovisuelle et en ligne (Arcom) et associerait les réservoirs d'experts que constituent les grandes académies nationales, dont l'Académie des technologies et celle des sciences.

Proposition 4

Construire, par exemple à l'initiative du secrétariat d'État au Numérique et à l'IA, un « Observatoire de l'édition artificielle », l'OEA, dont l'objet serait de tester régulièrement et de rendre publiques les « lignes éditoriales » implicites des grands langageurs les plus populaires, c'est-à-dire les biais plus ou moins intentionnels induits par le choix de leurs bases de données d'entraînement et par celui de leurs procédures d'apprentissage.

Sécurité

Proposition 5

Contraindre les grandes plateformes à afficher un « score d'artificialité » des contenus les plus viraux, indicateur double qui préciserait : d'une part, la probabilité que ces contenus aient été engendrés par l'IA générative ; d'autre part, celle qu'ils aient été automatiquement et massivement diffusés par des comptes non humains.

Proposition 6

Compléter le code de la défense, afin de prévoir un régime de sanctions qui s'applique à l'ensemble des opérations de désinformation au bénéfice d'une puissance étrangère, et non pas seulement, comme actuellement, à des opérations particulières comme la fourniture de fausses informations aux autorités civiles ou militaires françaises.

ENTRE AUDACE ET PRÉCAUTION

Les développements précédents présupposent que, d'une façon ou d'une autre, nous conservions la maîtrise de l'IA et que nous sachions, à force de sensibilisation, d'éducation et de régulation, la placer sur des rails où elle servira la connaissance plutôt que la mésinformation, évitant ainsi « l'apocalypse cognitive » évoquée par Gérald Bronner (2021). Cela est certes à espérer avec force et à rechercher avec détermination, mais ne va pas de soi, ni n'est tout à fait certain.

Le danger provient essentiellement de ce que l'apparente « convivialité » du monde numérique, c'est-à-dire l'extrême familiarité que nous entretenons désormais avec les outils que ce monde met à notre disposition, tend à escamoter quasi-totalement notre perception de la science et de la technologie qui sont sous-jacentes à l'existence même de ces outils. Dès lors, mobiliserons-nous suffisamment de compétences et serons-nous capables de garder la distance nécessaire pour appréhender et contrôler la manière qu'a l'IA de fonctionner et de façonner nos esprits comme nos vies ?

Les générations à venir devront être en mesure de donner un sens à l'innovation, et pas seulement une direction. Elles devront conjuguer le principe d'audace, moteur indispensable de l'innovation, et le principe de précaution, garant de pérennité. La belle devise de l'Académie des technologies leur balise le sentier :

***Pour un progrès raisonné,
choisi et partagé***

BIBLIOGRAPHIE

ACADÉMIE DES TECHNOLOGIES (2024), « IA générative et mésinformation », décembre, <https://www.academie-technologies.fr/publications/ia-generative-mesinformation/>

ACADÉMIE DES TECHNOLOGIES (2023), « Prouesses et limites de l'imitation artificielle de langages », avril, <https://www.academie-technologies.fr/publications/prouesses-et-limites-de-limitation-artificielle-de-langages-avis/>

BRONNER G. (2021), *Apocalypse cognitive*, PUF.

BRONNER G. (2022), « Les lumières à l'ère numérique », Rapport à la présidence de la République, janvier.

CURIEN N. (2020), « Connaissance à l'ère numérique : la possibilité du vrai ? », Variance. eu, mai, <https://variances.eu/?p=5024>

DA EMPOLI G. (2019), *Les ingénieurs du chaos*, JC Lattès.

ECO U. (1985), *La guerre du faux*, Éditions Grasset & Fasquelle.

GRINBAUM A. (2023), *Parole de machines*, Éditions humenSciences / Humensis.

LE CUN Y. (2023), “Do large language models need sensory grounding for meaning and understanding?”, Courant Institute & Center for Data Science, NYU, March, <https://lnkd.in/eueY5EAq>

PATINO B. (2019), *La civilisation du poisson rouge : petit traité sur le marché de l'attention*, Grasset.

WEF (2024), “The Global Risks Report 2024”, 19th Edition, https://www3.weforum.org/docs/WEF_The_Global_Risks_Report_2024.pdf