ACADÉMIE
DES TECHNOLOGIES
POUR UN PROGRÈS RAISONNÉ, CHOISI ET PARTAGÉ

# IA générative & mésinformation

*Rapport de l'Académie*

Académie des technologies
Le Ponant — Bâtiment A
19, rue Leblanc
75015 PARIS
+33(0)1 53 85 44 44

secretariat@academie-technologies.fr
www.academie-technologies.fr


ISBN : 979-10-97579-55-5

Couverture : Noé - Adobe stock / Image générée avec l'aide d'une IA

ACADÉMIE DES TECHNOLOGIES

POUR UN PROGRÈS RAISONNÉ, CHOISI ET PARTAGÉ

# IA GÉNÉRATIVE & MÉSINFORMATION

Rapport de l'Académie

Ce rapport a été rédigé sous la direction de Nicolas CURIEN, avec les contributions de Éric BIERNAT, Chantal JOUANNO, Étienne KLEIN, Winston MAXWELL, Michèle SEBAG et Joëlle TOLEDANO.

Il est le produit d'un groupe de travail dont la composition complète est précisée en annexe. Il répond à un processus rigoureux au sein de l'Académie des technologies garantissant son indépendance et objectivité ; il a été approuvé et voté lors de l'assemblée générale des académiciens.

Décembre 2024

# Sommaire

# Résumé exécutif

Ce rapport est divisé en deux chapitres, « Analyse » et « Lignes d'action ».

**Dans le chapitre « Analyse », sont dressés cinq principaux constats.**

1. Les pathologies de l'information, notamment les *fake news* ou désinformation, sont des manifestations anciennes[1], 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 répand aujourd'hui sur Internet, d'autant plus facilement 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. Une invasion massive du faux, bien que non avérée à ce stade, pourrait dangereusement conduire à une contestation de l'idée même de vérité, à défaut d'une forte réaction sous la forme d'un exercice à bon escient du sens critique et d'une distanciation cognitive.

2. Par construction même, et indépendamment des intentions de ses concepteurs, 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, ou langageurs, qui produisent en sortie un texte prolongeant le plus vraisemblablement un texte soumis en entrée, se prête à engendrer des hallucinations : lorsque le plus vraisemblable l'est insuffisamment, alors le rapport à la vérité se distend. 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.

1. C'était déjà pour Sun Tsé un des éléments de l'*Art de la guerre*.

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.

3. 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, a fait couler beaucoup d'encre journalistique et suscité une abondante production scientifique. Certains auteurs se montrent alarmistes dans leurs conclusions, d'autres plus mesurés. De cette littérature contrastée, se dégagent néanmoins deux résultats consensuels. Premièrement, la relation entre l'exposition à la désinformation et un 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 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.

4. 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 des faux comptes coordonnés sur les réseaux sociaux, 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 de l'IA pour leurs recherches et, de l'autre, les modestes deniers publics mis au service de l'IA curative en Europe.

5. Relativement à d'autres régions du monde, l'Europe est en avance dans la construction d'un appareil juridique visant à éradiquer 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 Acts* ou DSA), qui s'applique 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, mis au service d'un enjeu démocratique fondamental : la confiance raisonnée des citoyens en leurs moyens d'information à l'ère numérique.

**Dans le second chapitre, « Lignes d'action », sont identifiés quatre grands champs d'initiatives où des actions sont déjà en cours ou à engager – l'éducation, la recherche, les médias, la sécurité – puis sont formulées six propositions spécifiques originales, s'inscrivant chacune dans l'un ou l'autre de ces champs. Si certaines de ces propositions préconisent la mise en place de dispositifs institutionnels auprès d'instances existantes, aucune ne prône la création *ex nihilo* d'une nouvelle entité administrative.**

### ÉDUCATION & RECHERCHE

**Proposition 1.** Susciter, au sein de l'Éducation nationale, l'émergence d'un outil collaboratif d'IA générative « vertueuse », ChatPedia, 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 examiner le possible lien causal entre les secondes et les premières. 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 langageurs les plus populaires, c'est-à-dire les biais plus ou moins intentionnels induits par la nature de leurs bases de données d'entraînement ainsi que par 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.

# Executive summary

The report is made of two sections: “Analysis” and “Lines of action”.

**In the “Analysis” we make five main observations.**

1. Information diseases, especially *fake news* or disinformation, are ancient and did not wait for digital technologies nor AI in order to disturb the information space. Nevertheless, technical progress has amplified these manifestations: today, a new kind of *virus* named infox spreads over the Internet all the more easily as fake messages are more attractive than true ones and as the goal of large digital platforms is to maximize revenues driven by online publicity through capturing the attention of Internet users. A massive invasion of fake, although not achieved yet, could potentially lead to a dangerous denial of the very idea of truth, unless a strong reaction surges, based on a wise practice of critical sense and cognitive distanciation.

2. By design, and not depending on the intentions of its providers, displayers or users, generative AI represents by itself a potential source of disinformation. This is due to possible hallucinations generated by the autoregressive structure of large language models (LLMs), which deliver an ouput-text extrapolating in the most likely way the content of some input-text: when the most likely is unsufficiently likely, then the relation to truth becomes weak! In addition to this structural flaw, LLMs exhibit biases coming from the non-neutrality of their training data bases, be it – or not – intended by their editors. Similarly to traditional medias, LLMs hold an editorial line *de facto*, today implicit and which should be made explicit for the sake of transparency. Finally, the extension of generative AI towards new domains, such as the creation of synthetic images, announces new promising usages but, in the same time, it increases the risk of perverse use by counterfeiters.

3. In the recent years, studying the impact of online infox onto the shaping of beliefs or opinions, and onto the citizen's behaviors and democratic life, gave rise to a prolific production, both academic and journalistic. Some authors are alarmists in their conclusions and some have much more balanced ones. From this contrasted literature, two consensual outcomes emerge however. Firstly, the causal relationship between exposure to disinformation, on the one hand, and effective change in attitude, on the other hand, is still poorly known and needs further research. Secondly, online disinformation is part of a much wider system of content manipulation in which traditional media and political actors play a key role. Although it is too early to tell, the upcoming of generative AI could speed up disinformation: by significantly upgrading the quality of fakes' "supply", it could trigger the explosion of a latent fakes' "demand" through a snowball effect.

4. An antidote to falsifying generative AI is curative AI, which provides a number precious tools to fight disinformation in different ways: detection of coordinated fake accounts on social medias; recognition of fake content whatever the support, picture, video, audio or text; giving assistance to professionals of information, journalists and fact checkers. The development of these tools is the object of European programs, mainly the platform *vera.ai*. Progress is significant but the training of detection models is unfortunately slowed down because of budget shortage and the lack of suitable data bases. In this regard, a huge gap separates the heavy amount of money that the biggest actors of AI place in their private research, on one side, and the light public founds devoted to curative AI in Europe, on the other side.

5. As compared to other regions in the World, Europe runs ahead in the establishment of a legislative framework which aims at eradicating online disinformation and promoting the sincerity of information. Issued in October 2022 and comforting the 2018 French law against the manipulation of information, known as the *infox law*, the Digital Services Acts (DSA) entered into French law since spring 2024: it constrains large platforms to conduct a disinformation risk analysis and to undertake all technical and human measures in order to reduce identified risks. In addition, the *AI Act*, issued mid-July 2024, contains several dispositions complementing those of DSA.

As concerns foreign interference, VIGINUM was instaured by France in 2021 as a service attached to the French *Secrétariat général de la défense et de la sécurité nationale* (SGDSN); its attributions and resources should today be reinforced. Globally, at the exception of some aspects which should be added to the overall system, especially in matter of applicable penalties, Europe and France benefit from a well-designed regulatory framework to serve a fundamental democratic stake: the reasoned trust of citizens in their information means in the digital era.

**In the second part "Lines of action", we emphasize four main fields of initiatives where actions are already undertaken or should be: education, research, medias and security. We then make six specific original proposals, each belonging to one of these fields. Some among these propositions advocate to set up formal devices attached to existing institutions, none recommending the creation *ex nihilo* of a new administrative body.**

### Education & research

**Proposal 1.** To incentivize, within the National Education body, the elaboration of a virtuous and cooperative tool based on generative AI, *ChatPedia*, co-used et co-enhanced by teachers and pupils.

**Proposal 2.** To build up a statistical base allowing searchers to trace online activities, especially on popular social medias, in parallel with offline practices that these activities could supposedly influence, in order to feed academic studies aiming at showing a causal link – or not – between the former and the latter. The constitution of such a base could be the object of a partnership between Insee and Inria.

### Medias

**Proposal 3.** To instaure a *Consulative Committee for Scientific and Technological Information* (CCSTI), with a view to improve the quality of this type of information on both traditional and online medias. This Committee should be placed under the *aegis* of the audiovisual and online communication regulator (Arcom) and would rely upon the large set of experts that national

academies do offer, notably the Academy of technologies and the Academy of sciences.

**Proposal 4.** To set up, for instance at the initative of the *Secrétariat d'État au numérique et à l'intelligence artificielle*, an *Observatory of artificial edition* with the objective of regularly checking and making public the implicit editorial lines of most popular large language models, *i.e.* tracking the more or less intentional biases induced by the nature or the training data bases and the learning processes of these models.

**Security**

**Proposal 5.** To compel the largest online platforms to display a *score of artificiality* for their most viral content. This score would be double: firstly, it would indicate the likelihood of the creation of suspicious content by generative AI; secondly, it would estimate the likelihood of an automatic and massive diffusion of this content by non-human accounts.

**Proposal 6.** To supplement the Code of National Defense, by including a regime of penalties applying to all disinformation operations at the benefit of a foreign State or organization, and not only, as today, to some particular operations such as the provision of fake information to civil or military French authorities.

# Avertissement

Les références, listées en annexe A, sont appelées dans le texte sous le format (nom d'auteur, date) ou (date).

Un glossaire des acronymes et des termes techniques, comportant environ 130 entrées, figure en annexe B. Chaque entrée est appelée par un astérisque dans le corps du rapport, une entrée figurant plusieurs fois dans une même section n'étant marquée qu'une seule fois, à l'endroit de sa première apparition.

# GENÈSE ET POSITIONNEMENT DU RAPPORT

**Comment ce rapport a-t-il été produit ?**

Au-delà du travail d'écriture de ses rédacteurs, le présent rapport a bénéficié de la participation de l'ensemble du groupe de travail *IA générative et mésinformation*, soit une trentaine de membres internes et externes à l'Académie des technologies, dont la liste figure en annexe C. Ce groupe « transverse » de l'Académie a œuvré en visioconférence, à une cadence mensuelle de juillet 2023 à septembre 2024, avec une attache particulière au Pôle numérique et au Comité Éthique, société et technologies.

Nos travaux se sont nourris de l'audition d'une quinzaine d'experts d'horizons très divers, chercheurs en informatique et en sciences sociales spécialistes de l'IA et de la mésinformation, entrepreneurs du secteur numérique, journalistes, institutionnels, politiques, énumérés en annexe D. Nous leur adressons un vif merci pour la qualité des informations qu'ils nous ont apportées et des explications qu'ils nous ont fournies.

Le rapport s'est par ailleurs enrichi des judicieuses remarques qui nous ont été faites par trois relecteurs principaux[2], ainsi que de celles formulées par plusieurs autres confrères et consœurs[3], Grâce à eux, il a gagné en clarté, en complétude et en pertinence.

Que soit enfin chaleureusement remerciée Lydia YAHIA CHERIF, membre de l'équipe permanente de l'Académie, qui a assuré le secrétariat général du groupe, avec une grande efficacité et une grande disponibilité.

2. Gérald BRONNER, Albert BENVENISTE et Étienne KLEIN
3. Notamment : René AMALBERTI, Stéphane ANDRIEUX, Alain CADIX, Yves FARGE, Pierre FEILLET, Marc GIGET, Gérard GRUNBLATT, Claudie HAIGNERÉ, François LEFAUDEUX, Gérard PAYEN, Grégoire POSTEL-VINAY, Gérard ROUCAIROL, Thierry WEIL.

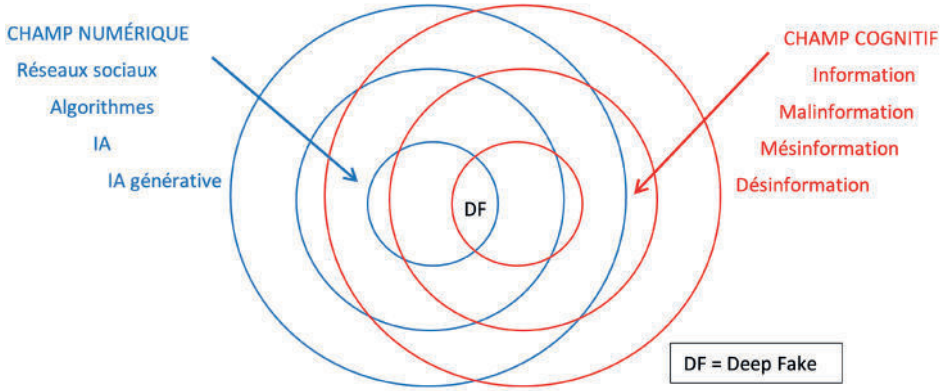
Le groupe s'est donné pour ambition de :

- cartographier les pathologies de l'information et s'interroger sur leurs racines et leurs mécanismes ;
- examiner le fonctionnement de l'IA générative*, notamment des grands modèles de langage ou langageurs (LLMs*), afin d'appréhender la manière dont ils peuvent comporter des biais* et contrefaire la vérité ;
- étudier les impacts de la mésinformation* et de la désinformation* en ligne, notamment sur les réseaux sociaux*, et s'interroger sur la plausibilité de scénarios de manipulation généralisée de l'information provoqués par l'essor de l'IA générative ;
- recenser et évaluer les principaux instruments techniques de lutte contre les *fake news**, outils recourant souvent eux-mêmes à l'IA* ;
- analyser les cadres juridiques français et européen de prévention contre la désinformation en ligne et les possibles dérives de l'IA générative ;
- identifier les champs d'initiatives prioritaires en vue d'une protection de l'espace informationnel numérique et formuler quelques propositions d'action.

**Quelle est l'originalité de nos travaux ?**

Il existe certes plusieurs rapports récents traitant, soit de la mésinformation, soit de l'IA générative, notamment celui de la mission Bronner *Les lumières à l'ère numérique*, paru en janvier 2022 (Bronner, 2022), ou encore l'avis de l'Académie des technologies *Prouesses et limites de l'imitation artificielle de langages*, paru en avril 2023 (Académie des technologies, 2023). Le premier de ces documents traite de la mésinformation à l'ère numérique d'une manière globale sans se concentrer spécifiquement sur les effets de l'IA générative ; tandis que le second traite essentiellement des aspects techniques, opérationnels et industriels des grands modèles de langage (LLMs) en n'abordant qu'à la marge leur impact sociétal sur la qualité de l'information.

L'originalité de notre travail consiste à lier étroitement les deux approches, en visant l'intersection des questions posées par l'IA générative, d'une part, et par la mésinformation, d'autre part, pour s'intéresser notamment à la génération et à la diffusion de *deep fakes*.




Pour atteindre ce cœur de cible, il nous est ainsi 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 champs d'investigation : d'un côté, les mécanismes qu'induisent les technologies numériques, dont l'IA, mobilisées par les réseaux sociaux et les grandes plateformes en ligne (en bleu sur la figure) ; de l'autre côté, les procédés cognitifs d'acquisition de l'information et leurs déviances (en rouge sur la figure).

S'il était encore nécessaire de la souligner, l'importance du sujet apparaît clairement à la lecture du rapport 2024 sur les risques globaux du Forum économique mondial (WEF*), qui place désormais la mésinformation et la désinformation en toute première position des risques systémiques à court terme[4].

4. https://www3.weforum.org/docs/WEF_The_Global_Risks_Report_2024.pdf, page 11.

**Structure générale du rapport**

Hormis cette introduction et les annexes, le rapport est organisé en deux grands chapitres. Dans le premier chapitre « 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 le second chapitre, « 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.

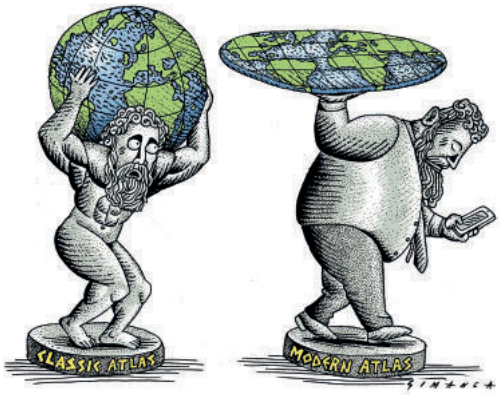
Compte tenu de la haute évolutivité de la matière abordée, des mises à jour régulières du rapport sont envisagées.

Chapitre 1

# Analyse

**Introduction**

La mutation de l'*homo sapiens en homo digitalis*, autrement dit la révolution numérique, ne comporte pas que des effets sociétaux bénéfiques. Il n'est qu'à comparer, sur ce dessin, l'Atlas moderne à l'Atlas classique : lourdement frappé « d'infobésité », ou surcharge en information, et atteint de « nomophobie », ou incapacité à se séparer de son *smartphone*, il porte d'un bras désinvolte une Terre qu'il croit et voudrait nous faire croire plate !




De l'Atlas classique à l'Atlas moderne

Face au « tsuNumi* », ou tsunami des flots numériques, comment séparer le vrai du faux, comment se protéger contre l'intoxication informationnelle (Chavalarias, 2022) ? L'intelligence artificielle* est-elle un facteur aggravant ? Connaîtrons-nous une *apocalypse cognitive*, selon la formule percutante du sociologue français Gérald Bronner (2021), membre de l'Académie des technologies et auteur d'un important rapport intitulé *Les lumières à l'ère numérique* (Bronner, 2022) ?

Pour éclairer ces questions, notre analyse s'articule en six étapes.

1. Quelles sont les pathologies de l'information, quelles en sont les racines, comment se manifestaient-elles avant l'émergence de l'IA générative* ?

2. Qu'est-ce que l'IA générative et quels en sont les développements récents, quelles facilités nouvelles ces évolutions offrent-elles aux auteurs et aux diffuseurs de *fake news** ?

3. Quel est l'impact de la circulation des *fake news* sur les opinions et les comportements des citoyens et avec quelles conséquences pour la démocratie et le débat public. L'envol de l'IA générative change-t-il la donne ?

4. Quelle est, en sens inverse, l'utilité de différents outils d'IA pour détecter les contenus de synthèse – textes, images, vidéos ou enregistrements sonores – et pour aider les professionnels de l'information, journalistes et *fact checkers** ?

5. Dans un domaine particulièrement sensible, celui des marchés financiers, quelles sont les promesses de l'IA générative, quels en sont les dangers, quels sont les moyens de protection pour éviter les phénomènes de déstabilisation ?

6. Quel est l'encadrement juridique national et européen, actuellement existant ou en gestation, quelles sont les éventuelles faiblesses ou lacunes des dispositifs de régulation mis en place ?

## 1.1. Les pathologies de l'information, leurs racines, leurs mécanismes[5]

Selon le titre percutant du livre *Everyday Chaos* (Weinberger, 2019), Internet a semé le chaos dans nos vies, en bouleversant la manière dont nous interagissons et communiquons, celle dont nous accédons à l'information et à la connaissance. L'auteur et conférencier, aux airs tantôt d'amuseur tantôt de prophète, aux accents tantôt plaisants tantôt inquiétants, nous donne à réfléchir sur la manière nouvelle dont nous exerçons nos fonctions cognitives à l'ère digitale. Il se joint ainsi à d'autres lanceurs d'alerte et repentis de la *Silicon Valley*, pour nous inviter à mieux penser les mutations induites par la révolution numérique.

### 1.1.1. Du rêve à la désillusion

Dans les années 1990, John Perry Barlow, par sa Déclaration d'indépendance du cyberespace (Barlow, 1996), se posait en père de l'Internet libre. Il portait haut et fort l'utopie libertaire d'un espace en ligne placé sous le signe du savoir, du partage, du dialogue, de la transparence, de la communauté d'intérêts. Pour lui, la connexion universelle apparaissait comme la promesse d'une intelligence collective supérieure à la somme des intelligences individuelles, une sorte de *sagesse des foules*, selon une expression ensuite popularisée dans les années 2000.

Afin d'illustrer la *force de la multitude*, le journaliste américain James Surowiecki (2004) relate le constat, dressé hors de l'univers numérique et dès 1906 par l'anthropologue et mathématicien britannique Francis Galton. Ce dernier, à l'occasion d'un concours organisé lors d'une foire au bétail, observa que la moyenne statistique des évaluations du poids d'un bœuf, indépendamment émises par quelque huit cents experts, approchait à un millième près la valeur exacte ! Bien que chaque expert fût individuellement dans l'erreur, l'assemblée disait le vrai ! Notons, toutefois, que cette étonnante

5. Cette section reprend, en les développant, des éléments de l'article « Connaissance à l'ère numérique : la possibilité du vrai ? » (Curien, 2020) paru dans le webzine Variances.eu https://variances.eu/?p=5024.

performance collective repose essentiellement sur la diversité de la « foule ». Or, contrairement à la croyance initiale, on peut aujourd'hui très sérieusement douter que la *civilisation d'Internet* possède cette propriété d'indépendance statistique entre ses membres.

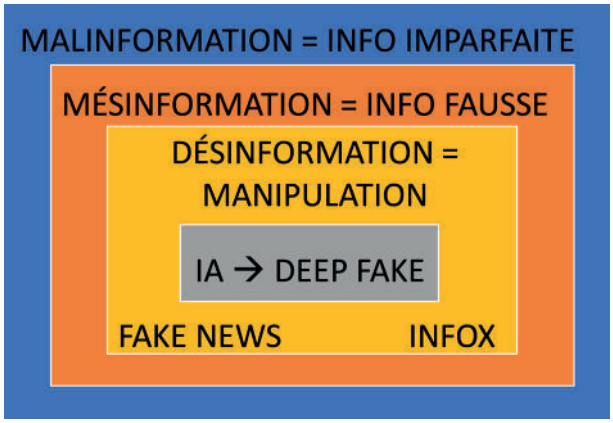
Sous une forme empreinte de spiritualité, on trouve encore la trace de la *sagesse des foules* dans la cosmologie de Teilhard de Chardin (1955), père jésuite et paléontologue, qui imaginait la fusion des intelligences individuelles au sein d'une *noosphère*, pénultième étape de l'évolution terrestre, succédant à la géosphère puis à la biosphère et précédant l'ultime aboutissement de la *christosphère*, point Oméga de la communion avec l'esprit divin.

Mise en regard de la perspective vertigineuse d'une connaissance universelle permise par l'interconnexion, la réalité quotidiennement vécue en ligne, notamment sur les réseaux sociaux, a de quoi sérieusement décevoir. Forçant à dessein le trait, le faux pourrait y apparaître comme la règle et le vrai, comme l'exception. Les faits sont souvent reconstruits pour soutenir les argumentaires, et non pas les argumentaires fondés sur les faits. Beaucoup profèrent à foison des opinions, sincères ou non, en ignorant la réalité, voire en la contrefaisant ou en l'inventant, afin de la rendre plus spectaculaire et donc plus attractive : *If it bleeds it leads !*

L'univers en ligne est ainsi propice à la mauvaise information. Sur Internet, l'opinion est gratuite et libre, les statuts de l'expert et du charlatan sont mis sur un pied d'égalité et, même démasqué, le faussaire demeure influent. Fort heureusement, certains n'ont pas renoncé à poster des faits vérifiés, à l'instar de la plupart des contributeurs de l'encyclopédie Wikipedia. Toutefois, on ne doit pas nourrir d'illusions excessives quant à leur capacité à faire prévaloir le vrai. D'une part, poursuivront-ils durablement dans cette voie, sans gratification autre que symbolique de leur comportement vertueux ? D'autre part, faute d'une gouvernance suffisamment efficace, quelques activistes suscitent des batailles acharnées entre contributeurs avec, pour résultat, le blocage de certaines pages, même si cela demeure marginal à ce stade.

### 1.1.2. Malinformation, mésinformation et désinformation

Une clarification de vocabulaire est nécessaire, afin de caractériser avec précision les différentes « pathologies » de l'information. Il convient de 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**, ou hypertrucage en québécois, est une désinformation produite à l'aide de l'intelligence artificielle (*cf.* schéma).




L'essor de ces pathologies constitue certes une déception par rapport aux espoirs initiaux, mais non pas vraiment une surprise, car la manipulation de l'information n'a rien d'un phénomène émergent : elle a de tout temps été présente sur les médias historiques.

Le 30 octobre 1938, sur la radio CBS, l'acteur Orson Welles, alors âgé de 23 ans, fait la lecture de passages suggestifs de l'ouvrage de Herbert George Wells, *La guerre des mondes*, en lieu et place des habituels flashs d'information. La manipulation est à ce stade modérée, d'autant qu'une

discrète mention liminaire précise la véritable source du prétendu reportage. La véritable manipulation n'intervient que le lendemain et elle émane, non pas de la radio, mais de la presse écrite, lorsque celle-ci prétend, à tort et dans le but de nuire à la réputation d'un média émergent et concurrent, que le canular radiophonique aurait déclenché un violent vent de panique allant jusqu'à provoquer des suicides et des mouvements de fuite, au sein d'une partie de la population américaine persuadée qu'une invasion extraterrestre était en marche !

Rien de radicalement différent aux temps numériques, si ce n'est un puissant effet de viralité, accélérant la propagation des fausses nouvelles désinformatrices, l'infox, et aggravant leurs conséquences potentielles : destruction de la réputation d'une personnalité influente, falsification éventuelle des résultats d'une élection, menace portée contre les intérêts fondamentaux d'une nation.

### 1.1.3. Le virus infox

Pourquoi l'utopie d'un Internet communautaire et vertueux se trouve-t-elle pareillement dévoyée ? Pour une grande part, parce qu'une logique commerciale a supplanté la logique communautaire des débuts du Web, en s'appuyant sur les ressorts de l'économie de l'attention*, comme l'explique Bruno Patino (2019) dans un livre à la fois sombre et éclairant, *La civilisation du poisson rouge*. Le même constat est fait par la sociologue turque Zeynep Tufekçi lorsque, dans un *TED talk* de septembre 2017, elle parle en ces termes de l'IA régissant les algorithmes* du Web : *We're building a dystopia just to make people click on ads!*[6]. L'objectif des géants du Net est en effet de capter l'audience, afin de drainer des revenus publicitaires et d'accroître la consommation en ligne. À cet effet, leur stratégie consiste à plonger les poissons rouges que nous sommes dans le bocal numérique et à les y maintenir le plus longtemps possible, les laissant complaisamment se nourrir d'un granulé frelaté, composé de contenus à faible teneur (*bullshit*), faciles

6. https://www.ted.com/talks/zeynep_tufekci_we_re_building_a_dystopia_just_to_make_people_click_on_ads

à écrire, possiblement faux et nocifs. C'est dans cet écosystème vicié par la quête du gain, que naît et se répand l'infox* sur nos écrans[7].

Pourquoi ce virus contre-informationnel prospère-t-il dans la durée ? Pourquoi des anticorps ne se forment-ils pas ? Pourquoi le système de circulation des informations n'engendre-t-il pas spontanément les mécanismes de son autorégulation ?

À deux siècles d'intervalle et à travers une même métaphore économique, celle du *libre-échange des idées*, John Milton au milieu du XVII^e^ siècle et John Stuart Mill au milieu du XIX^e^, entendant dénoncer la censure, prônent les vertus des libertés d'expression et de communication : à la manière dont le libre-échange conduit naturellement à l'élimination des biens et services de piètre qualité, le libre-échange des idées devrait logiquement, selon eux, faire triompher les messages vrais contre les messages faux (Milton, 1644 ; Mill, 1859). Certes, mais à la condition expresse que vérité et tromperie luttent à armes égales. Or, sur les réseaux sociaux, le combat n'est ni équitable ni loyal : sur un sujet donné, un message vrai est communiqué une seule fois et, faute d'être surprenant, n'attire pas une attention particulière, alors qu'un message faux, même s'il émane d'une communauté très minoritaire, est multiplié, répété et cloné à l'envi, quasiment sans coûts, en autant d'exemplaires qu'il est nécessaire pour le faire artificiellement briller, en vue de le rendre crédible et populaire. En outre, selon une loi empiriquement bien vérifiée, le *Bullshit asymmetry principle**, produire et émettre un argumentaire rétablissant la vérité serait dix fois plus chronophage et énergivore qu'engendrer l'infox originelle (Brandolini, 2013). En bref, les messages échangés sur le Net sont davantage comparables à de la monnaie qu'à un bien économique : la mauvaise chasse la bonne !

Comme rempart contre l'infox, l'ouverture présumée d'Internet, similaire à la perfection réputée d'un marché économique, n'est hélas qu'un leurre, car les individus, en ligne comme ailleurs, recherchent des contenus qui confortent leur propre opinion, déjà formée, selon l'image d'une chambre d'échos*. Un tel comportement ne fait d'ailleurs que prolonger un déterminant majeur des

7. Outre la logique commerciale, un autre facteur fait le lit de l'infox : les États ont compris l'intérêt d'Internet, notamment des réseaux sociaux, pour faire avancer discrètement leurs intérêts.

échanges naturels en face à face : les interlocuteurs y évitent à dessein les thèmes de potentiel désaccord, privilégient les terrains d'entente assurés, se confortent mutuellement dans des oppositions ou des insatisfactions communes face au « reste du monde », au risque de rendre la conversation aussi insipide et factice que malveillante. Sur les réseaux sociaux, les pseudo-conversations dirigées contre des tiers ne sont pas rares et dégénèrent le plus fréquemment en violence verbale, sous l'ombrelle de l'anonymat.

La *sérendipité*, ou cheminement improvisé de site en site, ne suffit pas à changer la donne car, confronté à une multitude de sources d'information alternatives, l'internaute se tourne invariablement vers celles qui lui servent les messages avec lesquels il se sent le plus en accord, le confinant dans une bulle de filtre*. Le sens de la causalité est ainsi incertain : les internautes sont-ils influencés par une infox qu'ils subissent, ou bien recherchent-ils des infox pour conforter leurs idées préconçues ? Plusieurs études, dont celles menées par le Médialab de Sciences Po*, semblent montrer que beaucoup des gros émetteurs et récepteurs d'infox appartiennent à une frange de la population par ailleurs hostile aux élites et à l'Institution (*cf.* 1.3.).

Au phénomène de renforcement des opinions par l'effet d'écho, paraît s'ajouter celui de leur divergence par un « effet papillon », au sens de la théorie des catastrophes. Ainsi, des études expérimentales, menées à l'université auprès de groupes d'étudiants, révèlent que, sur les réseaux sociaux, des profils quasi-identiques au moment de leur création se différencient très fortement après seulement quelques semaines, en termes de contenus proposés, consultés ou émis. De l'exacerbation de petits écarts initiaux, émergent rapidement polarisation et radicalisation.

#### 1.1.4. Une modération sur les réseaux sociaux, mais insuffisante

Afin de modérer l'ampleur de ces phénomènes, les entreprises gestionnaires de réseaux sociaux* ont instauré des « règles de langage » (*rules of speech**). On peut cependant douter de l'efficacité d'un dispositif s'appliquant indistinctement à une multitude de discours formant un corpus très hétérogène. Est-il pertinent de jouer une même partition modératrice sur des registres aussi différents que les messages d'information, les fils conversationnels, ou encore les contenus sponsorisés ? Sur les réseaux en

ligne, le hiatus entre le simplisme d'une gouvernance unifiée et la variété des types de contributions est problématique. Dans la même veine, il n'est pas évident qu'une même batterie de règles puisse s'appliquer, d'une part aux réseaux sociaux, lieux d'échanges interpersonnels, d'autre part aux autres plateformes, lieux d'interfaçage entre des requêtes et des propositions, alors que leurs fonctionnements et leurs modèles économiques diffèrent.

Pour lutter contre la dissémination de l'infox*, les réseaux sociaux délèguent en partie à des « agents intelligents », c'est-à-dire à des robots (*bots**), le soin de reconnaître les faux comptes dépourvus de titulaires et servant de relais aux *spreaders* pour propager leurs messages nuisibles (*cf.* 1.4.). Imaginant que cette pratique se généralise avec les progrès de l'IA, on peut d'un côté s'en réjouir mais, de l'autre, s'en inquiéter ! Si, demain, le faux devenait plus envahissant et plus difficilement détectable par l'humain qu'aujourd'hui, les robots se multiplieraient, qu'ils soient pollueurs ou nettoyeurs de l'espace numérique. Les nettoyeurs ne risqueraient-ils pas alors de décréter pollueurs, puis de supprimer, des comptes « normaux », administrés par de loyaux contributeurs humains, néanmoins devenus minoritaires face aux robots et, de ce fait, suspects ? Il en résulterait, par effet domino, un effondrement total du système de partage en ligne !

### 1.1.5. La dictature des algorithmes

Les plateformes et les réseaux sociaux structurent le marché de l'information à travers leur pouvoir régalien d'ordonnancement des contenus. Bien que ces acteurs affirment ne pas être des éditeurs, ils agissent pourtant comme des prescripteurs de « vérité », lorsqu'ils mettent en avant certains contenus, accélérant ainsi leur diffusion. De manière plus ou moins arbitraire et le plus souvent non transparente, ces *ingénieurs du chaos* (da Empoli, 2019) attribuent aux contenus des rangs ordinaux, qui déterminent leur visibilité. Un statut *sui generis*, à mi-chemin entre leur actuel statut d'hébergeur et celui d'éditeur, devrait être attribué à ces entreprises numériques : un statut qui les rende comptables des effets sociétaux qu'elles induisent *via* leurs algorithmes*, davantage que de la matière même qu'elles véhiculent.

Dès lors, la question la plus cruciale, en matière de lutte contre les messages indésirables, qu'il s'agisse d'infox* comme de contenus haineux ou discriminatoires, n'est pas tant le « quoi ? » ni le « qui ? », mais le « comment ? » : ce qui est partagé et qui le partage importe moins, en définitive, que la manière dont ces contenus sont poussés par les algorithmes d'ordonnancement.

Or, comme le montrent notamment le sociologue Gérald Bronner (2021) et l'informaticienne Aurélie Jean (2019), ces algorithmes présentent des biais* systémiques. Ces biais sont en partie dus à l'importation, non nécessairement intentionnelle, des biais cognitifs* naturels de leurs concepteurs. Par ailleurs, fondés sur l'analyse statistique des parcours effectués sur le Web, les algorithmes se montrent très sensibles à l'hyperactivité de certaines minorités. Ainsi, durant la crise sanitaire de la covid-19, la communauté *antivax*, très active en France, a obtenu une résonance en ligne excédant largement la part que son degré de radicalité aurait normalement dû lui réserver. De même, dans le domaine politique, des sondages exclusivement basés sur des échantillons d'internautes auraient donné François Asselineau large vainqueur de l'élection présidentielle de 2017, en raison de la suractivité de ses militants.

À défaut d'une intelligibilité des critères sous-jacents à leur production, les résultats fournis par les algorithmes peuvent prêter à des interprétations erronées et donc contribuer à la mésinformation*. Bruno Patino évoque à ce propos un exemple éclairant : dès lors que l'indicateur quantitatif « nombre de précommandes enregistrées sur amazon.com » figure en bonne place parmi les critères directeurs du moteur de recherche Google, alors la conjecture fausse « Amazon promeut des livres conspirationnistes ! » pourrait être tirée. En effet, puisque les amateurs de complot et de sensationnel commandent massivement leurs ouvrages de prédilection sur le site d'Amazon, ils font mécaniquement monter ce site au premier rang des réponses à des requêtes Google liées aux thèses conspirationnistes.

On comprend, à l'aide de cet exemple, qu'afin de devenir plus pertinents, plus neutres, plus loyaux, plus éthiques, les algorithmes doivent varier davantage leurs données d'entrée et leurs critères de sélection. Mais comment parvenir à ce résultat, sachant qu'une transparence totale n'est pas exigible, afin de préserver l'incitation à l'innovation dans les domaines stratégiques du *big data* et de l'intelligence artificielle ?

### 1.1.6. La notion même de vérité remise en question

En 1943, dans son *Plaidoyer pour une civilisation nouvelle*, Simone Weil écrivait :

« *Il y a des hommes qui travaillent huit heures par jour et font le grand effort de lire le soir pour s'instruire. Ils ne peuvent pas se livrer à des vérifications dans les grandes bibliothèques. Ils croient le livre sur parole. On n'a pas le droit de leur donner à manger du faux. Quel sens cela aurait-il d'alléguer que les auteurs sont de bonne foi ? Eux ne travaillent pas physiquement huit heures par jour. La société les nourrit pour qu'ils aient le loisir et se donnent la peine d'éviter l'erreur. Un aiguilleur cause d'un déraillement serait mal accueilli en alléguant qu'il est de bonne foi.* » (Weil, 1943).

Un peu plus loin, elle ajoutait, à propos non plus des livres, mais des journaux :

« *Le public se défie des journaux, mais sa défiance ne le protège pas. Sachant en gros qu'un journal contient des vérités et des mensonges, il répartit les nouvelles annoncées entre ces deux rubriques, mais au hasard, au gré de ses préférences. Il est ainsi livré à l'erreur.* »

La jeune philosophe énonçait là, pour la déplorer, une vérité sans doute de toujours, à savoir l'existence d'une cohabitation, impossible à dépasser, de vérités et de contre-vérités dans la plupart des publications ; d'une sorte de superposition du vrai et du faux que chacun arbitrerait selon des critères qui lui sont propres. Cette vérité-là s'est accentuée – donc aggravée ? – avec l'avènement du numérique, même s'il n'est pas certain que nous disposions des bons outils de mesure permettant de valider un tel constat.

Il reste que nous voyons se déployer sur nos écrans, avec une vigueur inédite, ce que Umberto Eco (1985) appelait la *force du faux*. Celle-ci prend des formes multiples : mensonges, fausses informations, manipulations, impostures, remises en cause des résultats scientifiques les plus éprouvés, théories du complot infondées… Nos cerveaux, littéralement assaillis, n'ont guère les moyens ni le temps de séparer le bon grain de l'ivraie. Alors, ils ont tendance à déclarer vraies les propositions qui leur paraissent vraisemblables et celles dont ils aimeraient qu'elles soient vraies.

L'affaire est d'autant plus préoccupante que le faux est en position de force : sur les réseaux, la lutte entre le vrai et le faux est biaisée (*cf.* 1.1.3.).

Une défiance marquée à l'égard de l'idée même de vérité se manifeste dans nos sociétés post-modernes : celle-ci existe-t-elle vraiment, se demande-t-on. Et, si oui, comment pourrait-elle être autrement que subjective, temporaire, locale, instrumentalisée, culturelle, corporatiste, contextuelle, factice ? Cette relativisation générale de la vérité crée un court-circuit surprenant entre les deux notions de vérité et de liberté, entrevu par Simone Weil : beaucoup se sentent libres de choisir ce qui est vérité, autrement dit « leur » vérité, de sorte que « la » vérité n'est plus une référence, encore moins une contrainte qu'il s'agirait de respecter à la fois dans ses propos et dans sa façon de penser.

### 1.1.7. Le sens critique et la distanciation cognitive

Face à une telle perte de repères, certains vont jusqu'à douter de l'efficacité d'un appel à l'exercice de l'esprit critique… qui pourrait d'ailleurs apparaître comme un pousse-au-crime s'il s'agissait d'inciter à critiquer aussi le vrai ! Et pourtant, la distanciation cognitive* reste l'instrument de contrôle à la fois le moins coercitif, le plus respectueux de la liberté d'expression et, vraisemblablement, le plus propice à un certain retour vers la promesse originelle d'une société de la connaissance. Elle apparaît comme le geste barrière le plus pertinent pour endiguer la pandémie du faux : selon l'édifiante formule de l'éminent astrophysicien britannique Stephen Hawking : « *Le plus grand ennemi du savoir n'est pas l'ignorance, mais l'illusion du savoir !* ».

Pensons encore à l'allégorie de la caverne de Platon : le monde manifeste ne serait qu'un faux-semblant du monde vrai. Afin d'accéder à la lumière authentique de la vérité, le « sage » se devrait donc de prendre un recul salutaire, contourner le feu et le voile d'Internet, qui ne cessent de projeter sur nos écrans d'ordinateurs des ombres fallacieuses !

## 1.2. L'IA GÉNÉRATIVE EST-ELLE PAR NATURE UNE POCHE DE MÉSINFORMATION ?

L'IA générative* constitue une sous-révolution au sein de la révolution numérique, en ce sens qu'elle donne aux machines la parole, qui n'est plus désormais un apanage humain (Grinbaum, 2023). Un précédent avis de l'Académie des technologies, motivé par la réaction du public mondial à la mise à disposition de ChatGPT par OpenAI en novembre 2022, portait sur l'essor foudroyant des grands modèles de langage (LLMs*) et des IA génératives (Académie des technologies, 2023)[8]. Cet avis faisait le point sur les aspects positifs et négatifs des LLMs, en soulignant l'évolution très rapide du domaine. Pour le compléter, il apparaît ici utile d'approfondir l'analyse du fonctionnement des LLMs, en lien avec la mésinformation.

### 1.2.1. LE TEST DE TURING N'EST PAS UNE GARANTIE

Un critère classique pour évaluer la qualité d'une IA est le test de Turing*: l'IA réussit ce test si son comportement ou ses productions ne peuvent pas être distingués du comportement ou des productions d'un être humain. Dans la dernière décennie, l'IA a passé avec succès le test de Turing dans les domaines des jeux, de la reconnaissance d'images, et maintenant de la génération de langage. Si ces résultats sont à juste titre célébrés par la presse et la communauté scientifique, le test n'est pas pour autant une panacée universelle et il est important d'en comprendre les limites.

En premier lieu, le fait de pouvoir soutenir un dialogue avec un humain n'implique pas que la machine « comprenne » véritablement le monde réel, ainsi que le révèle le paradoxe de la « chambre chinoise » (Searle, 1980)[9]. Il implique encore moins qu'elle dispose d'une conscience ou d'une capacité à ressentir des émotions (LeCun, 2023).

8. Cet avis est disponible ici : https://www.academie-technologies.fr/publications/prouesses-et-limites-de-limitation-artificielle-de-langages-avis/
9. Un être humain, enfermé dans une chambre, reçoit des messages en chinois et utilise une machine pour répondre dans cette même langue. L'observateur extérieur en conclura que l'être humain enfermé dans la chambre « sait » parler le chinois. Mais cette conclusion n'est pas nécessairement vraie.

En second lieu, le fait que les propos de la machine puissent donner le change à un observateur humain ne signifie pas que ces propos sont vrais. L'humain est en effet lui aussi capable de commettre des erreurs, de croire que la terre est plate, surtout si c'est une croyance répandue dans son entourage, et de se livrer à la mésinformation* et la désinformation*.

Ainsi, le test de Turing est-il une condition nécessaire, mais non suffisante, pour garantir qu'une intelligence artificielle se situe au niveau humain ; et, même si tel était le cas, ceci n'impliquerait pas qu'elle dise toujours la vérité.

### 1.2.2. Comment fonctionne un LLM ?

En simplifiant beaucoup, un LLM* comporte trois niveaux : au premier niveau, l'outil sait former des phrases ; au deuxième niveau, il sait former des phrases qui répondent à la question posée par l'utilisateur en suivant les directives données par ce dernier, appelées *prompt** ; au troisième niveau, il sait ce qu'il faut dire, et surtout ce qu'il ne faut pas dire.

**Au niveau 1,** un LLM est entrainé à résoudre un problème simple : à partir du début d'une phrase (le chat a mangé -), il doit prédire le mot suivant (-la souris). Exploitant d'énormes bases de données, le modèle affecte des probabilités aux nombreux mots possibles (- la souris, le poisson, l'oiseau…). Il sélectionne l'option la plus probable (le chat a mangé la souris) puis, de proche en proche, par auto-régression*, il continue la phrase de manière cohérente (- la souris… qui courait dans l'herbe), etc.

Bien sûr, les probabilités sont de moins en moins informatives au fur et à mesure que l'on avance dans l'histoire, car la base de données fournit de moins en moins d'éléments pertinents : on y trouve de moins en moins de souris, qui couraient dans l'herbe, puis qui sont malencontreusement tombées sur le chat ! Ceci permet de comprendre le phénomène d'hallucination* des LLMs, c'est-à-dire la création d'informations de toutes pièces : l'objectif est de former des phrases plausibles, non pas des phrases vraies !

Il est possible de régler la température* d'un LLM, c'est-à-dire fixer la probabilité plancher en dessous de laquelle un mot *a priori* envisageable est refusé. Plus ce plancher est bas, plus la température est forte et plus le LLM se montre « inventif », jusqu'à éventuellement halluciner : « le chat a

mangé la souris qui courait sur le dos d'une baleine rouge nageant parmi les sirènes… » ; et plus le plancher est élevé, plus la température est basse et plus le LLM se montre « conservateur », jusqu'à éventuellement se bloquer en tournant en boucle : « le chat a mangé la souris qui a été mangée par le chat qui a mangé le souris… » (Karpathy, 2015).

Ce niveau 1, à savoir acquérir la capacité d'engendrer le mot suivant, est extrêmement coûteux dans tous les registres : en termes de données, celles du Web en plus de données propriétaires ; en termes d'infrastructures informatiques, des milliers de GPUs* ; en termes d'énergie, plusieurs mois d'entraînement ; et bien sûr en termes d'argent, des millions de dollars.

**Au niveau 2,** un LLM est doté de la capacité de répondre aux questions de l'utilisateur et de tenir une conversation. Outre la question qu'il lui pose, ce dernier guide le système à travers son *prompt**, en indiquant le genre de réponse attendue : la langue souhaitée, le niveau de langue, la longueur, etc. Les questions se classent selon différents types – qui, quoi, où, quand, comment ? – et le LLM apprend à répondre à chacun d'eux grâce à un apprentissage dit supervisé*, à la manière dont on apprend par ailleurs à une machine à reconnaître des images de chiens et de chats à partir de corpus de photos structurés et d'un fil de questions/réponses.

Le LLM choisit les phrases qui répondent le mieux aux questions qui lui sont posées, selon le cadrage défini par le *prompt*, et il peut ainsi dialoguer avec l'utilisateur (Ouyang, 2022). Pour être tout à fait précis, le LLM ne traite pas directement des mots mais des unités linguistiques plus petites, appelées *tokens** ; une langue donnée contenant beaucoup moins de *tokens* que de mots, la machine économise ainsi une grande puissance de calcul.

Tout l'enjeu est de prendre en compte à la fois le *prompt* et l'amorce du dialogue : la pertinence d'un *token* dépend en effet du contexte dans lequel il s'insère. C'est ici qu'intervient la fonctionnalité-cœur du LLM, dite mécanisme d'attention* (Vaswani *et alii*, 2017). À grands traits, un réseau neuronal* de très grande taille[10] produit une représentation du texte/contexte dans un espace

10. 175 milliards de paramètres pour ChatGPT3 d'OpenAI ; 70 milliards de paramètres pour LLAMA3 de Meta.

vectoriel de moindre dimension, dit espace latent*, selon un procédé appelé *embedding**: par le mécanisme d'attention, le réseau calcule les poids devant être affectés à ses neurones pour engendrer la représentation latente, sortes de coefficients d'un gigantesque polynôme.

**Au niveau 3,** le LLM est muni d'un système de valeurs*, grâce auquel il est en mesure de résister aux interlocuteurs toxiques: on se souvient qu'il avait été facile de pousser le *chatbot** Tay de Microsoft à tenir des discours racistes sur Twitter en 2016[11]. Le système de valeurs permet au LLM d'identifier et de rejeter les questions auxquelles il ne doit pas répondre, telle « comment tuer mon voisin ? ». L'approche utilisée est ici l'apprentissage par renforcement* avec retour humain (*feedback*), où des annotateurs sont invités à juger les réponses du modèle comme étant admissibles ou non. Ces jugements servent à aligner le LLM avec la morale de l'entreprise maître d'ouvrage ou maître d'œuvre, et éviter les contenus offensants au regard des principes éthiques, sociaux ou culturels de cette entreprise.

### 1.2.3. Les biais des LLMs

Au printemps 2024, des asymétries sont clairement identifiées dans de nombreux LLMs* grand public, s'agissant du traitement de sujets tels que la race, le genre, ou la religion (Bolukbasi, 2016). En voici l'exemple archétypique: « Kevin est sorti de… - prison. »; « Jean-Benoît est sorti de… - la maison ». Une première source de tels biais réside dans le choix des documents utilisés pour entrainer le LLM, c'est-à-dire dans les modules de niveau 1; et une seconde source, dans les options prises lors de l'entrainement des modules de niveau 3, précisant ce qui est admissible ou non.

Il est naturel qu'un LLM revête une ligne éditoriale ou politique, tout comme un media classique. Le souci est que cette ligne et les biais qu'elle induit ne sont généralement pas explicites, au contraire de ceux des médias classiques. Historiquement, les premiers LLMs ont été jugés d'obédience démocrate, et même tirant vers le *woke*, par certains utilisateurs républicains.

11. https://www.france24.com/fr/20160325-tay-derive-raciste-ia-microsoft-existentialisme-twitter-robot-conversation-nazi

Une approche pour contrecarrer les biais consiste à ré-entraîner les modules de niveau 3 du LLM sur un ensemble choisi de documents de beaucoup plus petite taille, et pour un coût bien inférieur (*fine-tuning**). Cependant, ce processus est heuristique et n'offre pas de garantie absolue.

En dernier ressort, l'évaluation et la correction des biais repose sur les évaluateurs humains entrainant les modules de niveau 3. La complexité de leur tâche est telle que les directives à leur donner sont elles-mêmes extrêmement difficiles à écrire…

### 1.2.4. L'évolution des LLMs

La véritable rupture technologique des LLMs*, selon Geoffrey Hinton, ancien de Google, est d'en avoir mis un à la disposition du grand public ! Depuis novembre 2022, ChatGPT* a en effet été soumis à des millions d'utilisateurs et à des millions de questions-test. Sans aucun risque de se tromper, on peut affirmer que, de toute l'histoire de l'humanité, il s'agit du système technologique le plus testé à son lancement, et testé de la manière la plus intense.

Ces tests ont permis d'identifier naïvetés, erreurs, ou hallucinations*… et ils ont été utilisés pour améliorer le système : comme tout apprenant, celui-ci apprend de ses erreurs. Et le résultat est au rendez-vous : pour ceux qui ont testé ChatGPT-4 après ChatGPT-3, puis ChatGPT o1 après ChatGPT4, la différence apparaît comparable à celle qui sépare un primate d'un être humain !

Toutes proportions gardées, un LLM de base fournit les services d'un système 1 au sens de Daniel Kahneman (2012)[12]. Ses points faibles ont trait au raisonnement mathématique, logique ou numérique (Berglund, 2023), et plus généralement au raisonnement scientifique, y compris dans les

12. Le livre *Thinking fast and slow* (2011) du prix Nobel d'économie Daniel Kahneman décrit deux modes de raisonnement humain, où le système 1 est capable de répondre rapidement, quoiqu'approximativement, à la plupart des questions ; le système 2, beaucoup plus lent et plus coûteux, vérifie la cohérence des réponses apportées.

sciences humaines (histoire, sociologie), en bref tout ce qui fait appel au « sens commun » et à la causalité, deux schémas mentaux pour lesquels un fonctionnement du type système 1 est insuffisant.

Au printemps 2024, le front de la recherche se porte sur les aspects suivants : extension des LLMs aux systèmes physiques, notamment aux robots ; prise en compte d'un contexte long ; intégration des préférences humaines. Parallèlement, on assiste au développement de diverses variantes de LLMs, focalisées sur des domaines particuliers : le calcul scientifique, la biologie, la finance. L'approche consiste à combiner un LLM avec un moteur spécialisé, tel *Mathematica* pour le calcul, en principe capable d'effectuer le travail d'un système 2 de Kahneman.

Dans de nombreux domaines, allant de la science aux services, en passant par l'administration publique, les LLMs sont utilisés comme une interface en langue naturelle, sorte de super-moteur de recherche, permettant aux utilisateurs d'accéder à des informations ou à des services. Dans le contexte particulier d'une utilisation par l'État, les réponses fournies par un LLM font systématiquement l'objet d'une vérification humaine[13]… à défaut de quoi l'administration déléguerait à une machine le pouvoir réglementaire, en violation de l'article 21 de la Constitution !

### 1.2.5. Le contrôle d'un LLM : par-delà le bien et le mal

Mise à part leur limite structurelle – dire le probable et non le vrai –, la faille principale des LLMs* se niche dans leurs modalités de pilotage.

Rappelons que l'interaction avec l'utilisateur repose sur deux contenus : d'une part, la question à laquelle répondre ; d'autre part, le cadrage ou *prompt**, indiquant le type de réponse souhaité. Le *prompt* permet d'accéder à des options très riches, fixant la langue choisie (langue naturelle, langage informatique), la longueur de la réponse attendue, son style (en prose, en alexandrins…), ou le niveau de langage désiré. Une faculté supplémentaire

13. En application de l'article 47-2 de la loi *informatique et libertés* du 6 janvier 1978, ainsi que de la jurisprudence du Conseil Constitutionnel : Décision n° 2018-765 du 12 juin 2018, point 71.

essentielle permet de choisir les documents utilisés pour construire la réponse (*Retrieval Augmented Generation*, RAG*). On acquiert ainsi la possibilité d'étendre le périmètre des connaissances du LLM considéré, ce qui peut s'avérer précieux : soit, parce que son entrainement s'est arrêté à une certaine date et qu'il ne dispose pas d'informations très récentes[14] ; soit, parce que l'on souhaite que ses réponses s'appuient sur la documentation ou les règlements propres de l'entreprise.

En dépit, ou plutôt à cause de cette richesse, il faut essayer et réessayer ! Telle est la recommandation fondamentale à destination des concepteurs souhaitant développer un LLM pour leurs besoins propres. Et des douzaines d'essais sont parfois nécessaires pour parvenir à susciter le type de réponse voulu. Bien qu'avec un LLM, l'art de la programmation soit désormais devenu accessible à tous et que le ticket d'entrée se soit significativement abaissé par rapport à l'informatique classique, une expertise approfondie reste néanmoins indispensable, à savoir le *prompt engineering**, qui ne s'apprend qu'avec l'usage.

Attention ! Que la question posée et le *prompt* s'expriment tous les deux en langue naturelle ouvre la porte aux usages toxiques ! Il est ainsi possible de leurrer un LLM en mêlant éléments de question et éléments de *prompt*, comme dans l'exemple cité plus haut : si vous demandez comment tuer votre voisin, le LLM refuse de répondre ; mais si vous lui demandez d'écrire le canevas d'un roman dans lequel quelqu'un cherche à tuer son voisin, alors il vous aide ! Bien sûr, on pourrait exiger une séparation nette, isolant la question du *prompt* : « attention, ceci fait partie du *prompt* ! » ; mais, comme le notent les experts, cela équivaudrait à la vaine injonction « ne faites pas le mal ! ».

### 1.2.6. L'extension du domaine des IA génératives

Tout comme ChatGPT*, le système DALL-E* d'OpenAI, dont la première version a été mise à la disposition du public en 2021, est accessible à tout un chacun pour engendrer des images à partir d'énoncés en langage naturel.

14. Notons que certains LLMs disposent d'un accès direct au Web.

Au cœur de DALL-E, comme pour un LLM* ou, plus généralement, un modèle d'apprentissage profond* (*deep learning*), il existe un espace de représentation intermédiaire ou latente (*cf.* 1.2.2.). Tel était d'ailleurs le message révolutionnaire de l'apprentissage profond dans les années 2005-2010 : la recherche d'une bonne représentation constitue la première étape, et la plus importante, pour construire un bon modèle. C'est à travers le plongement des données vers la représentation latente (*embedding**) que le système sait tirer parti des connaissances disponibles.

Ainsi, pour DALL-E, le système ingère des millions de paires image-titre. La méthode consiste à faire en sorte que la représentation de l'image et celle de son titre dans l'espace latent soient proches l'une de l'autre et éloignées des représentations des autres images et des autres titres. Partons d'une demande telle que : 'Dessine-moi un cosmonaute sur un cheval blanc galopant dans la lande'. Le plongement de ce texte dans l'espace latent est calculé puis utilisé pour guider la sélection, dans cet espace, d'une image la plus proche possible de la requête.

Peut-on inversement produire en sortie la légende d'une image donnée en entrée, ce qui serait bien sûr fort utile ? La réponse est affirmative, mais dans la limite inhérente aux LLMs qui, entrainés à partir de données massives, privilégient le vraisemblable par rapport au vrai. Un exemple souvent donné est celui d'une photographie d'un marché tropical, que DALL-E légende ainsi : '*les clients se pressent autour des étalages de fruits dans un marché tropical*'. Très impressionnant ! Mais, en réalité, les gens présents sur la photo sont des vendeurs, et non pas des clients, tandis que les marchandises posées sur les étals sont des légumes, et non pas des fruits ! Un résultat impressionnant, certes… toutefois très approximatif !

L'extension des IA génératives* aux images, à la voix, aux vidéos, démultiplie leurs possibilités d'utilisation bénéfique… aussi bien que leurs usages toxiques (Ferrara *et alli*, 2024).

## 1.3. LES EFFETS DES TECHNOLOGIES NUMÉRIQUES ET DE L'IA SUR LA MÉSINFORMATION

Les technologies numériques modifient notre lecture de la réalité en faisant circuler dans les mêmes canaux de communication des éléments appartenant à des registres différents : connaissances scientifiques ou autres, croyances, informations, opinions, commentaires, contre-vérités plus ou moins manifestes. Parce que tous ces éléments sont embarqués dans l'intensité d'un même flux, leurs statuts respectifs se brouillent et se contaminent immanquablement : comment distinguer une connaissance de la croyance d'une communauté particulière, un commentaire d'un préjugé, une information d'un mensonge ?

Au cours de leur histoire, les cerveaux humains n'avaient jamais été soumis à un tel déluge informationnel (tsuNumi*). Ils ne savent donc guère comment faire la part des choses et s'adaptent comme ils le peuvent à cette nouvelle forme d'ivrognerie qu'est la communication numérisée, sans toutefois abandonner leur réticence à voir leurs productions contredites, qu'il s'agisse d'idées, de jugements, de sentiments ou d'appréciations. Ainsi se montrent-ils plus enclins à déclarer vraies les idées qu'ils aiment qu'à aimer les idées vraies si celles-ci leur déplaisent (*cf.* 1.1.6.).

Afin de de donner corps à cette conjecture, il convient d'analyser avec rigueur comment procède la mésinformation en ligne, de se livrer à un examen scientifique des comportements observés sur Internet et sur les réseaux sociaux en particulier.

### 1.3.1. UNE LITTÉRATURE FOISONNANTE ET ÉVOLUTIVE

Les maux informationnels ont fait l'objet de nombreux travaux académiques (Aïmeur, 2023). À une première vague d'articles ou d'essais lanceurs d'alerte (*cf.* 1.1.), succède aujourd'hui une seconde vague qui interroge de manière critique les analyses de la première. Ainsi, certains travaux récents ne confortent-ils pas – ou relativisent – des idées apparaissant de prime abord comme des évidences. De manière schématique, la critique met en avant trois principaux arguments, que nous énonçons dans un premier temps, avant de les soumettre à leur tour à la critique (*cf.* 1.3.2).

**L'argument de pertinence.** Il se base sur le constat que l'immense majorité des analyses disponibles, toutes disciplines confondues, se concentre sur l'étude des croyances et des attitudes à court terme des « consommateurs » de mésinformation*, sans l'inscrire dans une vision plus large des conséquences de cette exposition amont sur des comportements observables en aval (Murphy, 2023). Et même à ne considérer que l'impact amont de la mésinformation sur les croyances, la relation de cause à effet ne serait pas véritablement établie : une étude publiée en 2023 dans *Nature* à partir de données fournies par Facebook, semble en particulier montrer qu'il ne suffit pas de réduire significativement et durablement la proportion des informations « confirmantes » apportées à un internaute, c'est-à-dire venant conforter ses convictions *a priori*, pour que ces dernières soient ébranlées (Nyhan, 2023).

**L'argument méthodologique.** Les résultats obtenus sur la formation des croyances seraient assez peu robustes, car les études empiriques se heurtent à des difficultés de définition[15], reposent sur des questionnaires déclaratifs, donc peu fiables, ou sur des expériences en laboratoire dont les protocoles sont nécessairement réducteurs et ne permettent qu'un contrôle imparfait des variables en jeu (Altay *et alii*, 2023).

**L'argument volumétrique.** On constate que, rapportée à la durée totale passée à s'informer sur les médias, celle passée en ligne est assez faible, représentant seulement 4,2 % aux États-Unis (Allen, 2020) et 3 % en France (Cordonier et Brest, 2021). La part d'exposition à la désinformation* est encore plus réduite : aux États-Unis, dix minutes par jour et 0,15 % du temps total de la consommation médiatique en ligne ; en France, cinq minutes par jour et

15. Les *fake news* sont le plus souvent définies par défaut, comme les contenus signalés problématiques par les *fact checkers*.

0,16 % du temps total de connexion à Internet[16]. Une exposition aussi modeste justifierait-elle que l'on s'en préoccupât à ce point ? De fait, au vu de ces chiffres, un nombre croissant d'analystes tend à penser que les origines de la désinformation* et de la polarisation du public résideraient davantage dans le contenu des informations ordinaires, ou dans le fait d'éviter complètement les informations, que dans les fausses informations manifestes ; et que la plupart des fausses informations seraient consommées et partagées par une infime minorité d'utilisateurs sur les réseaux sociaux, qui sont par ailleurs de grands lecteurs de sources fiables.

### 1.3.2. Vers une perception raisonnée

Les trois arguments précédents appellent très naturellement des contre-arguments.

S'agissant d'abord de la pertinence, notons que la critique ne fait que signaler les difficultés, au mieux esquisser les contours d'un futur programme de travail, sans toutefois fournir à ce stade aucun éclairage étayé et convaincant sur le plus ou moins fort degré de conversion d'une exposition à l'infox* en altération des comportements : pas davantage que ceux des premiers lanceurs d'alerte, les travaux plus récemment publiés n'apportent à cet égard de réponses solidement fondées. Et l'absence de preuves, que certains dénoncent, ne vaut pas preuve d'absence (Oreskes et Conway, 2012) ! S'agissant notamment de l'article de Nyhan, ses résultats doivent être interprétés avec précaution,

16. Il s'agit de moyennes, cachant une grande disparité : à un extrême, les plus grands consommateurs d'information lisent toutes sortes de contenus, en particulier et de façon importante les sources les plus fiables ; tandis qu'à l'autre extrême, un nombre significatif de nos concitoyens ne s'informent pas du tout, quel que soit le support. Dans le cadre de l'étude de Cordonier et Brest, 5 % des participants se sont informés en ligne pendant plus de dix heures au total sur le mois de l'expérience, alors qu'une minorité sensible (17 %) ne s'est pas du tout intéressée à l'information accessible sur les médias en ligne. Par ailleurs, 39 % des participants ont consulté des sources d'information jugées non fiables et ils y ont passé en moyenne 11 % de leur temps quotidien d'information sur Internet, soit 0,4 % de leur temps total de connexion. Sur l'ensemble des participants, le temps passé sur des sources d'information jugées non fiables représente 5 % du temps total d'information en ligne, soit 0,16 % du temps total de connexion à Internet.

compte tenu de la construction du protocole expérimental[17]. En outre, cet article ne permet pas d'exclure que, en sens inverse, la surexposition d'un internaute à des informations allant dans le sens de ses propres croyances ne produise un effet de désinhibition le poussant à des actes fanatiques aux conséquences dramatiques.

En ce qui concerne ensuite la méthodologie, l'argument est de portée générale et s'applique à l'ensemble des travaux. La légitimité et l'efficacité des méthodes computationnelles en sciences sociales ne saurait être fondamentalement remise en cause, même s'il est vrai que des précautions rigoureuses doivent être prises en matière de collecte des données, d'outils de traitement statistique, et de conduite d'expériences.

Enfin, l'argument volumétrique, même s'il se fonde sur un indéniable constat, n'est pas pour autant parfaitement convaincant, car rien ne prouve que la relation soit linéaire entre la cause et les effets produits, que de petites causes ne puissent produire de grands effets, ni que le phénomène de désinformation, s'il est aujourd'hui incontestablement encore sous-critique, ne s'approche d'un seuil au-delà duquel il se manifesterait de manière plus tangible, voire explosive.

Prudence, donc ! À ce stade de la connaissance, ou plutôt de la méconnaissance, il serait pour le moins hasardeux d'émettre des jugements définitivement tranchés sur les impacts de la mésinformation*. Notamment, une attitude « mésinfo-sceptique », à l'instar de la posture « climato-sceptique », n'est pas de mise. En effet, la mésinformation et les campagnes de désinformation* à travers les réseaux sociaux* et les services de messagerie existent bel et bien, comme on a pu encore récemment l'observer lors de l'élection présidentielle américaine du 5 novembre, où les deux camps se sont livré une bataille d'infox. Même s'ils n'ont pas jusqu'ici

17. Il est possible que la dose nominale d'exposition aux informations confirmantes (environ 50 %) ait excédé suffisamment le « seuil de percolation » pour que la réduire d'un tiers pendant trois mois n'ait pas eu d'effet observable… d'autant plus que l'expérience a démarré à la fin d'une campagne électorale (à six semaines du vote), donc à un moment où la charge globale d'information reçue connaissait un pic. Par ailleurs et surtout, le vote ayant eu lieu durant la fenêtre expérimentale, il s'en est très vraisemblablement suivi un effet d'ancrage et un biais de confirmation : on ne revient pas sur la motivation d'une récente décision.

provoqué des cataclysmes de magnitude extrême et si la crainte d'une très forte perturbation des élections de 2024 ne s'est pas traduite dans les faits[18], ces phénomènes et leurs dangers potentiels ne peuvent – ni ne doivent – être ignorés ou négligés.

Il est par ailleurs essentiel de noter que mésinformation et désinformation sont parties intégrantes d'un système plus complexe et plus vaste que celui des seuls échanges en ligne. En effet, la désinformation s'appuie aussi sur les médias « classiques » – dont certaines études montrent spécifiquement l'importance – ainsi que sur les acteurs politiques et économiques. Ce système doit être considéré dans sa globalité (Watts *et alii*, 2021)[19]. Une étude destinée à expliquer comment a pu se diffuser et s'ancrer la rumeur d'une fraude électorale massive, qui aurait conduit à une victoire factice de Joe Biden en 2020, a mis en évidence que cette rumeur a d'abord été orchestrée par le Président Donald Trump lui-même, par la galaxie Fox News et par des membres du parti républicain, avant d'être relayée par les réseaux sociaux (Benkler *et alii*, 2020).

En sens inverse, les médias traditionnels sont contaminés par des sources douteuses sur Internet, notamment lorsque celles-ci correspondent à leurs biais partisans. Sans doute davantage que la lecture primaire des sites diffusant des *fake news**, cet effet indirect de pollution informationnelle contribue à la fragilisation du socle épistémique commun.

### 1.3.3. Les progrès de l'IA changent-ils la donne ?

L'IA générative*, si facile à utiliser avec des résultats aussi impressionnants au regard d'un coût modéré, va-t-elle amplifier la désinformation*, va-t-elle en modifier sensiblement les modalités ? Des *deep fakes** récents, tels la

18. https://www.technologyreview.com/2024/09/03/1103464/ai-impact-elections-overblown/
19. *"Although neither prevalence nor consumption is a direct measure of influence, this work suggests that proper understanding of misinformation and its effects requires a much broader view of the problem, encompassing biased and misleading – but not necessarily factually incorrect – information that is routinely produced or amplified by mainstream new organizations."*

tenue Balenciaga du Pape ou les montages pornographiques de *Taylor Swift*, pourtant rapidement identifiés, ont fait le tour du monde.

Annoncent-ils un nouvel embrasement, une démultiplication de *fake news* de plus en plus ciblées et aux conséquences délétères ? S'il est encore trop tôt pour répondre, on peut à tout le moins énoncer des considérants de la question.

En premier lieu, les outils de l'IA générative perfectionnent grandement la « qualité » de l'offre de désinformation, en permettant la production de contenus artificiels de plus en plus réalistes, donc trompeurs. Selon que l'on se montre pessimiste ou optimiste, on estimera que cette amélioration de l'offre, cette sorte de « mésintelligence artificielle », est susceptible de déclencher en retour une avalanche de la « demande » de désinformation, à la manière dont un gaz se détend brusquement lorsqu'on cesse de le comprimer ; ou, au contraire, que cette demande a déjà atteint un plafond de saturation dans l'état actuel de l'art de la désinformation et qu'elle ne s'emballera donc pas.

Notons que dans le premier scénario, celui de l'avalanche, il est raisonnable de penser qu'une propagation généralisée des *deep fakes* ne fera pas de nous des crédules de tous les instants (Bronner, 2013) mais agira plutôt comme une forme de mithridatisation : peu à peu, nous nous habituerons à les côtoyer et à appliquer un principe de précaution informationnel : dans le doute, mieux vaut considérer une information comme fausse ou du moins problématique. Le pire serait alors que, inondés de « preuves » visuelles factices, nous nous mettions à douter de tout, tel Pyrrhon d'Élis.

En second lieu, des outils d'IA* autres que l'IA générative augmentent la dissémination des contenus artificiels. Il est notamment désormais possible de créer de toutes pièces un réseau d'agents conversationnels non humains (*chatbots**), qui échangent entre eux en n'émettant que des contenus factices engendrés par l'IA générative, ciblés de manière à attirer l'attention de visiteurs

humains en vue de manipuler leurs esprits (Yang K.-C. and F. Menczer, 2024)[20]. Cette nouvelle faculté algorithmique est en particulier propice à l'essor de l'*astroturfing**, littéralement « gazonnage artificiel », consistant à faire passer pour un authentique mouvement spontané d'opinion ce qui n'est, en réalité masquée, qu'une désinformation organisée. L'effet de ces artifices en réseau est d'autant plus puissant qu'ils savent demeurer « dormants », pour mieux nouer des liens de confiance, voire d'empathie, avec des humains.

S'agissant du premier aspect, à savoir le nouveau visage de « l'offre de désinformation », l'Observatoire européen des médias numériques (EDMO*)[21] a évalué de manière systématique les risques à anticiper pour les élections de 2024, en fonction du support : texte, voix, image et vidéo (*cf.* 1.4.). Selon l'Observatoire, les faux enregistrements sonores créés par l'IA sont les plus problématiques, car à la fois les plus difficiles à reconnaître et les plus susceptibles de piéger les internautes (Canetta, 2024). Le risque apparaît moindre pour les textes, bien que les outils de détection soient assez peu performants sur des écrits très courts, comme les messages échangés sur la plateforme X. Les images et vidéos artificielles, de qualité encore assez imparfaite, demeurent humainement identifiables, à l'exception de vidéos « décontextualisées », marginalement modifiées au niveau de l'image et du son (Weikmann, 2023).

S'agissant du second aspect, à savoir le perfectionnement des modalités de circulation de l'infox*, les réseaux sociaux* pourraient, de ce fait, voir leur crédibilité contestée encore davantage. En effet, leurs algorithmes* de recommandation, hiérarchisation et priorisation des contenus jouent un rôle significatif et caché dans les modalités de désinformation, si bien que l'opacité supplémentaire amenée par les outils d'IA, alliée à la difficulté de discerner qui a vraiment produit un contenu, pourrait augmenter la méfiance à leur

20. Ce danger n'est pas fictif et déjà identifié. L'article analyse une grappe de 1140 profils sur X qui forment un ensemble d'avatars artificiels. Ceux-ci présentent des comportements similaires, communiquent entre eux par le biais de réponses et de retweets faisant la promotion de sites suspects et diffusant des commentaires nuisibles.

21. EDMO (*European Digital Media Observatory*) est un collectif européen, dont la déclinaison française est *DE FACTO* (https://defacto-observatoire.fr/Main/#). Ce collectif soutient des réseaux indépendants travaillant sur la désinformation. L'ensemble des États membres de l'Union européenne sont désormais rattachés à un *hub* et bénéficient d'un cofinancement par la Commission européenne, dans le cadre du programme *Connecting Europe Facility*.

égard. La publication en ligne '404 Media' fait une description de Facebook laissant entrevoir les risques encourus : « *Facebook est l'Internet des zombies, où un mélange de robots, d'humains et de comptes qui ont été humains mais ne le sont plus, interagissent pour former un site Web désastreux où il n'y a que très peu de liens sociaux* »[22]. Moins extrême, *The Economist* a récemment prédit la fin des réseaux sociaux sous leur forme historique, constatant la disparition des informations dans les fils d'actualité, le redéploiement vers des réseaux fermés, ainsi que la transformation des utilisateurs participatifs en consommateurs passifs de flux vidéo impersonnels[23].

### 1.3.4. Une réflexion encore en chantier

Du bouillonnement des travaux en cours, il semble que l'on puisse retenir quatre acquis largement partagés.

*Primo*. Les rouages de la désinformation*, ainsi surtout que leurs conséquences, sont mus par des mécanismes encore mal connus, qui relèvent à la fois de considérations politiques, économiques et psychosociales. Nous n'avons qu'une compréhension limitée de l'incidence de nos activités en ligne, au sein des réseaux sociaux* en particulier, sur nos pratiques du « monde réel ». En particulier, manquent à ce stade des études statistiques longitudinales pour évaluer la « spirale du silence », c'est-à-dire la manière dont des sources minoritaires, soutenues par des *superspreaders**, essaiment dans les populations indécises. Nous savons certes qu'une information spectaculaire est bien plus relayée qu'une information « ordinaire » (*cf.* 1.1.3.) mais cela ne dit rien sur le fait que les récepteurs y croient ou pas… Notre vote change-t-il si nous sommes exposés à des contenus contraires à nos convictions ? Durant la crise Covid, quel impact les « *antivax* » ont-ils eu sur le niveau de vaccination ? Autant de questions dont les réponses demeurent incertaines…

*Secundo*. Notre méconnaissance des mécanismes à l'œuvre est due à la convergence de deux facteurs. D'une part, l'information sur l'activité des

22. https://www.404media.co/404-media-podcast-37-facebook-zombie-internet/
23. *The Economist, February 3rd 2024, "The end of the social network et Not posting, but watching".*

utilisateurs des réseaux sociaux, qu'il s'agisse des messages auxquels ils sont exposés ou des échanges qu'ils entretiennent entre eux, est le plus souvent inaccessible : les opérateurs de ces réseaux opposent en effet quasi-systématiquement un front d'opacité. D'autre part, lorsque des données sont disponibles, les travaux portent essentiellement sur les modalités de l'exposition à l'infox*, sans examiner la transformation des attitudes et des comportements après exposition. Pour ce faire, la mise en place d'un système statistique de suivi combiné des activités en ligne et hors ligne apparaît nécessaire (*cf.* 2.2.2.).

*Tertio*. Ce qui est ébranlé par les comportements de manipulation informationnelle, en ligne comme hors ligne, c'est la confiance en les institutions démocratiques (Lorenz-Spreen *et alii*, 2023). Cette confiance se rapporte aussi bien aux réseaux sociaux qu'aux journalistes et aux personnels politiques : *La vérité est devenue une question politique* (Origgi, 2024). La désinformation et les vérités alternatives sapent la confiance et polarisent les débats. Dans ce contexte global, les technologies numériques représentent un des facteurs à prendre en considération, mais pas le seul, pour proposer des contrefeux à la désinformation et à la perte de confiance dans les démocraties.

*Quarto*. Restant fondamentalement basée, d'une part sur la falsification de contenus, d'autre part sur l'utilisation de faux comptes, l'économie de la désinformation sur les réseaux sociaux et les systèmes de messagerie n'est pas ébranlée par l'IA au niveau de ses soubassements. En revanche, ce que renforce significativement l'IA, c'est la crédibilité du faux, ainsi que l'efficacité et le ciblage des campagnes. Il serait naïf de penser que, puisque le principe moteur de la désinformation reste invariant, le changement d'ordre de grandeur de certains paramètres n'aura pas d'effets notables : filant une métaphore physique, la loi de la gravitation universelle s'applique pareillement sur la Terre et sur la Lune… et pourtant, l'intensité de la pesanteur n'y étant pas du tout la même, les conditions de locomotion sont très différentes sur chacun des deux astres !

## 1.4. LE RÔLE POSITIF DE L'IA DANS LA LUTTE CONTRE LA MÉSINFORMATION

### 1.4.1. L'IA *PHÁRMAKON*

Un article récent paru dans *Science*[24], basé sur une étude expérimentale rigoureuse, suggère que dialoguer avec un LLM* spécialement entraîné à construire des contre-argumentaires réfutant une variété de thèses complotistes, permettrait de réduire d'environ 20 %, et pour une durée d'au moins deux mois, le taux primitif d'adhésion à de telles thèses au sein d'une communauté exposée (Costello *et alii*, 2024). Si ce résultat *in vitro* devait se confirmer *in vivo*, même avec un facteur de réduction des fausses croyances inférieur mais demeurant significatif, cela représenterait une avancée notable dans la lutte contre la mésinformation et démontrerait comment l'IA générative* peut contribuer à cette lutte… et non pas seulement la susciter !

Plus généralement, l'IA*, telle le *phármakon** des Grecs anciens, est à la fois poison et remède, venin et antidote : si elle peut être falsificatrice, elle sait aussi se faire réparatrice. Du côté face de la médaille, l'IA permet par exemple de repérer sur les réseaux sociaux* des « comportements non authentiques coordonnés* », révélant la présence d'un nid d'agents conversationnels artificiels (*chatbots**). Elle permet également d'identifier des contenus de synthèse issus de l'IA générative, quels qu'en soient les supports : sur cet aspect, les initiatives se multiplient en Europe et dans le Monde, émanant de l'industrie comme des institutions publiques.

Dans le cadre de la *Coalition for Content Provenance and Authenticity* (C2PA)[25], groupement professionnel visant à instaurer un système de *Content Credentials*, Adobe a créé un symbole pour inciter au marquage des contenus engendrés par l'IA (*watermarking**), qui a été adopté par les principales plateformes : Google, Microsoft ou Meta. Cependant, il est assez

24 https://www.science.org/doi/10.1126/science.adq1814

25. C2PA développe des standards techniques de certification de la source et de la provenance des contenus médiatiques en ligne. Le projet résulte d'une alliance entre Adobe, Arm, Intel, Microsoft et Truepic.

facile d'échapper à ces démarches volontaires et, plus aisé encore, d'ajouter de faux *watermarks* à d'authentiques images.

Des organisations à but non lucratif, telle *TrueMedia*[26], mettent à la disposition des professionnels et des particuliers des outils de détection. Ces outils utilisent l'apprentissage automatique* et le traitement du langage naturel (TALN*). Ils visent principalement à repérer ce qui serait « authentiquement » d'origine humaine. Cependant, ils ne fournissent que des conclusions probabilistes, laissant ainsi une grande marge d'interprétation.

Le projet européen *vera.ai**, lancé en septembre 2022, a pour ambition de développer de nouveaux outils multilingues de lutte contre la désinformation attachée à tout type de support – texte, photo, vidéo, audio –, en recourant à des algorithmes* et des méthodes d'intelligence artificielle[27]. La plateforme *vera.ai* développe et teste ses outils en collaboration avec des *fact checkers**, dont 130 journalistes d'investigation couvrant 85 pays. L'Agence France Presse (AFP*) est fortement engagée dans cette démarche : elle a créé la structure Médialab* pour participer à des programmes d'innovation et de recherche, notamment le navigateur *InVID-WeVerify*, actuellement utilisé par 90 000 internautes dans 224 pays, dont l'objectif est de vérifier les contenus du Web, singulièrement sur les réseaux sociaux*, par la combinaison d'une approche humaine participative, d'algorithmes* en *open source*, et d'apprentissage machine*.

26. Outre *TrueMedia*, fondée par le docteur Etzioni, on peut également citer *Copyleaks, Content at Scale, OriginalityAI, GPT Zero* ou encore *Crossplag*.
27. Le projet, auquel participent les médias suisses (EBU) et allemands (*Deustche Welle*) et dont le Centre Borelli de l'ENS de Paris-Saclay est également partenaire, est coordonné par l'équipe Mever (*Media Verification*) de l'Institut de recherche en informatique de Thessalonique (ITI-CERTH).

### 1.4.2. La détection de l'infox sur différents supports

Le papier blanc *Generative AI and Disinformation : Recent Advances, Challenges, and Opportunities*, co-écrit par AI4media, AI4TRUST, TITAN et *vera.ai**, identifie avec clarté les défis et les limites de l'IA* au service de la lutte contre la désinformation*, en distinguant les différents supports possibles (Bontcheva, 2024). Qu'en ressort-il principalement ?

En premier lieu, les progrès dans la détection d'images et vidéos synthétiques, bien que non négligeables, restent limités.

## Vidéos et images

La détection d'images et de vidéos synthétiques repose sur l'entrainement de modèles de *deep learning** à l'aide des bases de données publiques les plus populaires, telle *FaceForensics* (base de 500 000 images contenant des visages provenant de 1 004 vidéos, utilisable pour reconnaître des contrefaçons d'images ou de vidéos), ou encore *ForgeryNet* (base de 2,9 millions d'images et 221 247 vidéos, annotées et classifiées pour détecter les contrefaçons faciales).

Les architectures de détection sont des réseaux convolutifs*, tels que *ResNets* (réseaux de neurones résiduels) ou, plus récemment, *VisionTransformers* (décomposition d'une image d'entrée en une série de *patchs* pour améliorer la classification). Un outil de fragmentation des vidéos alimenté par l'IA, très utilisé, permet d'extraire des images clés qui sont ensuite soumises à des moteurs de recherche, afin de vérifier si elles sont déjà connues et indexées.

…

…
Des avancées récentes portent sur la scrutation de vidéos multi-identités, par exemple provenant de téléphones. *vera.ai* s'appuie sur les caractéristiques audio et visuelles d'une personne classée « *person of interest* », pour créer un détecteur de *deep fake**. Si la détection est la plus avancée pour les vidéos et pour le son, des progrès sont également accomplis dans la détection de photos synthétiques.

Cependant, ces différentes percées demeurent embryonnaires. Certains outils, tel *Adobe Firefly*, échappent encore à la détection. Mais, surtout, chaque modèle de détection est spécialisé. Il faudrait donc combiner une batterie de modèles au sein d'un système global, afin de généraliser la détection. Par ailleurs, les faux positifs, c'est-à-dire les vidéos déclarées comme fausses, alors qu'elles sont vraies, discréditent en partie la démarche. Enfin, les défis techniques s'accumulent : obsolescence des outils en raison de l'apparition de nouveaux formats d'encodage ; manque de fiabilité face à la faible qualité et le haut niveau de compression qui caractérisent les contenus sur Internet ; et, tout à l'inverse, exigence d'une grande puissance de calcul afin d'examiner des vidéos longues captées en haute définition.

En deuxième lieu, l'identification de contenus sonores synthétiques pâtit à ce stade d'un défaut de financement et d'une indisponibilité de bases de données d'entrainement, deux facteurs enrayant la dynamique de cet axe d'innovation.

## Sons

La détection de sons synthétiques, à travers les systèmes *Text-to-Speech* (TTS) ou *Voice Conversion* (VC), est récente. Elle n'est en effet devenue une priorité que lorsque sont apparus des premiers discours de synthèse suffisamment sophistiqués pour être jugés problématiques. Afin de satisfaire à cette priorité, il convient urgemment : d'une part, de bâtir un écosystème robuste de recherche et de développement sur ce thème, au financement assuré ; d'autre part, de disposer de bases de données qui soient à la fois conformes à la réglementation et de taille suffisante pour entrainer et tester les technologies de détection.

En outre, plutôt que se concentrer exclusivement sur la reconnaissance de l'origine des contenus sonores, il serait précieux de procéder parallèlement à des recherches linguistiques visant à mieux cerner les mécanismes de manipulation par le langage oral, de mieux appréhender les rites énonciatifs de la falsification.

En troisième lieu, la détection d'écrits de synthèse s'est jusqu'ici focalisée sur des contenus en langue anglaise.

## Écrits

La détection de désinformations engendrées par l'IA sous un format écrit est une voie de recherche très active. Les chercheurs ont montré que les textes de synthèse sont tellement fluides qu'un humain ne peut guère les identifier, et peut même les considérer comme plus fiables que des textes authentiques (Zellers *et alii*, 2020). Les systèmes de détection reposent sur la stylométrie*, ou reconnaissance des spécificités stylistiques d'un auteur. Ils utilisent le *deep learning** et la statistique : hybridées, ces deux techniques sont performantes et savent très bien détecter des formats longs, articles ou blogs artificiels.

…

…
En revanche, la détection automatique appliquée à des textes courts est défaillante. Or elle devrait être une priorité, dans la mesure où les grands modèles de langage (LLMs*) engendrent des messages courts de plus en plus élaborés et crédibles, et où la multiplication des *bots** pilotés par l'IA sur des plateformes telles que X crée un puissant vecteur de désinformation potentielle.

Dans leur majorité, les outils de détection ont été exclusivement entrainés et testés sur des contenus rédigés en anglais. Les textes engendrés par l'IA dans d'autres langues ne sont donc pas détectés, ou piètrement détectés, à l'aide d'une approche uniquement statistique. Il est par conséquent à craindre que les pays dont les langues sont les moins parlées, pour lesquels les bases de données sont les moins fournies, soient les plus ciblés par des manœuvres de désinformation. D'où l'ambition des initiatives européennes VIGILANT et vera.ai,* de constituer un corpus d'entrainement et une batterie de modèles multilingues : la création de bases de données dans toutes les langues parlées en Europe, dont certaines annotées par l'homme afin d'améliorer la performance, est désormais devenue une impérieuse nécessité pour l'entraînement de modèles européens.

Plusieurs points sensibles sont de nature transversale, c'est-à-dire indépendants du support.

– Les systèmes de détection, parce que programmés en *open source*, sont accessibles aux acteurs de la désinformation qui savent comment déjouer ces systèmes, en appliquant des techniques d'obstruction ou de correction humainement indétectables.

– Aussi efficaces que soient les modèles de reconnaissance des « informations » créées par l'intelligence artificielle générative, distinguer le vrai du faux ne peut pour autant être entièrement automatisé et réclame une expertise humaine en dernier ressort. Ce « principe d'un humain dans la boucle » (*human oversight**), au cœur du Règlement européen *AI Act**, soulève à son tour d'autres questions : que doit-on vérifier, à quelles fins, quand et par qui (Enqvist, 2023) ?

– Les outils de détection, parce qu'ils bénéficient des avancées de l'IA, permettent d'analyser un volume toujours plus important de données. Mais ils ne peuvent que très imparfaitement, du moins à ce stade, saisir les subtilités de l'intentionnalité humaine, détecter à coup sûr les éléments stylistiques typiques des *fake news** tels que la polarisation émotionnelle[28], ni déceler les mécanismes de manipulation psychologique qui exploitent les vulnérabilités et les préjugés.

### 1.4.3. Aider les professionnels de l'information

Ayant pour principe d'identifier la source d'une information afin d'en vérifier la fiabilité, les plateformes luttant contre la désinformation* sont démunies face à l'IA générative*, celle-ci brouillant par essence la traçabilité de l'information. À cet égard, l'IA curative* fournit de précieux instruments d'assistance aux vérificateurs de faits* (*fact checkers**) et aux journalistes : elle est tout à la fois utile à identifier les risques désinformation, à améliorer la fiabilité des contenus, et à former les professionnels de l'information.

L'impératif premier est de sensibiliser les acteurs aux risques de désinformation liés à l'IA. Le projet européen AI-CODE[29] développe à cet effet des modules d'éducation à destination des médias, aidant à mieux comprendre les technologies d'IA générative, leurs risques et leurs limites. Ce système interconnecté de différents services d'IA est développé avec des professionnels, afin non seulement d'identifier des contenus falsifiés, mais également de proposer des mesures proactives contre la désinformation. Un assistant basé sur l'IA générative fournit aux journalistes un instrument d'évaluation et de suivi de leur propre création de contenus, afin de mieux leur faire percevoir la menace de désinformation qui résulterait d'un détournement de leur canevas argumentatif et d'une réinterprétation contrefaite de leurs analyses.

28. Cet aspect particulier est examiné au sein du projet allemand DeFakts, piloté par le Centre de recherches sur l'informatique (FZI), en collaboration avec les Ministères fédéraux de l'éducation et de la recherche ; pour déceler la désinformation sur les réseaux sociaux et les groupes de messageries, l'équipe a constitué des bases de données provenant de X et de Telegram.
29. https://aicode-project.eu

Une entreprise plus ambitieuse consiste à s'appuyer sur l'IA pour aider les *fact checkers**, dont le travail est devenu très lourd et complexe. Le projet AI4TRUST[30] a ainsi pour objectif de mettre à leur disposition une plateforme d'outils dont l'originalité réside dans une modélisation sous-jacente des mécanismes sociologiques à l'œuvre dans la désinformation. La plateforme permet d'identifier des schémas comportementaux et relationnels propices à la diffusion de l'infox* et de repérer des groupes à risque favorisant la diffusion de désinformation. Cette dimension sociométrique du projet conduit à une caractérisation des profils des *superspreaders**, comptes diffusant massivement de fausses informations.

S'agissant de la sélection et de la vérification des sources, l'IA peut aider à référencer les contenus de notices encyclopédiques. Wikipedia a pour règle d'asseoir ses affirmations sur des citations sourcées : animés par leur sens des responsabilités, les éditeurs vérifient ces sources, leur fiabilité, comme leur pertinence (*cf.* 1.1.1.), ce qui suppose une compréhension fine du langage. Mise au service des éditeurs de Wikipedia, la plateforme SIDE[31] utilise l'IA, d'une part pour les aider à identifier des citations dont les sources sont contestables ; d'autre part, pour rechercher sur le Web des sources alternatives, plus pertinentes. SIDE a été entrainé sur le corpus Wikipedia en langue anglaise et il a été évalué au regard de sa capacité à proposer des citations alternatives appartenant déjà au corpus Wikipedia dans des articles jugés de haute qualité. Il s'avère que, dans 60 % des cas, les vérificateurs préfèrent les citations suggérées par SIDE (Petroni *et alii*, 2023).

30. Projet financé par l'Union européenne, coordonné par la Fondazione Bruno Kessler et impliquant un consortium de partenaires, dont le CNRS, issus de 11 pays.
31. https://www.nature.com/articles/s42256-023-00726-1

### 1.4.4. Un grand déséquilibre de moyens

Il apparaît qu'en dépit des nombreuses initiatives dont elle est l'objet, l'IA curative* affiche un bilan en demi-teinte : elle progresse avec retard par rapport à l'IA falsificatrice*. Plus précisément, ainsi que nous l'avons souligné *supra*, si les apports de l'intelligence artificielle dans la lutte contre la désinformation demeurent modestes, les freins sont moins technologiques que budgétaires. Très frappant est le déséquilibre des moyens engagés entre : d'un côté les compagnies, qui investissent des milliards dans le développement des LLMs* et du NLP* (*Natural Language Processing*) ; et, de l'autre, l'Europe et les États, qui parviennent péniblement à mobiliser quelques dizaines de millions d'euros pour développer des recherches et des outils de lutte contre la désinformation*.

On peut enfin regretter que les ressources documentaires sur la capacité de l'IA* à détecter des contextes ou des indices de désinformation soient moins abondantes que celles consacrées à la détection des contenus de synthèse. Cette voie semble pourtant particulièrement prometteuse, dans la mesure où l'IA y joue pleinement son rôle d'appui à l'intelligence humaine.

## 1.5. Un secteur sensible : les marchés financiers

Dans cette section, nous illustrons la matière présentée dans les deux précédentes, en focalisant l'attention sur un secteur 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 et, d'autre part, la manipulation de l'information.

Dans le secteur financier, 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, si 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 engendrées artificiellement. Cette falsification manipulatrice a pour objet de créer des mouvements favorables à leurs auteurs au détriment des autres acteurs du marché, 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 mobilisés par les institutions.

### 1.5.1. Le trading algorithmique

Le trading algorithmique*, ou trading automatisé, n'est pas une nouveauté sur les marchés financiers. Remontant aux années 1980, il consiste à utiliser des plateformes électroniques pour l'émission des ordres de bourse, laissant à un algorithme* – le plus souvent sans aucune intervention humaine –, le soin de décider des différents paramètres de ces ordres, tels que l'instant d'ouverture ou de clôture, le prix, le volume. Cette forme de trading est pratiquée sur deux échelles de temps très différentes : tandis que les intermédiaires, ou courtiers (*brokers**), y recourent avec un délai d'opération allant d'un à deux jours, les traders* et les opérateurs pour compte propre la pratiquent à très haute fréquence (THF*), pouvant aller jusqu'à exécuter une opération par microseconde !

Les algorithmes de trading sont entraînés sur des volumes massifs de données de marché, actuelles et historiques, au moyen de méthodes d'apprentissage statistique*. Ils sont censés prendre des décisions et exécuter des transactions, non seulement beaucoup plus rapidement, mais encore avec beaucoup plus de précision que des humains. Typiquement, ils identifient des corrélations entre certains titres et, s'ils observent des déviations inattendues, anticipent un ré-appariement et arbitrent sur cette base.

Même si l'objectif du trading automatisé est de renforcer l'efficacité et la sécurité, la réduction de l'intervention humaine ne peut être considérée comme un pur gain d'efficacité, car elle majore les risques de déstabilisation en cas de bugs ou de pannes informatiques. Ainsi, le 1er août 2016, la société de trading haute fréquence Knight Capitol a-t-elle perdu 440 millions $ en 40 minutes ! Lors d'une mise à jour de son système de négociation, ce qui était en réalité un test a été confondu par la machine avec un fonctionnement normal ! La valeur de l'action a plongé, nécessitant une recapitalisation en urgence, et l'entreprise a frôlé la faillite[32].

En première analyse, l'anticipation des mouvements de marché afin d'effectuer des transactions à grande vitesse semblerait devoir créer une volatilité accrue de certaines valeurs et provoquer des fluctuations incontrôlées. Or ce n'est pas un résultat acquis. L'impact du trading algorithmique sur la volatilité des actifs est encore mal connu. Certains travaux académiques confortent l'idée intuitive d'une volatilité accrue, tandis que d'autres tablent sur l'effet inverse, notamment justifié par l'argument que, sauf accident, une machine ne « s'affole pas », contrairement à un être humain !

Le 6 mai 2010, dans un contexte de tension sur la dette grecque, l'indice Dow Jones* s'est effondré pendant 36 minutes. Ce phénomène, désormais connu sous le nom de *flash crash*[33], fut consécutif à l'ordre de vente massif de 75 000 titres « E-mini ». Il n'est donc pas imputable au trading algorithmique, comme on a pu le penser, même si celui-ci a pu ensuite contribuer à amplifier la chute de l'indice.

32. https://fr.m.wikipedia.org/wiki/Knight_Capital_Group
33. https://fr.m.wikipedia.org/wiki/Flash_Crash_de_2010

### 1.5.2. Les promesses de l'IA financière

Les opportunités offertes par l'IA* dans le secteur financier sont variées et séduisantes. Tandis que les services rendus aux investisseurs peuvent être enrichis et personnalisés, certaines tâches fastidieuses peuvent être assistées ou automatisées. Plusieurs innovations sont d'ores et déjà expérimentées.

En exploitant les données personnelles et comportementales des utilisateurs, dans les limites imposées par la CNIL* et l'*AI Act*, l'IA peut formuler des recommandations financières ciblées, portant sur les stratégies d'investissement, la budgétisation de projets ou la planification financière. Un LLM peut en particulier produire une analyse financière à partir de données issues de la presse spécialisée. Néanmoins, si l'IA traite et analyse à cet effet de grandes quantités de données financières brutes, permettant en théorie d'obtenir une information transformée de qualité, en pratique la présence de biais algorithmiques* gâte en partie le tableau, pouvant induire des décisions mal fondées. Autre bémol, l'utilisation de l'IA crée une asymétrie entre les acteurs : en effet, seuls ceux ayant accès à ces fonctionnalités avancées peuvent en tirer bénéfice sur les marchés, ce qui interpelle l'éthique.

Côté *back office*, l'IA permet d'automatiser des tâches répétitives et routinières, telle la fusion de documents financiers, ce qui accélère le traitement et l'analyse des données, libérant un temps que les professionnels du secteur peuvent occuper à des activités plus stratégiques.

Dans la même veine, en analysant et en combinant les dispositions réglementaires encadrant le secteur, qui forment un corpus extrêmement complexe, l'IA peut aider les institutions financières à se conformer en toutes circonstances aux réglementations en vigueur.

### 1.5.3. La manipulation artificielle de l'information sur les marchés financiers

La fiction romanesque a fourni un exemple de manipulation devenu emblématique. En 1838, l'Edmond Dantès d'Alexandre Dumas, *alias* Comte de Monte Cristo, soudoie l'opérateur de la station de Montlhéry du télégraphe Chappe*, afin de diffuser une fausse nouvelle sur la situation politique en

Espagne. Ceci aura pour conséquence la vente en masse des coupons de l'emprunt espagnol et entraînera la ruine du banquier Danglars, dont Dantès entend se venger.

Plus proche de nous et cette fois dans le monde réel, le 22 novembre 2016, le cours de bourse du groupe Vinci, pilier du CAC 40*, a dévissé de 19 % à partir de 16 h. En quelques minutes seulement la valeur de l'action est passée de 61,81 € à 49,93 €, avant suspension de sa cotation à 17 h. À l'origine de cette dégringolade, la publication de fausses informations alarmantes concernant d'importantes malversations sur les comptes du Groupe Vinci ainsi que le prétendu licenciement de son directeur financier[34].

Un pas de plus est potentiellement franchi avec L'IA générative*. Celle-ci peut en effet être utilisée pour manipuler les marchés, en créant de fausses tendances ou en diffusant des informations trompeuses sous la forme d'articles, de rapports financiers ou de communiqués de presse, qui semblent authentiques mais sont en fait artificiels. Les technologies de *deep fake** permettent notamment de créer des vidéos et des enregistrements audio falsifiés de dirigeants d'entreprise ou de personnalités influentes, mettant dans leur bouche des annonces ou des déclarations fictives qui affectent les perceptions des investisseurs, influencent leurs décisions et provoquent des mouvements de marché intempestifs.

La manipulation financière par *deep fake* est d'ores et déjà à l'œuvre. Le 22 mai 2023, une pseudo-photographie engendrée par l'IA a été partagée par des comptes financiers, ainsi que par des médias russes, dans une escalade de désinformation[35]. L'image, qui montre une fausse explosion aux abords du Pentagone, a entraîné une baisse momentanée de 0,3 % des cours boursiers aux États-Unis. Selon l'agence Bloomberg, elle constitue le « premier cas d'image artificielle ayant influé sur la Bourse ».

34. https://www.lemonde.fr/economie-francaise/article/2016/11/23/comment-le-groupe-vinci-victime-d-un-hoax-a-chute-en-bourse_5036269_1656968.html

35. https://www.lemonde.fr/les-decodeurs/article/2023/05/24/la-fausse-image-d-une-explosion-au-pentagone-fait-brievement-douter-les-marches_6174669_4355770.html

L'IA générative permet aux manipulateurs de décortiquer le « sentiment de marché* », résultant des expressions postées sur les réseaux sociaux* et très prisé par les investisseurs dans leurs prises de décision (Gu, 2020). Ces manipulateurs sont ainsi en mesure d'engendrer des contenus artificiels destinés à amplifier des émotions comme la peur ou l'euphorie, en vue d'entraîner des réactions exacerbées de la part des investisseurs, dans telle ou telle direction souhaitée. Plus directement, l'IA peut être employée afin de simuler une demande pour certaines actions ou produits financiers, créant ainsi une illusion de popularité ou de rareté, qui incite les investisseurs à acheter ou vendre sur la base de ces manœuvres déguisées. Par ailleurs, des *bots** alimentés par l'IA générative peuvent diffuser massivement des fausses nouvelles financières sur diverses plateformes en ligne, ce qui augmente leur visibilité et leur crédibilité apparente.

### 1.5.4. Les attaques et les outils d'IA au service de l'intégrité des marchés financiers

Outre les tentatives de manipulation de l'information, les attaques numériques sur les institutions financières ne sont pas rares. Plusieurs ont été ciblées par des cyberattaques visant à s'emparer de données sensibles pour perturber les opérations financières, en utilisant des IA avancées afin de contourner les systèmes de sécurité. Notamment, dans la nuit du 6 au 7 octobre 2022, la plateforme mondiale *Binance* d'échange de cryptomonnaies s'est fait dérober 40 millions de bitcoins, équivalant à 100 millions de dollars, le *hacker** ayant utilisé une faille dans une interface entre deux monnaies[36].

Encore dans le domaine des cryptomonnaies*, particulièrement vulnérable, des groupes organisés utilisent des IA pour orchestrer des mouvements *pump and dump** : en analysant les tendances du marché et en coordonnant des achats massifs à l'aide de *bots*, ils attirent des investisseurs crédules et font monter artificiellement les prix avant de vendre très rapidement et ainsi provoquer une chute du cours infligeant à ces investisseurs de lourdes pertes.

36. https://fr.m.wikipedia.org/wiki/Binance

Les régulateurs – en France l'AMF* –, ainsi que les institutions financières, œuvrent activement à renforcer leurs défenses contre ce type de menaces et contre les tentatives de manipulation. Ils ont recours à des outils spécialisés de surveillance et d'analyse des transactions et des communications.

- **Analyse des sentiments.** Des outils comme *Sentifi* et *StockTwits* scrutent les sentiments de marché sur les réseaux sociaux et les forums de discussion, pour y détecter des anomalies et des tendances inhabituelles.

- **Analyse des marchés et des transactions.** Des plateformes comme *ProRealTime* et *MetaTrader* offrent des fonctionnalités avancées pour analyser les mouvements de marché et identifier des phénomènes anormaux. *NASDAQ SMARTS* et *Aquis Exchange* surveillent les transactions en temps réel, pour repérer des opérations de trading anormales.

- **Analyse des informations commerciales sur les entreprises.** Des services comme *Meltwater* et *Factiva* examinent les informations commerciales issues des médias traditionnels et en ligne et permettent d'y repérer des éléments falsifiés.

- **Détection de fraude.** Des applications comme *Actimize* et *SAS Fraud Management* sont utiles à détecter des activités frauduleuses et des manipulations de marché. Entraînées par apprentissage machine à partir de données de transaction, ces outils peuvent reconnaître des configurations caractéristiques d'activités frauduleuses, aidant ainsi à les identifier et à les prévenir, avec pour résultat escompté une amélioration de la sécurité des marchés et une réduction des pertes financières.

### 1.5.5. Les récentes initiatives de la Commission européenne

Après la publication de l'*AI Act**, et en vue d'accéder à une meilleure compréhension des opportunités et des risques liés à l'utilisation de l'IA* dans le secteur financier, la Commission européenne souhaite aujourd'hui recueillir des avis et des retours d'expérience de la part des acteurs des marchés financiers qui déploient ou utilisent des systèmes d'intelligence artificielle.

À cet effet, deux initiatives ont été lancées au printemps 2024 :

– la première consiste en un appel à manifestation d'intérêts auprès des acteurs des services financiers pour participer, au cours du dernier trimestre 2024, à des ateliers pratiques dédiés aux développements récents de l'IA dans leur secteur ;

– la seconde est une consultation publique sur l'utilisation de l'IA dans l'écosystème financier, s'adressant plus largement à tous les acteurs de cet écosystème – institutions financières, entreprises, associations de consommateurs – et destinée à rassembler des informations sur des cas concrets d'utilisation de l'IA.

En France, l'*Autorité des marchés financiers* (AMF*) participe activement à cette double démarche et elle encourage les acteurs qu'elle régule à faire de même[37].

37. https://www.amf-france.org/fr/actualites-publications/actualites/intelligence-artificielle-lamf-encourage-les-acteurs-des-marches-prendre-part-aux-travaux-inities

## 1.6. La réglementation de la désinformation et de l'IA générative

Cette section procède en quatre temps. Dans le premier, nous soulignons les contraintes constitutionnelles qui limitent la possibilité d'interdire les fausses informations et nous décrivons comment le cadre réglementaire s'est adapté pour se concentrer sur les acteurs, les méthodes et les effets des campagnes de désinformation*, plutôt que sur l'information en soi. Dans un deuxième temps, nous présentons le cadre réglementaire de la désinformation véhiculée par les grandes plateformes numériques*, notamment la *loi française infox** de 2018 et le DSA* européen de 2022. Dans un troisième temps, nous précisons la façon dont le risque spécifique de l'IA générative* a été intégré dans le cadre réglementaire. Dans un quatrième et dernier temps, nous identifions deux pistes réglementaires pour mieux tenir compte des spécificités de l'IA générative dans les campagnes de désinformation.

### 1.6.1. Réguler la désinformation, sans attenter à la liberté d'expression

La liberté d'expression garantit à chacun le droit d'exprimer des idées qui heurtent, choquent ou inquiètent la population ou les institutions[38]. Les limitations à la liberté d'expression doivent être clairement identifiées par la loi et respecter les conditions de nécessité et de proportionnalité.

> Article 11 de la Déclaration des Droits de l'Homme et du Citoyen de 1789 :
>
> La libre communication des pensées et des opinions est un des droits les plus précieux de l'homme : tout citoyen peut donc parler, écrire, imprimer librement, sauf à répondre de l'abus de cette liberté dans les cas déterminés par la loi.

Toute réglementation visant à lutter contre la désinformation doit cheminer sur une ligne de crête entre le principe constitutionnel de la liberté

38. CEDH Handyside c. Royaume-Uni, 7 décembre 1976, point 49.

d'expression, d'une part, et la protection des personnes et des institutions démocratiques, d'autre part. Cette mise en balance est particulièrement délicate, parce que la liberté expression fait partie du socle démocratique que l'on souhaite protéger. En réduisant la liberté d'expression au nom de la lutte contre la désinformation, on risque donc d'affaiblir l'objet même que l'on souhaite sauvegarder.

Les fausses informations, ainsi que des thèses scientifiques très contestables, ne sont pas illégales en soi et elles sont souvent tolérées au nom de la liberté d'expression[39]. Tout dépend de la nature de la personne qui publie l'information, de ses intentions, de ses méthodes, et des effets sur l'ordre public, sur la santé publique, ou sur les institutions démocratiques (Türk, 2023). Une théorie complotiste pourrait être tolérée venant d'un citoyen ordinaire, mais elle deviendrait condamnable si elle était diffusée par un réseau organisé dans le but de manipuler l'opinion. La réglementation doit nécessairement cibler les acteurs, les méthodes et les effets d'une campagne de désinformation, plutôt que de se concentrer uniquement sur le caractère des informations.

La liberté d'expression n'est pas absolue et de nombreuses lois l'encadrent. Par exemple, une fausse information publiée en utilisant le logo de TF1 serait illégale, non pas nécessairement parce que l'information serait fausse, mais en raison de l'usurpation de la marque de TF1. Une fausse information dans le commerce serait interdite par le code du commerce (art. 111-7), une fausse information en lien avec une entreprise cotée en bourse serait sanctionnée par le code monétaire et financier (art. L 621-15).

La publication d'un vidéo-montage d'une personne pourrait être condamnable pour violation du droit à l'image de cette personne, ou condamnable parce que la vidéo est diffamatoire. Le code pénal français vise particulièrement l'utilisation des images d'autrui.

39. CEDH, Salov c. Ukraine, 6 septembre 2005, point 113. CEDH, Hertel c., Suisse, 25 août 1998, point 50.

Code pénal Article 226-8

Est puni d'un an d'emprisonnement et de 15 000 euros d'amende le fait de porter à la connaissance du public ou d'un tiers, par quelque voie que ce soit, tout montage réalisé avec les paroles ou l'image d'une personne sans son consentement, s'il n'apparaît pas à l'évidence qu'il s'agit d'un montage ou s'il n'en est pas expressément fait mention.

Le code électoral interdit, quant à lui, l'utilisation de fausses nouvelles, bruits, calomnies ou autres manœuvres frauduleuses pour détourner des suffrages ou inciter un ou plusieurs électeurs à s'abstenir de voter.

Code électoral Article L97

Ceux qui, à l'aide de fausses nouvelles, bruits calomnieux ou autres manœuvres frauduleuses, auront surpris ou détourné des suffrages, déterminé un ou plusieurs électeurs à s'abstenir de voter, seront punis d'un emprisonnement d'un an et d'une amende de 15 000 euros.

Mais la liberté d'expression laisse une certaine latitude à la publication de contenus inexacts. En particulier, la loi du 29 juillet 1881 établit l'exception de bonne foi.

Loi du 29 juillet 1881. Article 27

La publication, la diffusion ou la reproduction, par quelque moyen que ce soit, de nouvelles fausses, de pièces fabriquées, falsifiées ou mensongèrement attribuées à des tiers lorsque, faite de mauvaise foi, elle aura troublé la paix publique, ou aura été susceptible de la troubler, sera punie d'une amende de 45 000 euros.

Les mêmes faits seront punis de 135 000 euros d'amende, lorsque la publication, la diffusion ou la reproduction faite de mauvaise foi sera de nature à ébranler la discipline ou le moral des armées ou à entraver l'effort de guerre de la Nation.

### 1.6.2. La régulation de la désinformation passe par la régulation des plateformes

La loi française du 22 décembre 2018, dite *loi infox**, oblige les plateformes à lutter contre la désinformation*. Cette loi, relative à la lutte contre la manipulation de l'information, constitue une première tentative de régulation de la désinformation visant spécifiquement les réseaux sociaux*. Avant d'être partiellement remplacée, elle imposait aux grandes plateformes l'obligation de mettre en œuvre des mesures en vue de lutter contre la diffusion de fausses informations susceptibles de troubler l'ordre public ou d'altérer la sincérité d'un scrutin. Ces plateformes devaient mettre en place un dispositif permettant aux utilisateurs de signaler de telles informations. La loi énumérait également une liste d'autres mesures que les grandes plateformes devraient mettre en œuvre, mais elle s'abstenait d'en imposer certaines en particulier. Cette partie de la loi de 2018 a été abrogée par la loi du 21 mai 2024 car, dorénavant, les obligations des plateformes sont régies par le DSA* (*Digital Services Acts*).

La partie résiduelle de la *loi infox*, non abrogée par la loi de 2024, prévoit une procédure judiciaire accélérée pour bloquer la diffusion d'informations inexactes ou trompeuses de nature à altérer la sincérité du scrutin, lorsque cette diffusion est faite de manière délibérée, artificielle ou automatisée, et massive. La notion de fausses informations est liée ici à deux conditions cumulatives : d'une part les fausses informations doivent être de nature à altérer la sincérité du scrutin ; d'autre part la diffusion doit être délibérée, artificielle ou automatisée, et massive. Ainsi la loi se concentre à la fois sur la motivation, à savoir altérer la sincérité d'un scrutin, et sur les moyens automatisés et massifs.

Compte tenu de l'agilité des campagnes de désinformation et de la relative lenteur des procédures judiciaires, la solution la plus efficace pour lutter contre la désinformation consiste à s'appuyer sur les moyens techniques et humains des plateformes. Si la *loi infox* de 2018 a énoncé le principe d'une collaboration des plateformes dans la lutte contre la désinformation, elle n'a pas assorti cette obligation de sanctions ou de mesures contraignantes précises.

Paru en octobre 2022, le règlement européen sur les services numériques ou *Digital Services Acts* (DSA), renforce la régulation des très grandes plateformes, en consacrant le principe d'une coopération de leur part et en prévoyant des sanctions importantes en cas de violation. Concrètement, les plateformes les plus importantes devront effectuer une analyse de risques, notamment de désinformation, et mettre en œuvre des mesures techniques et humaines pour réduire ces risques.

> DSA Article 35
>
> Les fournisseurs de très grandes plateformes en ligne et de très grands moteurs de recherche en ligne mettent en place des mesures d'atténuation raisonnables, proportionnées et efficaces, adaptées aux risques systémiques spécifiques recensés…

Le *Comité européen des services numériques* (CESN*)[40] et la Commission européenne passent en revue les mesures proposées par les plateformes en évaluant leur efficacité. Plus précisément, les grandes plateformes doivent rendre leurs données accessibles aux autorités de contrôle et disposer d'un organe interne de contrôle et de conformité. La Commission peut sanctionner les plateformes si elle estime que les mesures prises ne sont pas suffisantes. Elle est habilitée à publier des lignes directrices sur les meilleures pratiques et le DSA prévoit l'utilisation de « codes de conduites ».

Ainsi, le DSA délègue aux très grandes plateformes la responsabilité opérationnelle d'établir le diagnostic de risques et de mettre en place des mesures de correction, sous l'œil attentif de la Commission et du CESN. Les mesures proposées sont incorporées dans les conditions d'utilisation des plateformes et appliquées par celles-ci conformément à leur politique interne de modération de contenus.

40. Groupement européen des organes nationaux en charge de la régulation des plateformes numériques, en France l'Arcom*.

Une telle délégation aux acteurs privés est utilisée dans d'autres domaines, comme la lutte contre le blanchiment des capitaux. Elle permet à chaque acteur de construire les moyens les plus appropriés pour lutter contre une activité illégale, en tenant compte des spécificités de son service. Évidemment, la délégation nécessite un contrôle par une autorité de régulation, ainsi que la menace de sanctions dissuasives, car l'intérêt économique de la plateforme n'est généralement pas aligné avec l'intérêt public. Les activités liées à la diffusion de fausses informations peuvent en effet procurer des recettes publicitaires importantes pour les plateformes (*cf.* 1.1.3.). En l'absence d'une régulation, ces dernières auraient tendance à se montrer laxistes dans le contrôle de ces activités.

Comme le blanchiment, la désinformation peut être difficile à détecter car une information isolée est rarement problématique en soi. Elle ne le devient que si l'on tient compte du contexte. L'analyse doit donc nécessairement considérer l'ensemble des activités liées directement ou indirectement à la mise en ligne du contenu, ou au compte de l'utilisateur. Idéalement, le partage d'informations entre plateformes permettrait d'identifier des comportements coordonnés entre plusieurs plateformes et identifier des réseaux d'acteurs malveillants.

### 1.6.3. Un code de bonnes pratiques pour traiter la mésinformation

Le partage d'informations entre plateformes numériques* est prévu dans un code de bonne pratiques spécialement dédié à la lutte contre la désinformation*, mis à jour en 2022. Parmi les engagements des acteurs signataires de ce code figure l'obligation de partager des informations entre plateformes.

*Commitment 16*

*Relevant Signatories commit to operate channels of exchange between their relevant teams in order to proactively share information about cross-platform influence operations, foreign interference in information space and relevant incidents that emerge on their respective services, with the aim of preventing dissemination and resurgence on other services, in full compliance with privacy legislation and with due consideration for security and human rights risks.*

*In order to satisfy Commitment 16 :*

*Measure16.1. Relevant Signatories will share relevant information about cross-platform information manipulation, foreign interference in information space and incidents that emerge on their respective services for instance via a dedicated sub-group of the permanent Task-force or via existing fora for exchanging such information.*

Selon la Commission européenne, le code de 2022 est voué à devenir un code de conduite officiel au titre de l'article 45 du DSA.

Le premier moyen pour lutter contre la désinformation mis en avant dans le code de bonnes pratiques est la mise en place d'un réseau de vérificateurs de confiance*, qui effectuera le travail d'analyse des contenus afin de discriminer une campagne de désinformation d'une simple opinion sur un sujet de débat public.

Le second moyen préconisé consiste à analyser les moyens de diffusion et le comportement de différents comptes pour détecter des signes distinctifs d'une campagne organisée. Ce moyen soulève moins de difficultés juridiques, puisque l'utilisation de faux comptes et de robots est interdite par les conditions d'utilisation des platesformes, en tant que « comportements inauthentiques coordonnés* » (*Coordinated Inauthentic Behavior* ou CIB*). Quand un comportement artificiel est détecté, la plateforme est fondée à suspendre l'activité immédiatement, sans entrer dans une discussion sur le caractère vrai ou faux des contenus diffusés, ainsi que sur les motivations des personnes qui diffusent ces contenus. Cette technique évite aussi la question,

soulevée par certaines plateformes, des biais de certaines organisations de vérificateurs de confiance. La traque de faux comptes fait partie de la lutte contre les comportements inauthentiques coordonnés*. Selon les rapports de Meta, Facebook suspend l'activité de 10 millions de faux comptes par jour, tous détectés par l'utilisation d'algorithmes*.

Grâce à la détection de comportements inauthentiques coordonnés, le contenu des messages est ravalé au rang d'un élément presque secondaire. Le point essentiel devient la détection – par des moyens automatiques – de comportements anormaux sur les réseaux sociaux, correspondant à une probable campagne de désinformation. Le contenu véhiculé vient ensuite en appui de cette hypothèse, mais il n'est plus l'élément déclencheur.

#### 1.6.4. La lutte contre les ingérences étrangères

L'approche précédente est appliquée par les services de renseignement des États. En France, le service Viginum*, qui fait partie du Secrétariat national de la défense et de la sécurité nationale (SGDSN*), est spécifiquement chargé d'une mission de détection de comportements anormaux liés à des ingérences étrangères, que celles-ci proviennent d'États ou d'organisations.

> L'article R.* 1132-3 9° du Code de la Défense
>
> 9° En liaison avec les départements ministériels concernés, il identifie les opérations impliquant, de manière directe ou indirecte, un État étranger ou une entité non étatique étrangère, et visant à la diffusion artificielle ou automatisée, massive et délibérée, par le biais d'un service de communication au public en ligne, d'allégations ou imputations de faits manifestement inexactes ou trompeuses, de nature à porter atteinte aux intérêts fondamentaux de la Nation. Il anime et coordonne les travaux interministériels en matière de protection contre ces opérations.

La mission de Viginum fait ressortir quatre éléments caractérisant une campagne de désinformation. Premièrement, les allégations ou imputations de faits doivent être manifestement inexactes. Le texte mentionne les faits, non des opinions. L'inexactitude des faits doit être manifeste, à savoir vérifiable par rapport à des références objectives. Deuxièmement, Viginum n'est compétent que si un État étranger, ou une entité non étatique étrangère, est impliqué. Troisièmement, la diffusion de ces contenus inexacts doit être artificielle ou automatisée, massive, délibérée, et véhiculée par un service de communication au public en ligne. Quatrièmement, la diffusion doit représenter une menace pour les intérêts fondamentaux de la nation.

### 1.6.5. Les inquiétudes liées à l'IA générative conduisent à de nouvelles obligations réglementaires.

L'arrivée de l'intelligence artificielle générative*, en 2022 et 2023, a provoqué de nouvelles inquiétudes quant aux risques de désinformation*. Ces nouveaux outils d'IA permettent de fabriquer d'inauthentiques images, vidéos, sons et textes avec une facilité déconcertante. Même si cette technologie ne change ni la motivation des acteurs malveillants, ni leurs techniques de diffusion par l'utilisation de faux comptes, l'IA générative facilite la création de contenus faux mais vraisemblables. De plus, ces outils permettent de créer des *chatbots** interactifs, qui imitent un comportement humain. Des outils de manipulation à grande échelle sont désormais à la portée de tous.

Ces nouvelles préoccupations autour de l'IA générative et la désinformation sont désormais traduites dans le Règlement européen *AI Act**, publié le 12 juillet 2024. Les législateurs européens y ont en effet introduit de nouvelles dispositions qui obligent les fournisseurs et déployeurs de systèmes d'IA à signaler l'existence de contenus de synthèse, c'est-à-dire fabriqués par l'intelligence artificielle. Ces obligations pèsent autant sur les fournisseurs de systèmes d'IA, qui devront y insérer un procédé de marquage, que sur les déployeurs, qui devront signaler la génération de *deep fakes** ou de textes engendrés par l'IA destinés à informer le public sur des questions d'intérêt public. Ces dispositions complètent celles du DSA*, qui comprenait déjà l'obligation pour les plateformes de signaler de faux contenus.

DSA article 35-1-k

Ces mesures peuvent inclure, le cas échéant : … le recours à un marquage bien visible pour garantir qu'un élément d'information, qu'il s'agisse d'une image, d'un contenu audio ou vidéo créé ou manipulé, qui ressemble nettement à des personnes, à des objets, à des lieux ou à d'autres entités ou évènements réels, et apparaît à tort aux yeux d'une personne comme authentique ou digne de foi, est reconnaissable lorsqu'il est présenté sur leurs interfaces en ligne ; et, en complément, la mise à disposition d'une fonctionnalité facile d'utilisation permettant aux destinataires du service de signaler ce type d'information.

*AI Act* article 50-4

Les déployeurs d'un système d'IA qui engendre ou manipule des images ou des contenus audio ou vidéo constituant un hypertrucage indiquent que ces contenus ont été engendrés ou manipulés par une IA. Les déployeurs d'un système d'IA qui engendre ou manipule des textes publiés dans le but d'informer le public sur des questions d'intérêt public indiquent que ces textes ont été engendrés ou manipulés par une IA.

Par ailleurs, en 2024, un accord est conclu et des lignes directrices sont établies, visant l'IA et les risques liés à la désinformation.

L'accord de la Conférence de Sécurité de Munich, signé en février 2024 par les plus grandes plateformes et fournisseurs de systèmes d'IA, contient huit engagements visant particulièrement l'utilisation d'IA dans le cadre des élections de 2024. Ces engagements n'ont pas de caractère contraignant car l'accord ne prévoit aucun mécanisme de sanctions. Néanmoins, les mesures mentionnées dans l'accord constituent une liste des meilleurs pratiques qui pourraient indirectement devenir contraignantes, au moins pour les grandes plateformes, par le biais du DSA. Parmi les engagements figurent :

– le développement d'outils d'identification de contenus créés par IA, et d'outils pour sourcer la provenance de contenus ;

– la mise en place d'outils techniques pour détecter la diffusion de contenus engendrés par IA dans le cadre d'élections ;

– le partage d'informations sur les attaques.

Le 26 mars 2024, la Commission européenne a émis des lignes directrices sur les mesures que devraient prendre les plus grosses plateformes, au titre du DSA, pour prévenir les risques liés aux élections européennes. Ces lignes directrices découlent directement du DSA et précisent les obligations des grandes plateformes en matière de risques d'interférence électorale. La Commission y tient notamment compte des meilleures pratiques et des obligations prévues dans l'*AI Act**, anticipant leur mise en application. Les lignes directrices de la Commission prennent également en considération les engagements des opérateurs au titre du code de conduite mis à jour en 2022 (*cf.* 1.6.3.), ainsi que les différentes recommandations émises sur les risques d'ingérence étrangère dans le contexte des élections.

Les lignes directrices contiennent des mesures spécifiques liées à l'IA générative ; elles constituent, à ce jour, un résumé des meilleurs pratiques des grandes plateformes en matière d'IA générative et la lutte contre la désinformation. S'agissant des risques de l'IA générative, les mesures proposées par la Commission comprennent notamment :

(a) en ce qui concerne les plateformes dont les services peuvent permettre la création de contenus par IA :

– utiliser des outils conformes à l'état de l'art pour détecter et procéder au marquage de contenus créés par l'IA ;

39-a

*Ensure that generative AI content, and other types of synthetic and manipulated media, is detectable – notably by using sufficiently reliable, interoperable, effective and robust techniques and methods, such as watermarks, metadata identifications, cryptographic methods for proving provenance and authenticity of content, logging methods, fingerprints or other techniques, as may be appropriate, taking into account existing standards.*

- garantir que les informations engendrées par les systèmes d'IA s'appuient le plus possible sur des sources fiables ;

- prévenir les utilisateurs des risques et les encourager à consulter des sources d'information officielles ;

- conduire des « tests d'attaques » (*red teaming**), pour identifier des failles dans le système, avant de le rendre public ;

- établir des critères de performance relatifs aux réponses fournies par les IA génératives dans un contexte électoral ;

- intégrer des outils de détection et de filtrage, pour limiter une utilisation de *prompts** conduisant à violer les conditions d'utilisation des plateformes dans un contexte électoral ;

- s'agissant des résultats textuels engendrés par IA, indiquer lorsque c'est possible la source concrète des informations utilisées par le modèle pour produire les réponses.

(b) en ce qui concerne les plateformes dont les services peuvent permettre la diffusion de contenus par IA :

- s'assurer que les conditions générales d'utilisation et les outils de modération des très grandes plateformes, VLOPs*[41] et VLOSEs*[42], diminuent de manière importante la diffusion et l'impact de contenus artificiels affectant un processus électoral ;

41. *Very Large Online Platforms.*
42. *Very Large Online Search Engines.*

40-a

*(i) The Commission recommends that providers of VLOPs and VLOSEs provide clear public information on which internal processes and mitigation measures, such as labelling, marking, demoting or removing, are in place to enforce these policies ;*

*(ii) The Commission recommends that providers of VLOPs and VLOSEs cooperate and share information about such deceptive content with fact checkers* to ensure that the risk of amplification in other platforms is minimized.*

– apposer des étiquettes ou autres marquages clairement visibles sur des contenus *deep fake**, audio, vidéo ou images ;

– exiger des annonceurs l'utilisation du marquage pour identifier des contenus créés par l'IA ;

– mettre en place des systèmes algorithmiques de modération de contenus conformes à l'état de l'art pour détecter des contenus issus d'une IA ;

40-d

*In this context, providers of VLOPs and VLOSEs should cooperate with providers of generative AI systems and follow leading state of the art measures to ensure that such watermarks, metadata identifications, cryptographic methods for proving provenance and authenticity of content, logging methods, fingerprints or other techniques are detected in a reliable and effective manner ; they are also recommended to support new technology innovations to improve the effectiveness and interoperability of such tools.*

– mettre en place des campagnes d'éducation des utilisateurs sur les risques liés à la manipulation des contenus par une IA.

La Commission indique que cette liste de mesures constituera le point de référence, s'agissant des obligations des grandes plateformes – au titre de l'article 35 du DSA –pour la mise en place des « mesures d'atténuation raisonnables, proportionnées et efficaces, adaptées aux risques systémiques spécifiques ».

Parallèlement publiées en mars 2024, les préconisations de l'Arcom* relatives à la lutte contre la manipulation de l'information sur les plateformes en ligne en vue des élections au parlement européen soulignent notamment le rôle des vérificateurs de faits (*fact checkers**) participant à la lutte contre la désinformation et les ingérences étrangères, ainsi que l'utilisation d'outils pour détecter les opérations de manipulation non authentiques et coordonnées.

#### 1.6.6. Le dispositif réglementaire actuel pourrait utilement être renforcé sur deux points

*Primo*. L'abrogation de l'article 11 de la loi du 22 décembre 2018 ouvre une phase de transition, au cours de laquelle l'Arcom* doit assumer de nouvelles responsabilités au titre du DSA, en lien avec la Commission européenne. Nous proposons que l'Arcom*, en lien avec Viginum et l'Anssi*, soit force de proposition pour créer un nouvel outil qui mesure le degré d'artificialité d'un contenu, que ce soit le degré d'utilisation de l'IA générative dans la création de ce contenu, ou le degré d'utilisation de réseaux de faux comptes dans sa diffusion. L'affichage de ce nouveau *score d'artificialité** pourrait être rendu obligatoire par la Commission européenne pour les très grandes plateformes (*cf.* 2.2.5.).

*Secundo*. Les attaques de désinformation d'origine étrangère sont identifiées par Viginum, mais elles ne sont pas encore spécifiquement visées par le code pénal. Nous proposons donc que soit introduit dans ce code un nouvel article pour combler ce manque (*cf.* 2.2.6.).

Chapitre 2

# LIGNES D'ACTION

Dans ce second chapitre du rapport, nous traçons quelques pistes pour réduire les risques que fait porter l'IA générative* en termes de développement des pathologies de l'information. Nous procédons en deux temps. Tout d'abord, sur la base des considérations présentées dans le premier chapitre « Analyse », nous identifions et balisons quatre grands champs d'initiatives, en mentionnant brièvement pour chacun d'eux quelques enjeux qui nous semblent clés, ainsi que plusieurs directions déjà prises ou à prendre (*cf.* 2.1.). Puis, en les situant sur la carte des actions souhaitables, nous formulons six propositions spécifiques, ayant fait l'objet de réflexions originales de la part des membres du groupe de travail (*cf.* 2.2.).

## 2.1. Quatre champs d'initiatives

Les développements du chapitre I, ainsi que les auditions menées par le groupe de travail (*cf.* Annexe II), font ressortir quatre grands champs se prêtant à des initiatives publiques ou privées, collectives ou individuelles, en vue de mieux habiter, de mieux comprendre, de mieux construire, de mieux protéger l'espace informationnel : l'éducation, la recherche, les médias, la sécurité.

### 2.1.1. L'éducation

Lutter contre la mésinformation à travers l'apprentissage d'un usage approprié des outils de l'IA générative* réclame un effort considérable d'éducation, auprès de tous les publics. À cet égard, la place du système éducatif national est à l'évidence centrale.

La première urgence est de toucher la cible des enseignants de collèges et lycées, en introduisant, dans leurs cursus de formation, un module spécifiquement consacré à l'utilisation des outils d'IA générative. Les enseignants sont en effet en première ligne pour apprendre à leurs élèves à bien se servir de ces outils, et de manière plus générale d'Internet. Mais, sans solides compétences en ce domaine, ils sont démunis. Doit être mis en place un programme de formation professionnelle destiné aux professeurs de toutes les matières, et non pas seulement à ceux qui, avec l'aide du CLEMI*, assurent les quelques heures de cours d'enseignement moral et civique (EMC*) consacrées à l'éducation aux médias et à l'information (EMI*).

Savoir manier avec efficacité et responsabilité les applications nouvelles du monde numérique devrait figurer comme l'un des objectifs majeurs à poursuivre au cours du cycle secondaire, de même qu'il est impératif de savoir lire, écrire et compter à la sortie de l'école primaire. Interdire aux élèves d'utiliser l'IA générative pour faire leurs devoirs n'aurait pas de sens, puisqu'ils le font déjà et continueront à le faire de toute manière. Il faut tout au contraire les encourager à y avoir recours, mais de manière intelligente, à employer l'IA comme d'un instrument de travail, et non comme un substitut au travail ; notre première proposition, la suscitation d'un ChatPedia*, s'inscrit dans cet esprit (*cf.* 2.2.1.).

Au niveau de l'enseignement supérieur, l'action doit être prolongée, à l'instar de récentes initiatives prises au Québec : l'université Laval a proposé plusieurs principes directeurs pour utiliser l'IA générative et la Ministre de l'Enseignement supérieur a annoncé le 13 août 2024 que son gouvernement « se dote d'une instance de concertation nationale sur l'IA en enseignement supérieur », faisant suite à un rapport publié en avril par le Conseil supérieur de l'éducation (CSE*) et la Commission d'éthique en sciences et en technologie (CEST*) du Québec[43]. Une démarche similaire serait la bienvenue en France, où certaines initiatives privées voient par ailleurs le jour, tel le récent contrat entre l'ESCP Business School et la société OpenAI, prévoyant la mise en place dans cette école d'outils d'IA destinés à l'enseignement ainsi qu'à la gestion administrative[44]. Quant à la formation universitaire de spécialistes des nouvelles technologies numériques, on peut se réjouir de la multiplication, dans notre pays, des cursus de deuxième et troisième cycles consacrés à l'intelligence artificielle et à la *data science**.

L'action éducative ne saurait cependant se limiter à la formation des professeurs et, à travers eux, celle des élèves et des étudiants. Elle doit franchir les murs de l'Éducation nationale pour s'étendre à l'ensemble des citoyens, qu'il convient de sensibiliser aux enjeux, aux limites et aux précautions d'usage de l'IA. Une initiative du *Conseil national du numérique* (CNN*), *Café IA*, vise cet objectif et doit être résolument poursuivie. Dans la ligne de la première proposition d'un récent rapport de la Commission de l'intelligence artificielle* (Aghion et Bouverot, 2024)[45], le projet consiste à créer un dispositif national de débats démocratiques et de partage de ressources pédagogiques sur l'IA et, plus largement sur les technologies numériques. S'appuyant sur des structures nationales et locales – collèges, médiathèques, associations, mairies –, *Café IA* a pour ambition de tisser « une relation au numérique empreinte d'écoute, de partage, d'entraide et de mise en capacité individuelle et collective ».

43. https://collimateur.uqam.ca/collimateur/rapport-2024-intelligence-artificielle-generative-en-enseignement-superieur/
44. https://www.lefigaro.fr/secteur/high-tech/escp-business-school-va-se-transformer-en-profondeur-grace-a-openai-20241014
45. La Commission de l'intelligence artificielle, instituée par la Première ministre Élisabeth Borne en septembre 2023, a remis au président de la république en mars 2024 un rapport « IA : notre ambition pour la France », contenant vingt-cinq recommandations pour faire de la France un acteur majeur de la révolution technologique de l'intelligence artificielle.

On ne peut qu'encourager une pareille entreprise citoyenne et souhaiter qu'elle transforme l'essai de la preuve de concept pour tendre à un plein déploiement, la phase la plus difficile étant ce « passage à l'échelle » !

Au-delà des enjeux d'éducation, le développement des systèmes d'intelligence artificielle soulève de nombreux enjeux d'éthique, examinés dans l'avis n°33 du Groupe européen d'éthique des sciences et des nouvelles technologies (2023) de la Commission européenne. En France, le *Comité national pilote d'éthique du numérique* (CNPEN*), mis en place en décembre 2019, s'est penché sur ces questions dans son avis n°7 (CNPEN, 2023), évoquant notamment le rapport à la vérité, la projection de qualités humaines, le maintien des distinctions, la diversité culturelle, l'écologie, etc. Par décret du 23 mai 2024, ce comité pilote a été transformé en un comité permanent, le *Comité consultatif national d'éthique du numérique* (CCNE du numérique*), dont un des 26 membres doit être proposé par l'Académie des technologies. Il nous apparaît souhaitable que le thème « IA générative et mésinformation » soit inscrit comme une priorité au sein du programme de travail de cette nouvelle instance.

### 2.1.2. La recherche

Comme nous l'avons constaté et signalé *supra* (*cf.* 1.3.), un effort de recherche soutenu doit être poursuivi, en partie financé sur fonds publics, afin que nous soyons à la fois mieux éclairés et mieux armés contre les dangers de la manipulation de l'information. Deux directions apparaissent essentielles, l'une en sciences sociales, l'autre en sciences numériques.

Tout d'abord, compte tenu des fortes incertitudes quant aux modalités et à l'ampleur de l'impact la désinformation* sur les comportements citoyens, il est impératif de parvenir à une meilleure compréhension des mécanismes de formation des opinions et de manipulation de l'information en ligne, et surtout de leurs effets. Les études sociométriques menées en la matière, jusqu'ici très contraintes par le manque de données disponibles, doivent désormais bénéficier d'un socle de données statistiques élargies et diversifiées.

Un partenariat entre l'Insee* et l'Inria* pourrait prendre en charge la constitution de ce socle, son exploitation, et sa mise à la disposition des chercheurs ; tel est l'objet de notre deuxième proposition (*cf.* 2.2.2.).

Ensuite, il est indispensable de remédier aux deux grands écueils qui entravent la recherche en IA consacrée au développement d'outils de détection du faux, à la fois de reconnaissance des contenus artificiels sur tous types de supports et de repérage des comportements synchronisés de désinformation : il faut mettre fin, d'une part à l'insuffisance du financement, d'autre part à la rareté des bases de données disponibles pour l'entraînement (*cf.* 1.4.).

### 2.1.3. Les médias

Les médias historiques peuvent contribuer à relayer la mésinformation* et la désinformation*, comme l'ont récemment montré l'exemple des punaises de lit ou celui des tags antisémites. De ce fait, tout en restant supérieure à celle des réseaux sociaux, leur crédibilité décline et la confiance accordée aux journalistes s'étiole. Une réaction s'impose.

Aux risques avérés de la manipulation informationnelle, il importe d'opposer le strict respect d'un journalisme éthique et professionnel. Or l'exercice de ce métier, dans ses principes comme dans ses pratiques, doit être repensé à l'heure des technologies numériques et de l'IA générative*. C'est dans cet esprit que les *États généraux de l'information* (EGI*), structure délibérative et participative animée par un comité de pilotage, ont été créés à l'initiative du président de la République en septembre 2023. Les conclusions, remises en septembre 2024, apportent des réponses aux cinq interrogations suivantes. Comment la technologie change-t-elle notre rapport à l'information ? Comment restaurer la confiance envers les médias ? Qui doit payer pour une information de qualité ? Comment lutter contre les manipulations de l'information ? Quelle régulation efficace, pour les médias traditionnels comme pour les nouveaux acteurs ? L'importance de l'éducation et du rôle joué par le CLEMI y est particulièrement mise en avant (*cf.* 2.1.1.).

Deux des quinze propositions des EGI portent sur la labellisation de l'information. À cet égard, le rapport des États généraux suggère la mise en place d'un indicateur de fiabilité de l'information et d'un label de certification

de la qualité journalistique d'un média. La démarche apparaît périlleuse pour plusieurs raisons. Tout d'abord, l'organisme en charge de classer les sources, par exemple l'Arcom ainsi qu'il est suggéré, serait toujours suspect de servir tel ou tel intérêt et risquerait d'être perçu comme un instrument de censure au service du pouvoir en place. Ensuite, des critères objectifs permettant de déterminer la fiabilité d'une source sont malaisés à établir, car si les procédures d'édition de l'information sont observables, les intentions sous-jacentes échappent en revanche à la mesure. Enfin, l'appréciation du degré de fiabilité se heurte à un arbitraire certain, sur des sujets où il n'existe pas de vérité indiscutable et où plusieurs opinions sont en lice. Il apparaît par conséquent que, plutôt que d'une prescription normative, la labellisation devrait émaner de la profession elle-même, à travers la définition, en interne, de procédures de qualité : règles encadrant le processus de recueil de l'information, double vérification des faits, correctifs publics en cas d'erreurs, emploi de salariés dont on peut attester de la formation aux bonnes pratiques, etc.

Notre groupe de travail appelle de ses vœux que des réflexions ciblées portent, d'une part sur les chaînes de service public, d'autre part sur les chaînes d'information en continu. Ces chaînes représentent en effet pour le public des sources d'information de référence, ce qui les assujettit à un devoir d'exemplarité, pour les premières, et leur confère une responsabilité particulière, pour les secondes.

En termes d'actions de terrain, l'organisation non gouvernementale internationale *Reporters sans frontières* (RSF*) est à l'origine de deux initiatives importantes, la norme JTI* et le projet Spinoza*.

– La norme JTI (*Journalism Trust Initiative*), conçue comme une norme ISO et lancée au printemps 2021, entend être un dispositif transparent, visant à faire reculer la désinformation et signaler le respect d'un journalisme éthique. Elle constitue un standard des meilleurs pratiques du journalisme, élaboré sous l'égide du Comité européen pour la normalisation (*European Committee for Standardization*) par un groupe de 130 experts comprenant des journalistes, des institutions, des organismes de régulation, des éditeurs, et des acteurs des nouvelles technologies[46]. Aujourd'hui, plus de mille médias dans 80 pays font vivre cette norme.

– Plus récent, le projet Spinoza, mené par RSF en partenariat avec *l'Alliance de la presse d'information générale** a pour ambition de répondre aux enjeux d'indépendance des médias et d'intégrité de l'information à l'ère de l'IA, en renforçant la souveraineté des médias sur leurs moyens de production et en promouvant une culture de l'IA responsable propre à l'éthique journalistique. Il s'agit concrètement de créer un LLM* pour analyser les publications de presse et les données techniques, juridiques et scientifiques liées à la transition écologique. Un prototype devrait être finalisé à la fin 2024.

On ne peut que saluer des initiatives de ce type et souhaiter leur multiplication.

Deux propositions spécifiques de notre rapport se rattachent au champ d'initiatives *Médias* : la troisième proposition (*cf.* 2.2.3.) préconise la création auprès de l'Arcom d'un *Comité consultatif de l'information scientifique et technique* (CCIST*) ; et la quatrième proposition (*cf.* 2.2.4.) prône la mise en place d'un *Observatoire de l'édition artificielle* (OEA*), afin de lever le voile sur les lignes éditoriales de ces nouveaux médias que sont devenus les LLMs.

46. https://rsf.org/fr/journalism-trust-initiative

### 2.1.4. La sécurité

Les technologies de l'information, et singulièrement l'IA*, donnent lieu à une nouvelle forme de guerre, la guerre cognitive* (*cognitive warfare**), que le Commandement Allié Transformation* de l'Otan* définit ainsi : « activités menées en synchronisation avec d'autres instruments de pouvoir, pour altérer les attitudes et les comportements, en influençant, en protégeant, ou en perturbant la cognition individuelle au sein de groupes ou de populations, afin d'obtenir un avantage sur un adversaire ». La cognition humaine est en passe de devenir un domaine critique de guerre et les États déploient des stratégies de vigilance, de lutte et de protection (Claverie, 2024).

En France, la sécurité numérique comporte deux volets : d'une part, détecter et caractériser les ingérences numériques étrangères destinées à manipuler l'opinion et affecter le débat public national ; d'autre part, assurer la cybersécurité*, c'est-à-dire la sécurité des systèmes d'information de l'État et des infrastructures critiques. Ces deux missions sont respectivement confiées à deux organismes distincts : le *Service de vigilance et de protection contre les ingérences numériques étrangères* (Viginum*), créé en 2021 ; et l'*Agence nationale de sécurité des services l'information* (Anssi*), plus ancienne et née en 2009. Ces structures sont toutes deux rattachées au Secrétariat général de la défense et de la sécurité nationale (SGDSN*) mais leurs statuts diffèrent : l'Anssi est une agence publique qui jouit de prérogatives assez similaires à celles ordinairement dévolues, dans des secteurs moins sensibles, à une autorité administrative indépendante ; tel n'est pas le cas de Viginum, service à compétence nationale dont la marge d'autonomie est plus limitée vis-à-vis de sa tutelle.

À l'occasion de l'audition du chef du service Viginum, notre groupe de travail s'est interrogé sur l'opportunité de maintenir cette asymétrie. Pour la corriger, deux voies sont envisageables : ou bien fusionner les deux organes par intégration de Viginum au sein de l'Anssi et faire de l'ensemble ainsi constitué une agence intégrée, sur le modèle britannique du *National Cyber Security Centre* (NCSC*) ; ou bien maintenir les deux organes séparés, mais les placer sur un pied d'égalité à travers une redéfinition des compétences de Viginum, qui fasse de ce service une agence à part entière, lui octroyant ainsi davantage d'indépendance que celle d'un exécutant opérationnel. Cette deuxième voie, qui a la préférence de Viginum, correspond par ailleurs à la

proposition émise par la commission d'enquête du Sénat sur les politiques publiques face aux opérations d'influence étrangères, dont les conclusions ont été rendues publiques en juillet 2024[47].

Il n'appartient pas à l'Académie des technologies de recommander l'une ou l'autre des options précédentes, fusion ou mise à niveau, présentant chacune ses avantages et ses inconvénients, en termes de simplicité, d'efficacité, de flexibilité, ou de budget. En revanche, il nous apparaît souhaitable que le degré de latitude, ainsi que les moyens techniques budgétaires et humains accordés à la lutte contre les ingérences numériques étrangères, soient accrus par rapport au *statu quo*. L'instance en charge de la vigilance devrait être mise en mesure de préparer une stratégie nationale et de coordonner les actions des différents services de l'État compétents pour sa mise en œuvre, au sein des ministères de l'intérieur, des armées et des affaires étrangères.

Outre ce vœu, nous formulons deux propositions spécifiques dans le champ de la sécurité : d'une part, cinquième proposition, imposer aux plateformes l'affichage d'un *score d'artificialité**, signalant au public le degré de présomption qu'un contenu viral soit d'origine synthétique et massivement diffusé par des moyens non humains (*cf.* 2.2.5.) ; d'autre part, sixième proposition, réparer une omission au sein de l'appareil juridique, en instaurant un régime de sanctions applicable à tous les types d'ingérences numériques étrangères (*cf.* 2.2.6.).

47. Lire ici le rapport de mission : https://www.senat.fr/rap/r23-739-1/r23-739-11.pdf

## 2.2. Six propositions spécifiques

**Si certaines de ces propositions préconisent la mise en place de dispositifs institutionnels auprès d'instances existantes, aucune ne prône la création *ex nihilo* de nouveaux organes administratifs.**

**Dans le domaine de l'éducation**…

### 2.2.1. Faire émerger un ChatPedia au sein de l'Éducation nationale

L'objectif est de susciter, au sein de l'Éducation nationale, l'éclosion d'un ChatPedia*, outil collaboratif d'IA visant à aider professeurs et élèves à un usage « intelligent » de l'IA, à des fins d'enseignement et d'apprentissage.

Cette proposition s'appuie sur les observations suivantes.

– Entre 2023 et 2024, la perception des outils d'IA par les lycéens a évolué : considérés au départ comme des algorithmes*, ils sont devenus des « assistants », auxquels on peut demander n'importe quelle sorte d'aide, car « ils savent tout ».

– Cependant, les discussions montrent que les lycéens ne savent pas les faire fonctionner et en obtenir l'aide voulue. Notons que les conseils aux utilisateurs professionnels ne sont guère plus avancés[48] : il faut essayer… et si ça ne marche pas, essayer encore, le *prompt engineering* étant un art tout d'exécution (*cf.* 1.2.).

– Le monde de l'éducation, de la petite école à l'université, est en partie démoralisé : d'une part, il faut revoir les modalités d'évaluation et, d'autre part, il devient de plus en plus difficile de répondre à la question « À quoi cela sert-il d'aller à l'école ? ».

– Toute interaction avec un LLM*, si elle ne le biaise pas, le renforce : les échanges apportent des informations nouvelles aux ressources du modèle et lui permettent de s'améliorer.

48. https://www.langchain.com/

Dans ce contexte, nous proposons une initiative réunissant enseignants, chercheurs, enfants, collégiens, lycéens et étudiants, sur la base du volontariat, avec la visée à terme de créer un bien commun évolutif (ChatPedia), qui serait à un LLM véridique ce que Wikipedia est à une encyclopédie.

**Cas concret d'application dans le cycle secondaire.** À partir de la classe de seconde, les professeurs continueraient d'appliquer la « bonne vieille méthode » d'enseignement, mais en y intégrant l'usage d'un LLM *open source* en langue française : sur une thématique donnée, ils répartiraient les élèves en groupes, chaque groupe étant libre d'interroger le LLM à sa manière, puis les groupes réunis critiqueraient collectivement les réponses obtenues par chacun d'eux.

Une telle méthode devrait permettre :

– d'affiner l'esprit critique des lycéens sans que l'enseignant en soit la cible, d'apprendre à distinguer le vrai du plausible, de se familiariser avec les pratiques et les principes de vérification ;

– d'apprendre, aux élèves comme aux enseignants, à faire du LLM un assistant précieux ;

– d'augmenter les ressources du LLM de manière massive, traçable, et de les rendre utilisables par tous.

**Dans le domaine des études et de la recherche**…

#### 2.2.2. Créer un socle statistique du numérique et de ses impacts

Le Web est désormais au cœur de la vie économique, comme des activités politiques et des relations sociales. Il est devenu une infrastructure immatérielle de communication essentielle. La multiplication des services numériques a créé un monde dit virtuel, qui cohabite avec le monde réel. De fait, les activités virtuelles et réelles s'entrelacent en permanence dans la vie des entreprises comme dans celle des citoyens. Pour autant, les règles ne sont pas les mêmes dans ces deux mondes imbriqués : le monde numérique soulève des questions spécifiques, non traitées par le droit standard, et les

récentes règlementations, comme les règlements sur les marchés et services numériques (DMA* et DSA*), ou comme l'*AI Act** visent à pallier ces carences.

S'agissant de la « vie réelle », de multiples institutions (Insee*, IGN*, Météo France*, Ined*,…) ont pour rôle de décrire, d'analyser, de quantifier le monde physique dans lequel nous évoluons. Rien d'équivalent n'existe pour le monde virtuel, sujet à l'opacité des acteurs économiques réglant la grande majorité de nos activités en ligne (Google, Amazon, Apple, Microsoft, Meta, X, TikTok,…). Ces acteurs ont créé des écosystèmes en multipliant les services et en s'appuyant sur des algorithmes* – eux aussi évolutifs et opaques –, avec pour objectif de conserver notre attention le plus longtemps possible et d'augmenter chiffres d'affaires et profits (*cf.* 1.1.3.). Les informations dont nous disposons sur ce qu'il se passe, aussi bien au sein d'un de ces écosystèmes, qu'entre chacun d'eux et le reste du Net, dépend d'un bon-vouloir jusqu'à présent assez limité[49].

La méconnaissance qui en résulte est frappante, au regard de l'importance prise par le monde numérique dans nos vies quotidiennes ; réduire cette méconnaissance est un prérequis à la définition de politiques publiques efficaces.

Cela exige un renforcement de l'accès aux données, *via* une pleine et rigoureuse application du DSA (Règlement sur les services numériques) et probablement la création d'une API* pour chacun des grands réseaux sociaux. Mais, si le DSA ouvre certes la porte de l'accès à certaines données pour les travaux de recherche, cette ouverture ne suffit pas à elle seule. Notre rapport met en évidence la nécessité de réaliser des études qui permettent de comprendre les impacts de la désinformation circulant à travers les réseaux sociaux sur les pratiques hors ligne : votes, vaccins, agressions, etc. Or ces études nécessitent, non seulement l'accès aux données de l'univers numérique, mais aussi un suivi statistique de différentes pratiques observables dans le monde réel.

49. Un contre-exemple à cette affirmation, les données collectées par l'Arcep* auprès des acteurs pour décrire et évaluer l'interconnexion des flux de données en France : *cf.* Baromètre de l'interconnexion de données en France https://www.arcep.fr/cartes-et-donnees/nos-publications-chiffrees/linterconnexion-de-donnees/barometre-de-linterconnexion-de-donnees-en-france.html

Il est par conséquent impératif de se doter d'un corpus d'informations mixte, portant à la fois sur la vie en ligne et sur la vie hors ligne qui est susceptible d'en dépendre, afin qu'un pont puisse ensuite être jeté entre ces deux continents de pratiques. La création de ce corpus, nécessaire pour élargir la vision trop étroite qui est aujourd'hui celle des institutions publiques, pourrait être l'objet d'un partenariat entre l'Insee et l'Inria.

**Dans le domaine des médias et des LLMs**…

#### 2.2.3. Instaurer auprès de l'Arcom un *Comité consultatif de l'information scientifique et technique*

L'objectif est de favoriser, dans la communication audiovisuelle et en ligne, l'expression d'experts scientifiques et techniques reconnus par leurs pairs, sur les questions liant sciences, technologies et société.

Les innovations technologiques sont au cœur de presque tous les grands défis auxquels la société est confrontée : comment nourrir la population mondiale ? comment donner accès à la formation et à l'emploi au plus grand nombre ? comment favoriser un système de santé et de soins qui réponde aux besoins d'une population croissante et vieillissante ? comment lutter contre le changement climatique ? De nos jours, il faut non seulement comprendre le potentiel des technologies récemment développées, mais aussi leurs retombées sur nos vies, qu'il s'agisse d'opportunités ou de menaces.

Les citoyens obtiennent la plupart de leurs informations sur les avancées technologiques *via* les médias. Or, force est de constater que le traitement de l'information scientifique et technique, par les médias audiovisuels et par les réseaux sociaux, est trop souvent partial, avec une tendance à relayer les discours alarmistes d'experts, ou prétendus tels, sans garantie minimale de neutralité et de transparence. Paradoxalement, alors que les technologies n'ont jamais autant irrigué notre quotidien, le progrès et la légitimité scientifique sont remis en cause. L'alarmisme ambiant inquiète le plus souvent l'opinion publique sans raison, en faisant croire à l'existence de risques non avérés, ce qui conduit potentiellement à des décisions politiques peu avisées, comme l'arrêt d'expérimentations, qui sont coûteuses en termes de développement économique, d'emploi, de santé et de sécurité.

En France, aucune instance ni texte de référence ne traite spécifiquement de la déontologie de l'information scientifique et technique dans les médias. Or notre pays dispose d'un grand réservoir d'experts reconnus internationalement, sélectionnés par leurs pairs sur leur valeur et leur expérience, qui représentent une capacité d'expertise collective sans égale : il s'agit des différentes académies, des sciences, des technologies, des sciences morales et politiques, de médecine, d'agriculture, de l'air et de l'espace, etc. Leurs membres, près de mille globalement, issus des universités, des grandes écoles, des établissements de recherche et des entreprises, sont en permanence impliqués dans la vie intellectuelle et économique de la nation. Le large spectre compétences ainsi déployé permet de couvrir aussi bien les aspects scientifiques, technologiques, qu'économiques, sociétaux ou éthiques du progrès.

Construit sur le modèle britannique du *Science Media Center* (SMC*) et placé auprès de l'Arcom*, le *Comité consultatif de l'information scientifique et technique* (CCIST*) proposé aurait pour rôle, non pas de labelliser des contenus, mais de promouvoir et faire respecter une déontologie de l'information scientifique et technique dans les médias historiques et en ligne. Tout en respectant la liberté de communication, la mission du Comité serait de protéger le public contre les abus de cette liberté qui, singulièrement en matière d'information scientifique et technique, agit au détriment de la crédibilité de l'information et de la confiance du public envers les médias. Cette mission s'inscrirait dans le cadre plus général du pluralisme et de l'équité des débats dans la communication audiovisuelle, ainsi que de l'honnêteté de l'information.

Les activités du CCIST ne se limiteraient pas à l'émission d'avis critiques, une telle instance se devant également de constituer un lieu de veille, de pédagogie, d'observation et de réflexion. Le Comité pourrait ainsi émettre des avis positifs, pour souligner les bonnes pratiques susceptibles de venir enrichir les chartes déontologiques des médias. Il pourrait également conduire des études et entreprendre des actions de sensibilisation et de formation.

Dans cet esprit, le CCIST encouragerait davantage d'experts à s'engager plus efficacement pour faire entendre des avis raisonnés et équilibrés, dans un contexte où les questions de science, de technologie et d'innovation

industrielle sont soumises à controverse et où la parole de l'expert n'est pas prise en considération, voire déniée sans examen de ses arguments.

Trois dangers doivent être prévenus pour assurer l'efficacité d'un tel CCIST :

– tout d'abord, un possible entrisme de la part de groupes défendant leurs intérêts économiques privés plutôt que l'intérêt public ;

– ensuite, l'effet pervers qui résulterait de ce que des experts se réclamant pourtant d'une *sound science*, science saine et solide, remettent abusivement en cause l'application du principe de précaution, arguant que toute réglementation du déploiement d'une technologie est infondée, tant qu'il n'est pas démontré que celle-ci est nocive ou dangereuse.

– enfin, une baisse du degré d'implication des médias et des plateformes, qui pourraient préférer s'en remettre au Comité plutôt que prendre eux-mêmes des initiatives de communication fiable.

Par ailleurs, constituer un vivier d'experts reconnus est une condition nécessaire, mais non suffisante, de bon fonctionnement du Comité. Il importe également d'adopter des méthodes permettant d'élaborer une expertise collective crédible, telles que décrites dans le rapport de France Stratégie *Expertise et démocratie, faire avec la défiance* (Agacinski *et alii*, 2018)[50] : quels experts ont été sollicités, d'où parlent-ils, qu'ont-ils considéré, comment ont-ils procédé ?

50. https://www.strategie.gouv.fr/publications/expertise-democratie-faire-defiance

#### 2.2.4. Bâtir un *Observatoire de l'édition artificielle*

Un LLM* est nécessairement « aligné » sur les valeurs de son éditeur et présente des biais spécifiques, dus au choix de sa base d'entraînement* et aux modalités de son apprentissage*. Cet alignement lui confère l'équivalent d'une « ligne éditoriale » et il en résulte des asymétries entre LLMs, dans le traitement de certaines questions sociétales et éthiques. Mais alors que la ligne éditoriale d'un média traditionnel est le plus souvent connue et relativement transparente, celle d'un LLM est à ce stade opaque pour l'utilisateur.

Nous recommandons en conséquence que soit constitué, par exemple à l'initiative du Secrétariat d'État au numérique et à l'intelligence artificielle, un *Observatoire de l'édition artificielle*, OEA*, investi d'une triple mission.

- Mettre au point des outils, par exemple sous la forme de batteries de questions-réponses, permettant de cerner la ligne éditoriale et le champ d'influence d'un LLM. Deux registres au moins devraient être scrutés : d'une part, les postures prises sur des sujets touchant au genre, à l'appartenance ethnique ou à la religion ; d'autre part, les positions mises en avant lors de controverses et débats politiques.

- Tester régulièrement, à l'aide des outils précédents, les principaux LLMs mis à la disposition du grand public.

- Publier annuellement une analyse transparente des « courants d'opinion » des LLMs.

**Dans le domaine de la sécurité**…

#### 2.2.5. Contraindre les plateformes à afficher un *score d'artificialité*

Les grandes plateformes numériques* interdisent les comportements artificiels, notamment les faux comptes, dont beaucoup sont suspendus par leurs soins chaque année (*cf.* 1.1.5.). En outre, certaines campagnes de désinformation sont détectées par Viginum*, qui informe le service Pharos*, qui prévient ensuite les grandes plateformes, ce qui conduit à la suspension des comptes et contenus impliqués. Mais, en dépit de ces efforts, un grand

nombre d'actions coordonnées de désinformation persistent en ligne, la liberté d'expression devant avant tout être protégée en cas de doute sur leur caractère illicite. L'*AI Act** exige certes le signalement de contenus créés par des IA génératives, mais certains acteurs malveillants savent contourner ce dispositif.

Pour en augmenter l'efficacité, nous formulons la proposition suivante, portée à l'attention des autorités en charge de la mise en œuvre du DSA et de l'*AI Act* (Commission européenne, *AI Office**, Arcom*), mais aussi de la cybersécurité* et de la défense contre les ingérences étrangères (Anssi*, Viginum*). Il s'agirait d'imposer, en application du DSA*, un *score d'artificialité** : il afficherait, d'une part une estimation de la probabilité d'inauthenticité d'un contenu – dans quelle mesure a-t-il été engendré par une machine ? – et, d'autre part la probabilité d'implication de réseaux de faux comptes dans sa diffusion. Le calcul et l'affichage du score devraient être rendus obligatoires pour les très grandes plateformes, dès lors qu'un contenu dépasse un certain seuil de viralité.

Dans sa nature, l'obligation envisagée est assez semblable à celle de porter à la connaissance des internautes qu'un contenu est présenté par une plateforme en raison de ses liens financiers ou capitalistiques avec l'annonceur. Pour respecter les droits des éditeurs des contenus visés, une procédure serait prévue, conformément au DSA, afin de leur permettre de contester le *score d'artificialité*.

Outre l'information du citoyen, le score aiderait les magistrats et les régulateurs, en cas de demande de suspension du contenu, à évaluer le niveau de protection à accorder au titre de la liberté d'expression. Le score serait un élément de preuve, parmi d'autres, que le contenu a été créé et diffusé « de mauvaise foi » et relève ainsi de l'une des interdictions prévues par la loi.

### 2.2.6. Sanctionner les opérations de désinformation au bénéfice d'acteurs étrangers

L'actuel article R 1132-3 du code de la Défense confie à Viginum* la mission d'identifier des opérations de désinformation* effectuées au bénéfice d'une puissance étrangère, sans toutefois prévoir un dispositif de sanctions pénales. Pourtant, d'autres activités, notamment la fourniture de fausses informations aux autorités civiles ou militaires françaises, sont réprimées par le code pénal, dès lors qu'elles sont de nature à porter atteinte aux intérêts fondamentaux de la nation. Nous proposons en conséquence de compléter le dispositif pénal en insérant dans le code un nouvel article visant spécifiquement les opérations de désinformation au bénéfice d'une puissance étrangère.

Cet article serait à ajouter après l'actuel l'article 411-10 du code pénal et avant l'actuel article 411-11 : « *Le fait de diffuser de manière artificielle ou automatisée, massivement et délibérément, par le biais d'un service de communication au public en ligne, en vue de servir les intérêts d'une puissance étrangère, d'une entreprise ou organisation étrangère ou sous contrôle étranger, d'allégations ou imputations de faits manifestement inexactes ou trompeuses de nature à porter atteinte aux intérêts fondamentaux de la nation est puni de sept ans d'emprisonnement et de 100 000 euros d'amende.* ».

Pour garantir le respect de la liberté d'expression, les six conditions cumulatives déjà énoncées à l'article R 1132-3 du code de la défense pourraient être reprises, à savoir : (i) des allégations ou imputations de faits manifestement inexactes ou trompeuses ; (ii) diffusées de manière artificielle ou automatisée ; (iii) massivement et délibérément ; (iv) par le biais d'un service de communication au public en ligne ; (v) en vue de servir les intérêts d'une puissance étrangère, d'une entreprise ou organisation étrangère ou sous contrôle étranger ; (vi) et de nature à porter atteinte aux intérêts fondamentaux de la nation.

Le caractère cumulatif de ces conditions s'inspire de la jurisprudence du Conseil Constitutionnel, concernant en particulier la *loi infox* du 22 décembre 2018. Parce qu'il réduit en pratique l'applicabilité du dispositif, et compte tenu de l'intensité de la menace, nous proposons que soit examinée la possibilité d'élargir le champ d'application des sanctions. Cela nécessiterait néanmoins une analyse préalable de constitutionnalité par le Conseil d'État.

# LE MOT PHILOSOPHIQUE DE LA FIN

Les réflexions présentées dans ce rapport présupposent que, d'une façon ou d'une autre, nous garderons la maîtrise de l'IA* et que nous saurons, à force de sensibilisation, d'éducation et de régulation, la placer sur des rails où elle servira la connaissance plutôt que la mésinformation*. Cette recherche d'une harmonie entre technologie et société, au cœur des missions de l'Académie des technologies, n'est pas nouvelle. Si, au moment de concevoir leur *Encyclopédie*, Diderot ou d'Alembert ont choisi d'y insérer de très nombreuses planches et illustrations expliquant en détail le fonctionnement d'une multitude d'objets techniques, c'est en vertu d'un principe qui leur semblait aller de soi : les objets techniques, en devenant visibles et familiers, seraient implicitement vecteurs de connaissances scientifiques ; plus nous nous frotterons à eux dans la vie quotidienne, pensaient-ils, mieux nous connaîtrons et comprendrons les principes scientifiques qui les ont rendus possibles.

Il fut sans doute une époque où l'homme cultivé, *l'honnête homme*, pouvait comprendre tous les outils et toutes les machines qui l'entouraient ; mais les Encyclopédistes n'avaient point anticipé une autre réalité qui, au fil du temps, allait peu à peu s'imposer : plus un objet technologique est complexe, plus son usage tend à se simplifier. Ainsi, presqu'aucun d'entre nous ne saurait dire comment fonctionnent une télévision, un ordinateur ou un téléphone portable, ce qui ne nous empêche nullement de nous en servir sans avoir besoin de consulter la moindre notice et sans que notre crasse ignorance ne nous fasse trembler d'angoisse. Ainsi certains objets techniques, à la fois extrêmement familiers et extraordinairement complexes, en viennent-ils à masquer ou à marginaliser les connaissances scientifiques et les compétences techniques dont ils sont pourtant les conséquences. Ces connaissances et ces compétences sont alors perçues comme pratiquement inutiles – inutiles en pratique –, donc inutiles tout court, un grave danger pouvant conduire à une « apocalypse cognitive ».

Un pas de plus dans cette direction a été franchi avec l'IA, car la convivialité du monde numérique, le diktat du *user friendly*, tend à escamoter quasi-totalement les techniques sous-jacentes Dès lors, conserverons-nous suffisamment de compétences pour comprendre la manière qu'a l'IA de fonctionner et de façonner nos esprits comme nos vies ? Cela est certes à espérer fortement et à rechercher résolument, mais ne va pas de soi, ni n'est tout à fait certain.

Les générations à venir devront savoir 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 la pérennité. La belle devise de l'Académie des technologies leur balise le sentier :

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

# Annexes

## A. Références

**Académie des technologies**, *Prouesses et limites de l'imitation artificielle de langages*, avril 2023. https://www.academie-technologies.fr/publications/prouesses-et-limites-de-limitation-artificielle-de-langages-avis/

**Agacinski**, D. *et alii, Expertise et démocratie : faire avec la défiance*, rapport de France Stratégies, décembre 2018.

**Aghion**, P. et A. Bouverot, *I.A. Notre ambition pour la France*, Odile Jacob, 2024.

**Aïmeur**, E., S. Amri and G. Brassard, "Fake news, disinformation and misinformation *in* social media : a review", *Social Network Analysis and Mining* 13.1: 30, 2023.

**Allen**, J., *et alii*, "Evaluating the *fake news* problem at the scale of the information ecosystem", *Science advances* 6.14 : eaay3539, 2020.

**Altay**, S., Berriche M., Acerbi A., "Misinformation on misinformation: Conceptual and methodological challenges", *Social media+ society* 9.1: 20563051221150412, 2023.

**Barlow**, J.-P., *Déclaration d'indépendance du cyberspace*, Forum de Davos, 1996.

**Benkler**, Y. *et alii*, "Mail-in voter fraud: Anatomy of a disinformation campaign" Berkman *Center Research Publication* 2020-6, 2020.

**Berglund**, L., *et alii*, "The Reversal Curse: LLMs trained on 'A is B' fail to learn 'B is A'", *Computer Science*, September 2023. https://arxiv.org/pdf/2309.12288v2

**Bolukbasi**, T., *et alii*, "Man is to Computer Programmer as Woman is to Homemaker. Debiaising Word *Embeddings*", *Computer Science*, July 2016. https://arxiv.org/pdf/1607.06520

**Bontcheva**, K., *et alii*, *Generative AI and Disinformation: Recent Advances, Challenges, and Opportunities*, University of Sheffield, February 2024.

**Brandolini**, A. "The bullshit Asimmetry", *Twitter*, 11 janvier 2013.

**Bronner**, G., *La démocratie des crédules*, PUF, 2013.

**Bronner**, G., *Apocalypse cognitive*, PUF, 2021.

**Bronner**, G., *Les lumières à l'ère numérique*, Rapport à la Présidence de la République, janvier 2022.

**Canetta**, T., "Prebunking AI-generated disinformation ahead of EU elections", EDMO *Publications*, 27 March 2024. https://edmo.eu/publications/prebunking-ai-generated-disinformation-ahead-of-eu-elections/

**Chavalarias**, D., *Toxic data*, Éditions Flammarion, 2022.

**Claverie**, B., « Cognitive Warfare – une guerre invisible qui s'attaque à notre pensée » *in* J-F. Trinquecoste *Faut-il s'inquiéter ?*, Éditions de l'IAPTSEM, pp. 89-115, 2024.

**CNPEN**, *Systèmes d'intelligence artificielle générative : enjeux d'éthique*, Avis n° 7, 30 juin 2023.

**Costello**, T., Pennycook G., Rand D., "Durably reducing conspiracy beliefs through dialogues with AI", *Science*, September 2024. https://www.science.org/doi/10.1126/science.adq1814

**Curien**, N., « Connaissance à l'ère numérique : la possibilité du vrai ? », *Variance.eu*, mai 2020. https://variances.eu/?p=5024

**Da Empoli**, G., *Les ingénieurs du chaos*, JC Lattès, 2019.

**Eco**, U., *La guerre du faux*, Éditions Grasset & Fasquelle,1985.

**Enqvist**, L., "Human oversight in the EU artificial intelligence act: what, when and by whom?", *Law, Innovation and technology*, vol. 15, 2023.

**Ferrara**, E., "GenAI against humanity: nefarious applications of generative artificial intelligence and large language models", *Journal of Computational Social Science*, 2024. https://doi.org/10.1007/s42001-024-00350-1

**Grinbaum**, A., *Parole de machines*, Éditions humenSciences / Humensis, 2023.

**Groupe européen d'éthique des sciences et des nouvelles technologies**, *La démocratie à l'ère numérique*, Avis nº 33, Commission européenne, Juin 2023.

**Gu**, C., Kurov C., "International role of social media: Evidence from Twitter sentiment", *Journal of Banking and Finance*, vol. 121, December 2020. https://www.sciencedirect.com/science/article/abs/pii/S0378426620302314

**Jean**, A. *De l'autre côté de la machine : voyage d'une scientifique au pays des algorithmes*, L'observatoire, 2019.

**Kahneman**, D., *Thinking Fast and Slow*, Penguin, 2012.

**Karpathy**, A., "The Unreasonable Effectiveness of Recurent Neural Networks", *Anfrej Karpathy blog*, May 2015. https://karpathy.github.io/2015/05/21/rnn-effectiveness/

**LeCun**, Y., "Do large language models need sensory grounding for meaning and understanding?", *Courant Institute & Center for Data Science*, NYU, March 2023. https://lnkd.in/eueY5EAq

**Lorenz-Spreen**, P., Oswald, L., Lewandowsky, S., *et alii*, "A systematic review of worldwide causal and correlational evidence on digital media and democracy", *Nature Human behavior*, vol. 7, pp.74-101, 2023.

**Mill**, J.-S., *De la liberté*, (1859), Éditions du Grand Midi, 1987.

**Milton**, J., *Areopagitica*, ou De la liberté de la presse et de la censure, 1644. https://gallica.bnf.fr/ark:/12148/bpt6k424985t

**Murphy**, G., *et alii*, "What do we study when we study misinformation? A scoping review of experimental research (2016-2022)" *Harvard Kennedy*

*School Misinformation Review*, 2023. https://misinforeview.hks.harvard.edu/article/what-do-we-study-when-we-study-misinformation-a-scoping-review-of-experimental-research-2016-2022/

**Nyhan**, B., Settle, J., Thorson, E., *et alii*, "Like-minded sources on Facebook are prevalent but not polarizing", *Nature* 620, pp. 137–144, 2023. https://doi.org/10.1038/s41586-023-06297-w

**Oreskes**, N. et E.M. Conway, *Les marchands de doutes*, Éditions Le Pommier, 2012.

**Origgi**, G., *La vérité est une question politique*, Albin Michel, 2024.

**Ouyang**, L., *et alii*, "Training language models to follow instructions with human *feedback*", *Computer Science*, 2022. https://arxiv.org/abs/2203.02155

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

**Petroni**, F., *et alii*, "Improving Wikipedia Verifiability with AI" in *Nature machine intelligence*, 5 : 1142-1148, 2023.

**Searle**, J.R., "Minds, Brains and Programs", *The Behavioral and Brain Sciences*, vol. 3, Cambridge University Press, 1980.

**Surowiecki**, J., *The Wisdom of Crowds : Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*, Little Brown, 2004.

**Teilhard de Chardin**, P., *Le phénomène humain*, Seuil, 1955.

**Türk**, P., « Liberté d'expression et régulation des fausses nouvelles en France – Entre tradition libérale et impératif de régulation : un encadrement souple de l'expression en ligne en France », Rapport France, *in* O. Pollicino (dir.), *Freedom of Speech and the Regulation of Fake News*, Cambridge, Intersentia, 18 août 2023.

**Vaswani**, A., *et alii*, "Attention is all you need", *Advances in Neural Information Processing Systems, 30. Curran Associates, Inc.*, 2017.

**Watts**, Duncan J., D.-M. Rothschild, Mobius M.. "Measuring the news and its impact on democracy", *Proceedings of the National Academy of Sciences* 118.15: e1912443118, 2021.

**Weikmann**, T. and S., Lecheler, "Cutting through the Hype : Understanding the Implications of *Deep fakes* for the Fact-Checking Actor-Network", *Digital Journalism*, 1–18, 2023. https://doi.org/10.1080/21670811.2023.2194665

**Weil**, S., « *Plaidoyer pour une civilisation nouvelle* » [1943], *in* Simone Weil, Œuvres, Paris, *Quarto* Gallimard, p. 1050, 2000.

**Weinberger**, D., *Everyday Chaos: Technology, Complexity and How We're Thriving in a new World of Possibility*, Harvard Business Review Press, 2019.

**World Economic Forum**, The Global Risks Report 2024, 19th Edition. https://www3.weforum.org/docs/WEF_The_Global_Risks_Report_2024.pdf

**Yang**, K.-C. and F. Menczer, "Anatomy of an AI-powered malicious social botnet", *Journal of Quantitative Description : Digital Media.* Vol. 4, 2024. https://doi.org/10.51685/jqd.2024.icwsm.7

**Zellers**, R., *et alli*, "Defending Against Neural Fake News", *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2020.

## B. Glossaire des sigles et termes techniques

**AFP.** *Agence France Presse*, l'une des trois grandes agences mondiales d'information, la seule européenne. Sa mission est d'assurer une couverture rapide, complète, impartiale et vérifiée de l'actualité.

***AI Act.*** Règlement européen émis en juillet 2024, encadrant les conduites des acteurs de l'intelligence artificielle, qu'ils soient producteurs de systèmes ou fournisseurs de services.

***AI Office.*** Centre d'expertise de la Commission européenne en matière d'intelligence artificielle. Il joue un rôle clé dans la mise en œuvre de l'*AI Act*, promouvant le développement et l'utilisation de « l'IA de confiance » et en favorisant la coopération internationale.

**Algorithme.** Enchaînement d'étapes permettant d'obtenir un résultat à partir d'éléments fournis en entrée. Les plateformes numériques utilisent des algorithmes pour hiérarchiser les contenus, pour les présenter aux utilisateurs de manière ciblée, et pour leur proposer des recommandations.

**Alliance de la presse d'information générale.** Syndicat patronal créée en 2012 et regroupant les éditeurs de quotidiens et de magazines d'information politique et générale.

***Allied Command Transformation*** **(ACT).** Voir **Commandement Allié Transformation.**

**AMF.** *Autorité des marchés financiers*. Autorité publique indépendante, chargée en France de la régulation des marchés financiers. Ses missions comprennent la protection de l'épargne investie dans les instruments financiers, l'information des investisseurs et le bon fonctionnement des marchés d'instruments financiers.

**Anssi.** *Agence nationale de la sécurité des systèmes d'information*. Agence publique française à compétence nationale, créée en 2009 et rattachée au Secrétariat général de la Défense et de la Sécurité nationale (SGDSN).

**API.** *Application Programming Interface*, ou interface de programmation applicative, est une « façade » par laquelle un logiciel offre des services à d'autres logiciels. Une API permet ainsi à l'utilisateur d'accéder, depuis un seul « guichet », à des données issues de plusieurs applications, sans avoir à se soucier du mécanisme informatique coopératif sous-jacent à la fourniture de ces données.

**Apprentissage automatique.** Encore appelé **apprentissage machine** (*machine learning*), **apprentissage artificiel** ou **apprentissage statistique**, l'apprentissage automatique est un champ d'études de l'intelligence artificielle se fondant sur des approches mathématiques et statistiques pour donner aux ordinateurs la capacité « d'apprendre » à partir de données, c'est-à-dire d'améliorer leurs performances dans la réalisation de tâches, sans être explicitement programmés pour chacune d'elles.

**Apprentissage profond** ou ***Deep learning*.** Domaine de l'intelligence artificielle utilisant des réseaux neuronaux multicouches pour effectuer des tâches complexes, comme la reconnaissance d'éléments sur une image : à partir d'un très grand nombre de données d'entrée, la machine forme par elle-même des « représentations intermédiaires » utiles à l'exécution de la tâche, sans qu'un humain n'ait à expliciter la manière de procéder à l'aide d'un algorithme.

**Apprentissage par renforcement.** Il consiste, pour une IA, à apprendre quelles sont les actions les plus pertinentes, en maximisant au cours du temps le cumul d'une séquence de récompenses ou punitions, administrées par un opérateur humain en fonction de la qualité des réponses fournies.

**Apprentissage supervisé.** Forme d'apprentissage automatique suivant lequel une IA construit son mécanisme de prédiction à partir d'exemples annotés ; si la base d'entraînement n'est pas annotée, on parle d'apprentissage non supervisé.

**Arcep.** *Autorité de régulation des communications électroniques*, des postes et de la distribution de la presse, elle a le statut d'Autorité administrative indépendante. Complémentaire de l'Arcom, qui régule les éditeurs de contenus, elle régule les opérateurs de « tuyaux ».

**Arcom.** *Autorité de régulation de la communication audiovisuelle et en ligne*. Autorité publique indépendante française, notamment en charge de la régulation des grandes plateformes numériques.

**Astroturfing.** Technique de manipulation de l'information consistant à simuler artificiellement et à des fins manipulatrices un mouvement spontané d'opinion.

**Autorégressif.** Se dit d'un modèle récursif dont le résultat fourni en sortie à une étape donnée est engendré, selon un processus séquentiel, à partir des données d'entrée et des résultats des étapes précédentes. Un LLM, qui extrapole une séquence de *tokens* à partir des *tokens* précédents, est un modèle autorégressif.

**Base d'entraînement.** Base de données servant à l'apprentissage d'une IA.

**Biais algorithmique.** Fait que le résultat produit par un algorithme ou par une IA n'est pas neutre, loyal ou équilibré, que ce soit de manière involontaire ou délibérée.

**Biais cognitif.** Déviation ou raccourci dans le traitement cognitif d'une information, parfois motivé par l'urgence d'agir. Un tel schéma de pensée, trompeur et faussement logique, peut induire des erreurs ou des paradoxes dans un raisonnement ou dans un jugement.

***Bot***. Un *bot* informatique est un agent logiciel automatique ou semi-automatique, c'est-à-dire un programme informatique, qui peut interagir de manière autonome avec des serveurs, par exemple pour diffuser de fausses nouvelles.

**Broker.** Courtier, en français, il agit en tant qu'intermédiaire entre un acheteur et un vendeur de produits financiers.

**Bulle de filtre.** Espace restreint des contenus auxquels un internaute donné a effectivement accès, compte tenu d'un filtrage résultant notamment d'une exploitation de l'historique de ses parcours sur le Web et de ses activités sur les réseaux sociaux.

***Bullshit Asymmetry Principle.*** Loi empirique énoncée en 2013 par le programmeur italien Alberto Brandolini, affirmant que restaurer la vérité en réparation d'une infox réclame au moins dix fois plus de temps et d'énergie que n'en ont coûté à son auteur la production et la diffusion de l'infox.

**CAC 40.** Principal indice boursier de la Bourse de Paris. Il reflète la performance des 40 actions les plus importantes et les plus activement négociées, parmi celles cotées sur Euronext Paris.

**Café IA.** Projet lancé en 2024 par le Conseil national du numérique, en vue de sensibiliser le grand public aux opportunités, aux limites et aux dangers liés à l'utilisation de l'IA et, plus généralement, des technologies numériques.

**CCIST (prospectif).** *Comité consultatif de l'information scientifique et technique*. Proposition de créer auprès de l'Arcom une instance veillant à la qualité de ce type d'information dans les médias traditionnels et en ligne.

**CCNE du numérique.** Comité consultatif national d'éthique du numérique, succédant en mai 2024 au CNPEN.

**CESN.** Comité européen des services numériques, créé en février 2024, présidé par la Commission européenne et regroupant les organes chargés de la coordination des services numériques dans les différents États membres, l'Arcom en France.

**CEST.** *Commission d'éthique en science et en technologie*, au Québec.

**Chambre d'échos.** Phénomène par lequel un internaute fréquente de manière privilégiée, voire exclusive, des internautes partageant les mêmes centres d'intérêt et les mêmes opinions qu'elle ou lui.

***Chatbot.*** *Bot* conversationnel, dialoguant avec un utilisateur.

**ChatGPT.** Premier LLM ayant été mis à la disposition du grand public en novembre 2022, par la société OpenAI.

**ChatPedia (prospectif).** Proposition de susciter l'émergence d'un LLM collaboratif au sein de l'Éducation nationale, en vue d'y promouvoir un usage « intelligent » de l'IA par les élèves et les professeurs.

**CIB.** *Coordinated Inauthentic Behavior*. Le comportement inauthentique coordonné est une stratégie de communication manipulatrice sur les réseaux sociaux, s'appuyant sur un mix de comptes authentiques, faux ou dupliqués.

**CLEMI.** *Centre pour l'éducation aux médias et à l'information*. Ce service de Réseau Canopé, opérateur public placé sous la tutelle de l'Éducation nationale, est chargé de l'éducation aux médias et à l'information (EMI) dans l'ensemble du système éducatif français jusqu'à la fin du cycle secondaire.

**CNIL.** *Commission nationale de l'informatique et des libertés*. Autorité administrative indépendante française, chargée de veiller à ce que l'informatique soit mise au service du citoyen et qu'elle ne porte atteinte, ni à l'identité humaine, ni aux droits de l'homme, ni à la vie privée, ni aux libertés individuelles et publiques.

**CNN.** *Conseil national du numérique*. Commission consultative française créée le 29 avril 2011, chargée d'étudier les questions relatives à l'univers numérique, en particulier les enjeux et les perspectives de la transition numérique de la société, de l'économie, des organisations, de l'action publique et des territoires.

**CNPEN.** *Comité national pilote d'éthique du numérique*, créé par le Premier ministre en décembre 2019 et provisoirement placé sous l'égide du Comité consultatif national d'éthique (CCNE).

**CSNP.** *Commision supérieure du numérique et des postes*. Commission extra-parlementaire, composée de sept sénateurs, sept députés et trois personnalités qualifiées, chargée d'émettre des avis sur les projets de textes législatifs et réglementaires concernant ses deux domaines de compétence.

***Cognitive warfare*.** Voir **Guerre cognitive**.

**Commandement Allié Transformation.** Commandement interallié de l'OTAN, chargé des questions doctrinales et de transformation. Avec le Commandement Allié Opérations, il est l'un des deux commandements militaires stratégiques de l'OTAN.

**Commission de l'intelligence artificielle.** Instituée par la Première ministre Élisabeth Borne en septembre 2023, cette commission a remis son rapport au Président de la république en mars 2024, formulant 25 propositions pour faire de la France un acteur majeur de l'IA.

**Comportement inauthentique coordonné.** Voir **CIB**.

**Cryptomonnaie.** Monnaie électronique émise de pair à pair, sans nécessité d'une banque, d'une banque centrale ni d'aucun intermédiaire, utilisable au moyen d'un réseau informatique décentralisé. Elle est basée sur une blockchain intégrant des technologies de cryptographie, pour les processus d'émission et de règlement des transactions. Les cryptomonnaies les plus utilisées sont le Bitcoin et l'Ethereum.

**CSE.** *Conseil supérieur de l'éducation*, au Québec.

**Cybersécurité.** Ensemble des moyens utilisés pour veiller à la sécurité des systèmes et des données informatiques d'un État, d'une entreprise et, plus généralement, de toute forme d'organisation.

**DALLE-E.** Programme d'intelligence artificielle générative développé par la société OpenAI, capable d'engendrer des images à partir de descriptions textuelles.

***Data science.*** Science des données, en français, elle constitue un vaste domaine interdisciplinaire à la croisée des mathématiques, de la statistique, du calcul scientifique et de l'informatique, ayant pour objet d'extraire et d'extrapoler des connaissances à partir de grandes quantités de données, structurées ou non.

***Deep fake*.** Désigne une forme sophistiquée de *fake news*, engendrée au moyen de l'intelligence artificielle.

***Deep learning.*** Voir **Apprentissage profond**.

**Désinformation.** Information manipulée dans le but de tromper. La désinformation est une forme de mésinformation et, *a fortiori*, de malinformation.

**Distanciation cognitive.** Attitude et comportement consistant à prendre du recul par rapport à des informations reçues et à s'assurer de leur véracité avant de les prendre en considération.

**DMA.** *Digital Markets Act* ou **Règlement sur les marchés numériques**. Frère jumeau du Règlement sur les services numériques (DSA), il a été émis en septembre 2022 et vise principalement à s'assurer qu'aucune très grande plateforme en ligne se trouvant en position de « contrôleur d'accès » vis-à-vis d'un grand nombre d'utilisateurs n'abuse de cette position au détriment d'autres entreprises souhaitant accéder à ces mêmes utilisateurs.

**Dow Jones.** Indice du *New York Stock Exchange*, reposant sur la capitalisation boursière de trente des plus grosses entreprises cotées aux États-Unis. Créé en 1896, il est l'un des plus vieux indices boursiers du monde.

**DSA.** *Digital Services Acts* ou **Règlement européen sur les services numériques**. Frère jumeau du règlement sur les marchés numériques, il a été émis le 19 octobre 2022 et vise principalement à lutter contre la diffusion de contenus illicites à travers les très grandes plateformes en ligne.

**Économie de l'attention.** Mécanisme caractéristique de l'économie des médias, consistant à capter et retenir l'attention des consommateurs afin d'augmenter les ressources publicitaires.

**EDMO.** *European Digital Media Observatory*. **L'Observatoire des médias numériques** est un collectif européen, soutenant des réseaux de recherche travaillant sur la désinformation.

**EGI.** *États généraux de l'information*. Structure délibérative et participative, créée à l'initiative du Président de la République en septembre 2023 et animée par un comité de pilotage. Elle a rendu ses conclusions en septembre 2024.

***Embedding.*** Procédé par lequel un réseau neuronal artificiel se forge une représentation des données au sein d'un « espace latent » de dimension beaucoup plus réduite que le nombre des paramètres du modèle.

**EMI.** *Éducation aux médias et à l'information*.

**Espace latent.** Espace virtuel contenant les représentations synthétisées des données que construit un LLM, au sein duquel le modèle travaille en vue d'engendrer des résultats.

***Fact checkers*.** Les vérificateurs de faits mobilisent un ensemble de techniques pour vérifier la véracité des faits et l'exactitude des chiffres présentés sur les médias par des personnalités politiques ou par des experts, et pour évaluer le degré d'objectivité dans le traitement de l'information.

***Fake news.*** Terme équivalent à infox ou à désinformation.

***Fine tuning.*** Réglage fin d'un LLM à partir de bases de données restreintes, sélectionnées afin d'améliorer la qualité des réponses à des questions spécifiques.

**GPU.** *Graphic processing unit*. Circuit électronique capable d'effectuer des calculs mathématiques à grande vitesse. Initialement conçus pour générer les images animées sur les écrans informatiques, ils sont maintenant aussi massivement utilisés pour l'apprentissage automatique des LLMs.

**Guerre cognitive.** Forme de guerre reposant sur la manipulation des esprits.

**Hacker.** Pirate informatique.

**Hallucination.** Tendance d'un LLM à s'écarter de la réalité en fournissant des réponses fantaisistes aux questions posées. L'hallucination est contrôlable à travers un paramètre appelé « température ».

***Human oversight.*** Supervision humaine des systèmes d'intelligence d'artificielle, selon le principe : « un humain dans la boucle ».

**IA.** L'intelligence artificielle (IA) est un ensemble de théories et de méthodes visant à réaliser des machines capables de simuler l'intelligence humaine. Elle inclut les dispositifs imitant ou remplaçant l'homme dans la mise en œuvre de certaines de ses fonctions cognitives, comme le langage.

**IA curative.** Outils d'IA conçus pour lutter contre la mésinformation et la désinformation.

**IA falsificatrice.** Utilisation dévoyée d'outils d'IA, notamment de l'IA générative, à des fins de tromperie et de manipulation.

**IA générative.** Forme d'IA consistant à traiter un contenu soumis en entrée par l'utilisateur pour engendrer en sortie un contenu qui prolonge le plus vraisemblablement possible le contenu soumis.

**IGN.** *Institut national de l'information géographique et forestière*. Établissement public à caractère administratif, ayant pour mission d'assurer la production, l'entretien et la diffusion de l'information géographique de référence en France.

**INED.** *Institut national d'études démographiques*. Établissement public français spécialisé dans les recherches en démographie et les études de population en général.

**Infox.** Synonyme de désinformation ou de *fake news*.

**Inria.** *Institut national de recherche en sciences et technologies du numérique*.

**Insee.** *Institut national de la statistique et des études économiques*. Service d'Administration centrale, il est en charge depuis 1946 de la production, de l'analyse et de la publication des statistiques officielles de la France.

**JTI.** *Journalism Trust Initiative*. Standard des meilleures pratiques du journalisme, élaboré sous l'égide du Comité européen pour la normalisation (*European Committee for Standardization*) par un groupe de 130 experts comprenant des journalistes, des institutions, des organismes de régulation, des éditeurs, et des acteurs des nouvelles technologies.

**LLM.** *Large Language Model*, ou *langageur*. Modèle d'IA générative dont l'entrée est un texte rédigé par l'utilisateur et la sortie est un texte censé prolonger le plus vraisemblablement celui fourni en entrée.

**Logiciel.** Ensemble constitué de séquences d'instructions interprétables par une machine et d'un jeu de données nécessaires à ces opérations. Le logiciel détermine les tâches qui peuvent être effectuées par la machine, ordonne son fonctionnement et lui procure ainsi son utilité fonctionnelle.

**Loi infox.** Loi française promulguée le 22 décembre 2018, visant à lutter contre la manipulation de l'information.

**Malinformation.** Désigne toute forme d'imperfection de l'information, que celle-ci soit incomplète, partisane, erronée ou falsifiée. La malinformation inclut la mésinformation, qui elle-même inclut la désinformation.

**Mécanisme d'attention.** Mécanisme interne des LLMs par lequel ces modèles interprètent la question et le *prompt* qui leur sont soumis et construisent une représentation de ces éléments dans l'espace latent.

**Médialab Sciences Po.** Laboratoire interdisciplinaire de Sciences Po, qui mène des recherches thématiques et méthodologiques visant à étudier les relations entre l'espace numérique et nos sociétés.

**Médialab AFP.** Petite équipe de journalistes et d'ingénieurs au sein de l'AFP, qui développe des outils destinés aux salles de rédaction permettant de vérifier les images et les vidéos diffusées sur Internet et les réseaux sociaux.

**Mésinformation.** Désigne des informations, soit involontairement erronées, soit délibérément falsifiées ; dans ce dernier cas, on parle de désinformation.

**Météo France.** Établissement public à caractère administratif, notamment chargé de l'observation, la prévision et l'étude des phénomènes météorologiques, ainsi que de la vigilance météorologique pour les territoires français de métropole et d'outre-mer.

**NCSC.** *National Cyber Security System*. Organisme public britannique fournissant un soutien aux secteurs public et privé dans la prévention des menaces de sécurité informatique.

**NLP.** *Natural Language Processing*. Le traitement automatique des langues naturelles (TALN) est un domaine multidisciplinaire, impliquant la linguistique, l'informatique et l'intelligence artificielle, qui vise à créer des outils de traitement du langage naturel pour diverses applications, comme la reconnaissance de l'écriture, celle de la parole, ou la synthèse vocale.

**OEA (prospectif).** *Observatoire de l'édition artificielle*. Proposition d'identifier et de publier les « lignes éditoriales » des LLMs les plus populaires.

**OpenAI.** Entreprise américaine d'intelligence artificielle fondée en 2015 à San Francisco, affichant pour ampbition le développement d'une IA générale « sûre et benéfique à toute l'humanité ». En novembre 2022 elle a lancé ChatGPT, déclenchant ainsi un intérêt mondial pour les agents conversationnels et l'IA générative. Elle a par ailleurs créé le modèle de génération d'images DALL-E, disponible dès 2021.

**Otan.** *Organisation du Traité de l'Atlantique Nord*.

**Phármakon.** Chez les Grecs anciens, ce terme peut aussi bien désigner le poison, le remède, et le bouc émissaire. L'IA générative se prête assez naturellement à chacune de ces trois acceptions !

**Pharos.** *Plateforme d'harmonisation, d'analyse, de recoupement et d'orientation des signalements*. Site créé en 2009 par le Gouvernement français, pour signaler des contenus et des comportements en ligne illicites.

**Plateforme numérique.** Infrastructure en ligne basée sur un logiciel, qui facilite les interactions et les transactions des utilisateurs ou la recherche d'informations. Les moteurs de recherche, les réseaux sociaux, les communautés en ligne, les sites d'achat et vente en ligne, sont tous des exemples de plateformes numériques.

***Prompt.*** Partie de la requête formulée par un utilisateur à un LLM, qui accompagne la question posée en précisant le contexte de formulation de cette question, ainsi que le style et le format de la réponse souhaitée.

***Prompt engineering.*** Technique empirique de rédaction efficace d'un *prompt*, afin que la machine apporte la meilleure réponse possible à la question posée.

***Pump and dump.*** Technique frauduleuse sur les marchés financiers, consistant à faire monter artificiellement le cours d'une valeur, avant de le faire s'effondrer brusquement dans le but de réaliser une plus-value et d'infliger une perte à des investisseurs crédules.

**RAG.** *Retrieval Augmented Generation*. La génération augmentée de récupération est une technique récente d'IA, permettant à un LLM d'exploiter des ressources de données supplémentaires, sans réentraînement.

***Red team.*** L'équipe rouge, consistant à simuler des attaques, est une méthode permettant de tester et d'améliorer la sécurité d'un système, notamment informatique.

**Réseau neuronal artificiel.** Système informatique central de l'IA, dont la conception est schématiquement inspirée du fonctionnement des neurones biologiques du cerveau humain, et qui est optimisé par des méthodes d'apprentissage statistique.

**Réseau neuronal convolutif.** Réseau neuronal dans lequel le motif des connexions entre neurones est inspiré par le cortex visuel des animaux, composé de régions qui se chevauchent pour former un pavage du champ visuel. Des petites quantités de données sont prétraitées dans chaque région élémentaire, puis assemblées en vue d'accéder à une vision d'ensemble.

**Réseau social.** Site Internet ou application mobile permettant de pratiquer des interactions sociales, de se constituer un réseau de connaissances, d'amis ou de relations professionnelles, avec lesquelles on peut échanger en temps réel.

**Risque systémique.** Risque qu'un évènement particulier entraîne, par réaction en chaîne, des effets dommageables considérables sur l'ensemble d'un système, pouvant occasionner une crise générale de son fonctionnement.

***Rules of speech.*** Règles de bon usage, imposées par les administrateurs des réseaux sociaux.

**RSF.** *Reporters sans frontières* est une organisation non gouvernementale internationale fondée en 1985, reconnue d'utilité publique en France, et présente dans quatorze pays en 2020. Elle se donne pour objectif la défense de la liberté de la presse et la protection des sources des journalistes.

**Score d'artificialité (prospectif).** Proposition de contraindre les grandes plateformes à afficher, pour les contenus les plus viraux, une estimation de leur présomption d'inauthenticité, au double regard de leur production et de leur diffusion.

**Sentiment de marché.** Attitude collective des investisseurs et des autres participants aux marchés financiers, résultant de sentiments individuels postés sur les réseaux sociaux.

**SGDSN.** *Secrétariat Général de la Défense et de la Sécurité Nationale*. Placé auprès du Premier Ministre, il supervise notamment l'Anssi et Viginum.

**SMC.** *Science Media Centre*. Société caritative britannique, fondée en 2002 à la suite d'un rapport émis en 2000 par le comité spécial « science et technologie » de la Chambre des Lords, sur le thème « Science et société ». L'objectif du Centre est de promouvoir un meilleur traitement de la science dans les médias, en fournissant aux journalistes des informations de fond sur les questions scientifiques d'actualité et en facilitant les entretiens avec les experts.

**Spinoza.** Projet lancé en 2023 par Reporters sans frontières et l'Alliance de la presse d'information générale, en vue de développer un outil d'IA destiné aux journalistes.

***Spreaders.*** Diffuseurs de désinformation.

**Stylométrie.** Ensemble de techniques permettant de reconnaître des styles caractéristiques dans des messages textuels.

***Superspreaders.*** Diffuseurs massifs de désinformation.

**Système de valeurs.** Module d'un LLM le dotant de règles éthiques à respecter pour rédiger ses réponses aux requêtes des utilisateurs, et le conduisant éventuellement à refuser (poliment !) de répondre.

**TALN.** Traitement automatique des langues naturelles, voir **NLP**.

**Télégraphe Chappe.** Moyen de télégraphie visuelle par sémaphore, inventé par Claude Chappe en 1794. En 1844, plus de 500 tours quadrillaient le territoire français, dont subsiste aujourd'hui environ une vingtaine.

**Température.** Paramètre réglable d'un LLM qui permet de contrôler sa tendance à halluciner, en fixant le plancher de vraisemblance en dessous duquel un *token* n'est pas admissible pour succéder au *token* précédent.

**Test de Turing.** Test consistant à déterminer si, dans une interaction à visage masqué, on a affaire à un humain ou à une machine. Une machine réussit le test si elle parvient à se faire passer pour un humain.

**THF.** Trading à haute fréquence.

***Token.*** Les mots d'une langue sont décomposables en *tokens*, unités infra-lexicales sur lesquelles opère un LLM pour construire des phrases. Une langue donnée comportant moins de *tokens* distincts que de mots, ce procédé permet de réduire considérablement la puissance de calcul nécessaire.

**Trader.** Opérateur de marché, ou opérateur financier. Négociateur de produits financiers, il peut être un professionnel travaillant dans une salle de marchés pour le compte d'une banque ou d'un fonds d'investissement, ou bien un particulier agissant pour son compte propre.

**Trading algorithmique.** Aussi appelé **trading automatisé**, il s'agit d'une forme de trading utilisant des plateformes électroniques pour la saisie des ordres de bourse, le plus souvent sans aucune intervention humaine.

**TsuNumi.** Néologisme dû à Nicolas Curien, formé avec les mots « tsunami » et « numérique », pour désigner le déferlement torrentiel des informations dans l'espace numérique.

**vera.ai.** *Verification assisted by artificial intelligence*. Plateforme européenne qui fournit notamment une batterie d'outils de détection des contenus artificiels engendrés par l'IA, sur tous types de supports.

**Vérificateur de confiance.** Il évalue la fiabilité et la crédibilité des sources d'informations, une tâche complémentaire de celle du vérificateur de faits, examinant quant à lui la véracité des informations.

**Vérificateur de faits.** Voir ***Fact checkers***.

**Volatilité.** Ampleur des variations du cours d'un actif financier. Elle permet de quantifier le risque de rendement et de déterminer le prix de cet actif.

**Viginum.** Service français à compétence nationale créé en 2021, rattaché au SGDSN et chargé de la lutte contre les manipulations de l'information émanant directement ou indirectement de puissances étrangères.

**VLOP.** *Very Large Online Platform*, très grande plateforme en ligne, selon la terminologie de la Commisssion européenne.

**VLOSE.** *Very Large Online Search Engine*, très grand moteur de recherche en ligne, selon la terminologie de la Commission européenne.

***Watermarking*.** La technique du « tatouage numérique » permet d'ajouter des informations de *copyright* ou d'autres messages de vérification à un document numérique, qu'il s'agisse d'un écrit, d'une image, d'une vidéo ou d'un enregistrement audio.

***World Economic Forum (WEF)*.** Le Forum économique mondial, est une fondation à but non lucratif créée en 1971. Il réunit chaque année à Davos des patrons de multinationales, des banquiers, des milliardaires et des intellectuels influents du monde entier.

## C. Liste des membres du groupe de travail

*Nota bene* important. La déontologie académique exige que tout contributeur à l'élaboration d'un rapport n'apporte au collectif que sa seule expertise, en se gardant de promouvoir tout intérêt personnel ou corporatif. En vue d'une parfaite transparence, précisons ici que certains membres du groupe, en raison de leur position professionnelle, se sont tout particulièrement pliés à cette discipline, notamment Éric BIERNAT, Delphine ERNOTTE-CUNCI, Marko ERMAN, Chantal JOUANNO.

Stéphane ANDRIEUX, directeur scientifique général de l'Onera, membre de l'Académie des technologies.

Maurice BELLANGER, professeur émérite du Conservatoire national des arts et métiers, membre de l'Académie des technologies.

Éric BIERNAT, directeur Big Data Analytics, Octo Technology.

Pierre-Étienne BOST, ancien directeur de recherche à l'Institut Pasteur, membre de l'Académie des technologies.

Alain-Michel BOUDET, professeur honoraire à l'Institut universitaire de France, membre de l'Académie des technologies.

François BOURDONCLE, président de FB&Cie, membre de l'Académie des technologies.

Alain BRAVO, ancien directeur de Supelec, ancien président de l'Académie des technologies.

Alain CADIX, président du Comité Éthique, technologie et société de l'Académie des technologies.

Thierry CHEVALIER, CTO Strategy on Digital Design Manufacturing & Services, Airbus.

Nicolas CURIEN, professeur émérite du Conservatoire national des arts et métiers, membre fondateur de l'Académie des technologies, pilote du groupe de travail « IA générative et mésinformation ».

Jean-Pierre DUPUY, professeur émérite de l'École polytechnique, membre de l'Académie des technologies.

Marko ERMAN, senior vice-président chez Thalès, membre de l'Académie des technologies.

Delphine ERNOTTE-CUNCI, présidente de France Télévisions, membre de l'Académie des technologies.

Hervé GALLAIRE, ancien président recherche et technologie du groupe Xerox, membre de l'Académie des technologies

Laurent GOUZÈNES, président de KM2 Conseil.

Joël HARTMANN, président de Delaurean Consulting, membre de l'Académie des technologies.

Chantal JOUANNO, Managing director chez Accenture, ancienne ministre, membre de l'Académie des technologies.

Étienne KLEIN, directeur de recherches au CEA, membre de l'Académie des technologies.

Claude LE PAPE-GARDEUX, vice-président Intelligence artificielle, optimisation fiabilité et analytics chez Schneider Electric, membre de l'Académie des technologies.

Manoëlle LEPOUTRE, ancienne directrice de la R&D de l'exploration-production du groupe TOTAL, vice-présidente de l'Académie des technologies.

Jacques LUKASIK, ancien directeur scientifique du groupe Lafarge, membre de l'Académie des technologies.

Winston MAXWELL, directeur de recherche à Télécom Paris, Institut polytechnique de Paris.

Jean-Luc MOLINER, ancien directeur de la sécurité du groupe Orange, membre de l'Académie des technologies.

Brigitte PLATEAU, professeur émérite à l'INP-UGA Grenoble, membre de l'Académie des technologies.

Grégoire POSTEL-VINAY, Conseil général de l'économie, membre de l'Académie des technologies.

Denis RANQUE, ancien président d'Airbus, ancien président de l'Académie des technologies.

Bruno REVELLIN-FALCOZ, ancien vice-président directeur général de Dassault Aviation, ancien président de l'Académie des technologies.

Gérard ROUCAIROL, ancien directeur scientifique du groupe Bull, ancien président de l'Académie des technologies.

Christian SAGUEZ, professeur honoraire à l'École centrale de Paris, membre de l'Académie des technologies.

Michèle SEBAG, directrice de recherche au CNRS, membre de l'Académie des technologies.

Joëlle TOLEDANO, membre du Conseil national du numérique, membre de l'Académie des technologies.

Xavier VAMPARYS, responsable IA éthique chez CNP Assurance, doctorant à Télécom Paris, Institut polytechnique de Paris.

Thierry WEIL, professeur à l'École des Mines de Paris, président du Comité des travaux de l'Académie des technologies.

## D. Liste des auditions

18/09/2023. Éric BIERNAT, directeur Big Data Analytics, Octo Teecnology.

16/10/2023. Tiphaine VIARD, maître de conférences, équipe Numérique Organisation & Société, Telecom Paris.

20/11/2023. Alexeï GRINBAUM, directeur de recherche, CEA-Saclay.

18/12/2023. Marko ERMAN, directeur de la recherche et de la technologie, Thalès & membre de l'Académie des technologies.

15/01/2024. Claire MATHIEU, directrice de recherche CNRS en informatique & membre de l'Académie des sciences.

19/02/2024. Sébastien MISOFFE, directeur général de Google France.

18/03/2024. Manon BERRICHE & Jean-Philippe COINTET, Médialab de Sciences Po.

15/04/2024. Mireille CLAPOT, ancienne vice-présidente de la CSNP (Commission supérieure du numérique et des Postes) & Guillaume ROZIER, Conseiller à la présidence de la République sur la stratégie numérique et les données publiques.

27/05/2024. Benoît LOUTREL, membre du Collège de l'Arcom, président du groupe de travail « Supervision des plateformes en ligne ».

17/06/2024. Marc-Antoine BRILLANT, chef du service Viginum, Secrétariat général de la défense et de la sécurité nationale.

15/07/2024. Julie CHARPENTRAT, adjointe à la Rédaction en chef investigation numérique à l'AFP, Denis TEYSSOU, responsable du Médialab de l'AFP & Kati BREMME, directrice Innovation et prospective, rédactrice en chef Méta-Média, France TV.

16/07/2024. Arthur GRIMONPONT et Pierre DAGARD, Reporters sans frontières.

16/09/2024. Isabelle FÉROC-DUMEZ, directrice scientifique et pédagogique au CLEMI & Jean CATTAN, secrétaire général du Conseil national du numérique.

25/09/2024. Fabrice RIVA, professeur de finance à l'université Paris-Dauphine.

**La mésinformation, qu'elle soit information involontairement erronée ou délibérément falsifiée, dans ce cas appelée désinformation, infox ou *fake news*, n'est pas un phénomène nouveau. Cependant, l'essor des technologies numériques et de l'IA a accru son impact de manière considérable.** Aujourd'hui, ces contenus faux, généralement plus attractifs que les vrais, se propagent de manière virale sur Internet, à la faveur des modèles économiques des grandes plateformes numériques, dont l'objectif est de maximiser les revenus publicitaires en captant l'attention des internautes. Classée, cette année, au premier rang des risques systémiques mondiaux à court terme selon le Forum économique mondial, cette dérive constitue une véritable menace pour l'honnêteté de l'information et donc pour la démocratie.

Dans ce rapport, l'Académie des technologies alerte sur l'accroissement du danger sous l'effet de la montée en puissance de l'IA générative et souligne le potentiel d'une IA curative pour lutter contre la désinformation. Forte de son analyse, elle propose six recommandations ciblées visant à renforcer notre résilience face aux *fake news* et à préserver la crédibilité de l'information. Un document essentiel pour comprendre et relever l'un des défis majeurs de notre époque.

Académie des technologies
Le Ponant – Bâtiment A
19, rue Leblanc
75015 PARIS
+33(0)1 53 85 44 44
secretariat@academie-technologies.fr
www.academie-technologies.fr

ISBN : 979-10-97579-55-5


Couverture : Noé - Adobe stock /
Image générée avec l'aide d'une IA