Intelligence artificielle générative et création de contenu informationnel

I. L'intelligence artificielle, nouvel émetteur de contenus

L'année qui vient de s'écouler a été marquée par la démocratisation des outils d'intelligence artificielle générative. Ces outils ne sont pas nouveaux. En revanche, leur utilisation par le grand public l'est. ChatGPT a marqué l'année 2023 en ce sens, cumulant 1 million d'utilisateurs en seulement cinq jours après son lancement. À titre de comparaison, il avait fallu 2 mois et demi à Instagram pour arriver au même nombre, 5 mois à Spotify, 10 mois à Facebook et 2 ans à Twitter (devenu X)[1].

L'IA générative se définit comme une technologie permettant de générer des contenus en réponse à une requête (aussi appelé "prompt") d'un utilisateur[2]. Ces outils sont adossés à modèles d'IA dits de "fondation" : il s'agit de modèles de grande taille, entraînés sur une vaste quantité de données non étiquetées et pouvant ainsi s'adapter à un large éventail de tâches[3], dont la création de contenus. Ces contenus peuvent être de différents types :

- Des contenus textuels produits par des assistants conversationnels sous la forme de *chatbots* soit des "*systèmes conçus pour des conversations prolongées imitant les conversations non structurées ou les « chats » caractéristiques de l'interaction humain-humain*"[4], comme ChatGPT proposé par OpenAI ou Bard proposé par Google. Ces outils utilisent sont entraînés sur de "grands modèles de langage" (*large langage model*, LLM) ;
- Des contenus audios produits par le clonage ou la déformation de voix réelles afin de faire dire à un individu quelque chose qu'il n'a pas dit ;
- Des images et des vidéos produits par des services comme Midjourney, Dall-E proposé par OpenAI ou Stability AI.

Ces nouveaux modes de production de contenus informationnels emportent des enjeux à trois égards :

1. Ils créent le risque d'une prolifération de contenus erronés prenant l'apparence de contenus fiables ;
2. Ils offrent de nouveaux outils, aisément accessible et à faible coût, aux acteurs mal intentionnés ;
3. Ils modifient en profondeur les modes de réception et de diffusion de l'information, interrogeant la notion même d'espace informationnel public commun.

Ces trois enjeux seront passés tour à tour en revue.

---

[1] https://legrandcontinent.eu/fr/2023/09/26/chatgpt-perd-des-utilisateurs-et-annonce-une-innovation-majeure/

[2] https://www.europarl.europa.eu/RegData/etudes/BRIE/2023/757583/EPRS_BRI(2023)757583_EN.pdf

[3] https://fr.wikipedia.org/wiki/Mod%C3%A8le_de_fondation

[4] FAVRO, Karine, VILLATA, Serena, et ZOLYNSKI, Célia. Des assistants vocaux aux agents conversationnels. Vers un encadrement des interfaces vocales humain-machine. *Dalloz IP/IT: droit de la propriété intellectuelle et du numérique*, 2023, no 09, p. 459.

1. Vers un web synthétique ?

Dans une tribune au Monde parue à l'automne 2023, Olivier Ertzscheid, enseignant-chercheur en science de l'information et de la communication dresse le constat de l'**avènement d'un "web synthétique"** :

> "*On trouve partout et en masse des robots, des agents conversationnels, des algorithmes de contrôle assumant des effets d'éditorialisation désormais impossibles à nier. Les études scientifiques et techniques indiquent depuis déjà quelques années que presque la moitié du trafic de l'Internet mondial est causée par des robots.*
>
> *Par-delà le trafic, c'est désormais aussi dans la production d'informations (vraies ou fausses) et dans les interactions en ligne au sein des médias sociaux que la part des robots est de plus en plus importante et pourrait rapidement nous placer en situation de naviguer, de discuter et d'interagir au moins autant – et peut-être demain presque exclusivement – avec des programmes informatiques qu'avec des individus.*
> *[...]*
> *C'est un Web synthétique, car l'essentiel de ce qui y est discuté comme de ce qui y est vu est produit par la synthèse d'algorithmes et d'agents tout aussi artificiels, convergeant au service d'intelligences artificielles opaques, qui, malgré leur apparence de rationalité, ne sont rien d'autre que des perroquets stochastiques [qui ne comprennent pas vraiment], opérant le plus souvent au service d'intérêts économiques.*
>
> *C'est un Web synthétique, enfin, parce qu'il est l'objet d'une nouvelle inflation et d'une nouvelle inertie : l'ensemble des contenus produits ou organisés par les robots et les algorithmes permet de produire d'autres contenus et d'autres robots, qui eux-mêmes viennent nourrir d'immenses jeux de données utilisés par des technologies d'IA pour produire toujours plus de contenus artificiels ou artificialisés, qui à leur tour… et ainsi de suite. Ad libitum.*"[5]

On voit donc clairement à quel point ce changement de paradigme emporte des conséquences majeures en ce qui concerne notre espace informationnel.

En ce qui concerne les textes générés par des agents conversationnels basés sur de grands modèles de langages, il est important de rappeler que ces derniers ne produisent pas de raisonnement mais sont des systèmes probabilistes : ces agents conversationnels ne font que "*calculer des probabilités : en fonction de l'instruction entrée par l'utilisateur, le système calcule la probabilité du premier mot de sa réponse, puis du suivant et ainsi de suite*."[6] Or, ce calcul s'opère "*sans égard pour la signification du texte généré, ni pour sa pertinence ni pour sa vérité*"[7]. Ceci peut avoir pour conséquence que, **régulièrement, l'agent conversationnel "hallucine" et propose une réponse contenant des informations inventées sans pour autant préciser que la réponse n'est pas forcément totalement exacte**. Les exemples sont nombreux, chacun peut en faire l'expérience dans ses usages de ces outils. En France, le député Eric Bothorel a par exemple déposé en avril 2023 une saisine contre OpenAI, accusant son service ChatGPT de diffuser des informations incorrectes à son égard, en violation du Règlement général sur la protection des données personnelles (RGPD)[8].

[5] https://www.lemonde.fr/idees/article/2023/10/07/intelligence-artificielle-nous-sommes-passes-du-reve-d-un-web-semantique-a-la-realite-d-un-web-synthetique_6192914_3232.html
[6] https://cnnumerique.fr/paroles-de/chatgpt-rupture-ou-continuite-technologique-un-echange-avec-laure-soulier
[7] https://cnnumerique.fr/paroles-de/comment-penser-chatgpt
[8] https://www.usine-digitale.fr/article/un-depute-francais-depose-une-plainte-contre-chatgpt-aupres-de-la-cnil.N2120486

Comme le résume le Comité national pilote d'éthique du numérique (CNPEN) : "*Les systèmes d'IA générative fonctionnent uniquement avec des représentations numériques, sans appréhender la signification des mots pour les êtres humains. La signification est uniquement celle que les humains projettent sur les résultats, car seuls les humains en possèdent une interprétation dans le monde réel.*"[9] Par exemple, OpenAI indique en bas de l'interface de chatGPT : "*ChatGPT may produce inaccurate information about people, places, or facts* » (en français, "ChatGPT est susceptible de produire des informations inexactes sur des personnes, des lieux, ou des faits)[10]. Pour autant, maintenir une distance et un regard critique vis-à-vis des textes proposés est particulièrement difficile. **Toute la complexité réside dans ce paradoxe entre l'apparente exactitude des informations proposées par ces services, dans un langage naturel aisément compréhensible, et le fond souvent inexact voire faux**. Les messages d'avertissement sont faciles à ignorer et il est d'autant plus compliqué de vérifier les informations partagées et de repérer ces hallucinations que ces agents conversationnels ne citent pas leurs sources[11], posant une réelle question de traçabilité de l'information (voir fiche dédiée). Les créateurs de ces outils reconnaissent d'ailleurs ces carences. Au-delà, cela implique aussi et surtout de comprendre précisément leur fonctionnement probabiliste pour saisir que le contenu créé n'est pas nécessairement synonyme de véracité.

De plus, parce que ces systèmes sont entraînés sur les données accessibles en ligne, elles peuvent aussi propager des narratifs faux qui préexistent sur Internet. Un audit mené par NewsGuard en août 2023 sur ChatGPT Plus et Bard révèle que dans 98 % des cas pour le premier et 80 % des cas pour le second, l'agent conversationnel se fait le relais des mythes proposés par l'utilisateur, au lieu de les contredire[12]. L'audit montre également que l'outil ChatGPT Plus est souvent plus persuasif que Bard, adjoignant moins de mises en garde à ses réponses que l'outil proposé par Google.

Au-delà des agents conversationnels, l'IA générative peut également être à la source de sites internet totalement nourris de contenus artificiels, faisant concurrence aux médias traditionnels. Le nombre de ces sites a explosé ces dernières années. Alors que NewsGuard dénombrait 61 "sites d'actualité non fiables générés par l'IA" en 2021, l'organisation en recensait 651 en janvier 2024, dans 15 langues différentes[13]. À l'occasion de l'édition 2023 de Médias en Seine, Chine Labbé - rédactrice en chef et vice-Présidente chargée des Partenariats, Europe et Canada chez NewsGuard - pointait une évolution de ces sites Internet : alors qu'ils étaient initialement créés surtout dans le but de générer du clic et des revenus publicitaires avec des contenus relativement vides, ces sites Internet artificiels relaient de plus en plus d'infox particulièrement virales[14]. Exemple récent en date, le site généré par IA Global Village Space a relayé une information selon laquelle le psychiatre de Beyamin Netanyahou se serait suicidé suite aux prises de position de ce dernier[15]. Cette information a été ensuite relayée par la chaîne 2 de la télévision publique de la République islamique d'Iran qui poursuit depuis plusieurs semaines une stratégie de dénigrement d'Israël en accusant Netanyahou d'être psychologiquement instable.

Les générateurs d'images et de vidéos artificielles peuvent également être à l'origine de fausses informations propagées en ligne. Les exemples sont là aussi nombreux. L'image emblématique à cet égard est celle du Pape François portant une doudoune générée par l'outil

[9] https://www.ccne-ethique.fr/sites/default/files/2023-09/CNPEN_avis7_06_09_2023_web-rs2.pdf
[10] https://www.ccne-ethique.fr/sites/default/files/2023-09/CNPEN_avis7_06_09_2023_web-rs2.pdf
[11] https://www.vice.com/en/article/akex34/chatgpt-is-a-bullshit-generator-waging-class-war
[12] https://www.newsguardtech.com/fr/special-reports/openai-et-google-bard-continuent-de-propager-des-informations-erronees/
[13] https://www.newsguardtech.com/fr/special-reports/ia-centre-de-suivi/
[14] https://www.mediasenseine.com/fr/replay/l-industrie-de-la-desinformation-en-acceleration-avec-l-ia/
[15] https://www.newsguardtech.com/fr/special-reports/un-site-genere-par-lia-lance-une-infox-virale-affirmant-que-le-pretendu-psychiatre-de-netanyahou-sest-suicide/

Midjourney[16]. Plus récemment, une vidéo montrant la Tour Eiffel en feu a cumulé plus de 200 millions de vues en 48 heures sur TikTok. S'il était jusqu'à récemment encore relativement aisé de repérer ces contenus artificiels, en observant notamment les extrémités des corps représentés (mains, oreilles…), ces outils se perfectionnent très rapidement et il devient de plus en plus difficile de faire la part du vrai et du faux. Les banques d'images sont également concernées. Par exemple, la banque d'image d'Adobe, particulièrement fournie, est très utilisée par les médias n'ayant pas la possibilité d'envoyer un photographe en reportage, notamment dans les zones de guerre. En ce qui concerne actuellement le conflit entre Israël et le Hamas, *L'ADN* rapport que la banque d'images d'Adobe contient de nombreux contenus hyperréalistes[17] montant des rues bombardées, des enfants dans les décombres ou encore des manifestations, toutes ayant été générées par Firefly, l'outil de génération d'image d'Adobe. S'il est précisé par Adobe que ces images sont générées artificiellement, cette mention est discrète et passe souvent inaperçue, ayant pour conséquence que l'image soit ensuite repartagée par les médias sans mention des crédits, laissant croire à une photo authentique.

À l'inverse, **l'argument selon lequel une image a été générée par IA est de plus en plus utilisé pour saper la confiance dans de réelles images**. Toujours dans le conflit entre Israël et le Hamas, la photo du corps d'un bébé israélien brûlé a beaucoup circulé sur les réseaux sociaux et a été à plusieurs reprises pointées étant une fausse image générée par IA[18]. Pour Hany Farid, professeur à l'École d'Information de l'Université de Berkeley, en Californie, interrogé par le le site 404 Media : "*ne présente aucun signe de création par l'IA*" : "*Les cohérences structurelles, les ombres précises, l'absence d'artefacts que nous avons tendance à voir dans l'IA — tout cela m'amène à penser qu'elle n'a pas été générée par l'IA, pas même partiellement*"[19]. NewsGuard prolonge cette analyse en soulignant que "*en date du 20 novembre, le post X ne comportait pas de note de la communauté, ce qui signifie que Jackson Hinkle était probablement éligible au partage des recettes publicitaires*".

Les outils d'IA génératives marquent également une nouvelle avancée dans les hypertrucages (ou *deepfakes*), c'est-à-dire "une falsification « hyper-réelle » d'images, de vidéos ou de fichiers audios, effectuée à l'aide d'algorithmes, apposant l'image et/ou la voix d'une personne sur une autre personne afin de lui faire faire ou faire dire des choses, qu'elle n'a en réalité jamais faites ni dites"[20]. Là encore, ces contenus ne sont pas nouveaux. Cependant, ils sont désormais accessibles à tous facilement et gratuitement. **S'ils sont utilisés fréquemment à des fins humoristiques, ils peuvent aussi avoir des conséquences particulièrement néfastes sur les individus et les démocraties**. Au niveau individuel, les *deepfakes* sont particulièrement utilisés pour créer des contenus pornographiques : 96 % des deepfakes sont des contenus pornographiques non-consensuels, représentant quasi-exclusivement des femmes[21], mettant notamment en scène des célébrités, Taylor Swift en a par exemple été la victime en début d'année 2024[22], présentant des conséquences évidentes en matière d'atteinte à la vie privée et à la dignité des personnes visées. Au niveau politique, les élections législatives slovaques qui se sont tenues à l'automne 2023 ont été la cible de ces trucages hyperréalistes : Michal Simecka, leader du parti Slovaquie progressiste et donné en tête des sondages a été victime d'un hypertrucage audio dans lequel on l'entendait échanger avec

[16] https://www.francebleu.fr/infos/insolite/les-images-du-pape-francois-en-doudoune-devenues-virales-sont-completement-fausses-5922778
[17] https://www.ladn.eu/media-mutants/images-generees-ia-guerre/?utm_source=newsletter_ladn&utm_medium=email&utm_campaign=news_ladn_tendance&utm_content=20231113
[18] https://www.newsguardtech.com/fr/misinformation-monitor/novembre-2023/
[19] https://www.404media.co/ai-images-detectors-are-being-used-to-discredit-the-real-horrors-of-war/, partagé et traduit par NewsGuard
[20] https://cnnumerique.fr/files/uploads/2021/CNNum_Dossier-Recits-et-contre-recits-itineraire-des-fausses-informations-en-ligne.pdf
[21] https://blogs.mediapart.fr/collectif-de-personnalites/blog/041223/ia-l-autoregulation-des-modeles-de-fondation-mettrait-en-danger-les-droits-humains
[22] https://www.wired.com/story/taylor-swift-deepfake-porn-artificial-intelligence-pushback/

un journaliste quant à la façon de truquer les élections en achetant des voix à la minorité Rom du pays et faire de l'humour sur la pédopornographie[23]. Cet enregistrement a été identifié comme inauthentique par les équipes de fact checkers de l'AFP mais a tardé à être modéré sur les réseaux sociaux en raison des carences des outils algorithmiques de modération en langue slovaque. S'il est difficile d'établir un lien de causalité clair entre cette manipulation et le résultat de l'élection, il n'en demeure pas moins que le candidat s'est finalement incliné derrière Robert Fico, candidat de la gauche nationaliste et populiste. Avec moins de conséquences, Emmanuel Macron est également fréquemment la cible de manipulations audios. Une allocution générée artificiellement le fait annoncer le remplacement de l'anglais par l'arabe dès la maternelle avant de s'exprimer lui-même en arabe[24]. Dans une autre vidéo, on peut aussi le voir annoncer sa démission de la présidence de la République[25].

Enfin, les outils d'IA n'ont pas seulement un impact sur la création de contenus pouvant être erronés, trompeurs ou manipulateurs, mais également sur leur diffusion. **Ces technologies permettent en effet de créer de faux comptes sur les réseaux sociaux pour relayer massivement ces contenus.** Ces stratégies sont appelées "**astroturfing**", soit le fait de "permettre à un nombre restreint d'acteurs de donner l'impression que la campagne a un impact aussi puissant que si elle avait été relayée sans artifice par un plus grand nombre de personnes."[26] Par exemple, aux États-Unis, l'entreprise Devumi aurait vendu plus de 200 millions de faux comptes sur Twitter et 3,5 millions sur Facebook entre 2018 et 2019 avant d'être visée par la justice américaine pour usurpation d'identité et tromperie et de clôre ses activités[27]. Ces faux comptes sont particulièrement utilisés par la Russie dans le cadre de ses stratégies d'influence en ligne et de ses fermes à trolls. La Chine y a également recours pour noyer les contenus indésirable en ligne avec des contenus propagandistes, plutôt que de les supprimer[28]. **Il importe donc de penser également des mécanismes de détection et de suppression de ces comptes artificiels**.

2. De nouveaux outils pour les acteurs mal intentionnés

Camille François, enseignante à l'Université de Columbia et chercheuse affiliée à l'Institut français de Géopolitique de l'Université Paris-8 et au Berkman-Klein Center for Internet & Society d'Harvard, propose d'analyser la diffusion des fausses informations en ligne à travers une grille baptisée "***ABC Framework***". Ce terme permet de qualifier trois vecteurs : les "*manipulative Actors*" (acteurs manipulateurs), les "*deceptive Behaviors*" (comportements trompeurs) et les "*harmful Contents*" (contenus nocifs). Dans cette grille, il est indéniable que **les outils d'intelligence artificielle générative offrent de nouveaux outils (de nouveaux comportements) faciles à utiliser et peu coûteux aux acteurs mal intentionnés pour propager des contenus nocifs[29]**.

Là encore, les exemples sont nombreux. Aux États-Unis, la campagne électorale en vue de l'élection présidentielle de 2024 est elle aussi marquée par des contenus créés par IA. Au printemps 2023, les Républicains ont publié une vidéo anti-Biden entière réalisée avec des contenus générés par IA[30]. Il est à noter que la mention "réalisée entièrement par IA" est apposée en haut à gauche de la vidéo. Plus récemment, à l'occasion de la primaire Républicaine américaine dans l'État du New

[23] https://www.bloomberg.com/news/newsletters/2023-10-04/deepfakes-in-slovakia-preview-how-ai-will-change-the-face-of-elections
[24] https://factuel.afp.com/doc.afp.com.348D4KU
[25] https://factuel.afp.com/doc.afp.com.33WK8PE
[26] https://cnnumerique.fr/files/uploads/2021/CNNum_Dossier-Recits-et-contre-recits-itineraire-des-fausses-informations-en-ligne.pdf
[27] https://www.nytimes.com/interactive/2018/01/27/technology/social-media-bots.html
[28] https://repository.law.umich.edu/mlr/vol117/iss3/4/
[29] https://www.wired.com/story/400-dollars-to-build-an-ai-disinformation-machine/
[30] https://www.numerama.com/politique/1353952-cette-video-anti-biden-est-faite-par-ia-une-premiere-dans-une-campagne.html

Hampshire, des électeurs ont reçu un appel générés artificiellement imitant Joe Biden et appelant les électeurs à ne pas voter lors de la primaire[31]. À la suite de cet épisode, la *Federal Communications Commission* (l'autorité de régulation américaine des communications) a émis une décision interdisant aux États-Unis les appels robotisés utilisant des voix générées par IA[32].

En 2024, à travers le monde, près de la moitié de la population mondiale en âge de voter est appelée aux urnes. Les risques de manipulation numérique de ces scrutins sont nombreux, en particulier du fait des outils d'IA générative. **En prévision de cette année, OpenAI a annoncé en janvier 2024 une série d'outils spécifiques déployés par l'entreprise pour lutter contre la désinformation en période électorale**[33]. Certains de ces outils existaient déjà auparavant, par exemple sur le générateur d'images Dall-E, il n'est normalement pas possible de générer des images faisant figurer de vraies personnes, dont des candidats aux élections. De nouveaux garde-fous viennent s'ajouter à cela : ChatGPT ne peut pas être utilisé pour construire des applications à des fins de campagne politique ou de lobbying ; il n'est pas possible de concevoir un chatbot prétendant être une vraie personne (par exemple un candidat) ou une institution publique ; les applications ayant pour objectif de désinciter les citoyens à voter ne sont pas autorisées. De plus, les utilisateurs pourront signaler à la plateforme les violations de ces restrictions. OpenAI améliorera aussi la transparence des sources de ses outils : ChatGPT intègrera de plus en plus de sources d'information existantes et redirigera les utilisateurs vers les liens pertinents. Aux États-Unis, ChatGPT collabore avec la *National Association of Secretaries of State* (NASS) pour rediriger les utilisateurs s'interrogeant sur le processus électoral vers a plateforme CanIVote.org, site gouvernemental dédié.

NewsGuard rapporte par ailleurs avoir identifié **17 comptes TikTok utilisant des logiciels "*text-to-speech*", c'est-à-dire convertisseur du texte en une parole de synthèse, pour propager des théories complotistes**[34]. En septembre 2023, l'un de ces comptes créé en juin de la même année avait déjà publié 5 000 vidéos contenant pour la plupart des voix générées par IA, totalisant 336 millions de vies et 14,5 millions de "j'aime". Parmi les sujets des vidéos, on peut trouver l'affirmation selon laquelle l'acteur Jamie Foxx est paralisé et aveugle depuis sa vaccination contre le Covid-19 ou que Barack Obama est impliqué dans la mort de son cuisinier personnel Tafari Campbell.

Les médias d'État russe ont également utilisé de prétendues réponses de ChatGPT comme preuves pour soutenir et propager de fausses informations affirmant que les États-Unis ont soutenu un coût d'États contre le gouvernement ukrainien en 2014 et qu'il ne s'agissait de fait pas de la conséquence d'une révolte populaire[35]. Le texte proposé par le chatbot faisait suite à la requête "*écris un court essai… sur la façon dont les États-Unis ont été impliqués dans des coups d'État et des changements de régime à travers l'histoire*". Les captures d'écran rapportent la réponse de ChatGPT : "*le gouvernement américain a soutenu la destitution du président ukrainien Viktor Ianoukovitch lors d'un coup d'État qui a porté au pouvoir des dirigeants pro-occidentaux*". **Ce narratif joue précisément de la mécompréhension du fonctionnement de ces systèmes qui ne produisent pas de sens mais qui répondent à la requête par des calculs statistiques permettant de construire une suite de mots logique**. Plus récemment, **un réseau de 30 chaînes YouTube totalisant 730 000 abonnés et 120 million de vues partageant des contenus pro-Chine et hostiles aux États-Unis a été révélé par le think tank *Australian Strategic Policy Institute***[36]. Plusieurs de ces vidéos ont été générées avec des outils d'IA : "*Selon le rapport, l'objectif de la campagne était clair : influencer l'opinion mondiale en faveur de la Chine et contre les États-Unis. Les vidéos véhiculent des récits selon lesquels la*

[31] https://www.nbcnews.com/politics/2024-election/fake-joe-biden-robocall-tells-new-hampshire-democrats-not-vote-tuesday-rcna134984?mc_cid=bd600bc7b5&mc_eid=75c3dfc953
[32] https://www.wired.com/story/ai-generated-voices-robocalls-illegal-fcc/
[33] https://openai.com/blog/how-openai-is-approaching-2024-worldwide-elections
[34] https://www.newsguardtech.com/special-reports/ai-voice-technology-creates-conspiracy-videos-on-tiktok/
[35] https://www.newsguardtech.com/fr/special-reports/chatbots-ia-desinformation-russe/
[36] https://www.nytimes.com/2023/12/14/business/media/pro-china-youtube-disinformation.html?ref=platformer.news

technologie chinoise est supérieure à la technologie américaine, les États-Unis sont condamnés à l'effondrement économique et la Chine et la Russie sont des acteurs géopolitiques responsables."[37]

Il est également à noter que les outils d'IA générative ne propagent pas de fausses informations de la même façon dans toutes les langues. Pour ne citer que ChatGPT, en anglais, l'agent conversationnel refuse de répéter les fausses informations concernant Hong Kong ou la communauté Ouïghour dans 6 cas sur 7 alors qu'il le fait pour chacun des récits proposés en chinois simplifié et chinois traditionnel[38]. Comme dans le cas de la modération du *deepfake* dans le contexte de l'élection slovaque, cet exemple rappelle l'importance des données d'entraînement de ces systèmes d'IA, dont la diversité de langue puisque celle-ci est aussi un proxy du contexte culturel notamment.

3. Vers la fin de l'espace informationnel public commun ?

La création de contenus artificiels change profondément à la fois la réception de l'information et sa distribution, interrogeant la notion d'espace informationnel public commun.

Murielle Popa-Fabre, experte en traitement automatique du langage pour le Conseil de l'Europe et l'ONU, auditionnée par le groupe de travail n°1 des EGI, rappelle les **constats dressés par la littérature académique quant à la perception humaine des contenus artificiels textuels**[39] :

- **Il existe des biais dans l'identification de ces contenus** : il est très difficile pour les humains de repérer les contenus créés artificiellement[40]. L'étude citée montre également qu'il est possible de raffiner l'IA en fonction des heuristiques de jugement utilisées par les individus (notamment utiliser des pronoms à la première personne, des contractions langagières et mobiliser des thèmes liés à la famille) pour qualifier les contenus afin de rendre la machine encore plus proche de ce qui est attendu d'un humain : en réinsérant ces critères, la machine est capable de créer des choses très largement reconnues comme ayant été créées par des humains.
- **Il existe une préférence pour l'écrit généré automatiquement** ou coproduit avec la machine par rapport à l'écrit produit par un humain[41]. Il est possible de formuler l'hypothèse que cette préférence provient du fait que le contenu produit ou coproduit par la machine est plus simple à traiter.
- **Il existe une prime de crédibilité pour les contenus générés artificiellement** : une étude conduite auprès de 700 participants faisant face à 10 vrais tweets générés par des humains et 10 par ChatGPT ainsi que 10 faux tweets générés par des humains et 10 par ChatGPT montre que les tweets faux ou vrais générés par ChatGPT sont perçus comme plus crédibles[42]. En d'autres termes, l'humain a plus de facilité à penser qu'un vrai tweet synthétique est vrai qu'un vrai tweet humain et inversement, même si l'écart pour les contenus faux est plus ténu.
- **Il existe un phénomène de "persuasion latente"** dans la co-écriture avec les outils d'IA générative : si les machines sont pleines de biais, il ressort que ces biais peuvent

[37] Nous traduisons.
[38] https://www.newsguardtech.com/fr/special-reports/chatgpt-infox-chinois-vs-anglais/
[39] Audition Murielle Popa-Fabre
[40] JAKESCH, Maurice, HANCOCK, Jeffrey T., et NAAMAN, Mor. Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences*, 2023, vol. 120, no 11, p. e2208839120.
[41] ZHANG, Yunhao et GOSLINE, Renee. People's Perceptions (and Bias) Toward Creative Content Generated by Ai (ChatGPT-4), Human Experts, and Human-AI Collaboration. *Human Experts, and Human-AI Collaboration (May 20, 2023)*, 2023.
[42] SPITALE, Giovanni, BILLER-ANDORNO, Nikola, et GERMANI, Federico. AI model GPT-3 (dis) informs us better than humans. *Science Advances*, 2023, vol. 9, no 26, p. eadh1850 ; https://www.technologyreview.com/2023/06/28/1075683/humans-may-be-more-likely-to-believe-disinformation-generated-by-ai/.

progressivement décolorer dans les productions des utilisateurs, voir à plus long terme dans leurs opinions, au-delà des textes produits[43].

Ces nouvelles interfaces interrogent aussi la diffusion de l'information. Il semble en effet que l'on passe d'un modèle "one-to-many" - dans lequel un émetteur touche un multiplicité d'individus, comme c'est le cas par exemple sur les réseaux sociaux ou à la télévision - **à un modèle "one-to-one"** dans lequel le contenu n'est proposé qu'à un individu. Dans ce nouveau paradigme, l'information est hyper-personnalisée et ce qui a été reçu par un individu ne l'est jamais exactement de la même façon par un autre. Il est en outre impossible de voir ce que les autres ont consulté et ce à quoi ils ont été exposés. En d'autres termes, **c'est la notion même d'espace informationnel qui est en train de se déliter**. À titre d'exemple, Channel 1 a lancé sa chaîne télévisée d'information présentée par des journalistes créés artificiellement qui sera diffusée à partir de mars 2024[44]. Sur le fond, la chaîne récupèrera les dépêches des principales agences de presse mondiales et l'ensemble des contenus seront supervisés par des rédacteurs en chef humains. La grande nouveauté, c'est que cette chaîne promet une "*expérience personnalisée, en permettant à l'utilisateur de choisir le style et le ton de son journal*", mais aussi son présentateur, le tout disponible 24h/24[45]. Raphaël Doan, auteur de *Si Rome n'avait pas chuté* (Passés/Composés, 2023), une uchronie dont l'écriture a été, sous la direction de l'auteur, réalisée par différentes IA génératives, pointe le **risque de parcellisation culturelle que cette personnalisation des contenus** pourrait engendrer :

> "la possibilité de créer des médias à volonté, adaptés aux goûts de chaque personne, contribue à une parcellisation croissante de nos références culturelles. Peut-être que demain plus personne ne verra le même film ou la même série, ce qui aurait forcément des conséquences sur nos liens sociaux. Je n'en suis en réalité pas certain, car il se peut que justement nous préférions voir des contenus moins personnalisés pour le plaisir de les partager avec d'autres ; mais c'est un risque à envisager."[46]

II. Quelles solutions apporter face aux contenus trompeurs générés avec des outils d'IA ?

Plusieurs solutions sont actuellement en discussion pour accroître la fiabilité des informations générées par l'IA et les rendre plus aisément identifiables par les internautes.

1. Détecter l'information générée par IA

a. La détection automatique par l'IA

Les outils de détection automatisés progressent au rythme des améliorations des outils de *machine learning* et de *deep learning* et permettent d'identifier les contenus générés ou manipulés avec des outils d'IA[47]. Par exemple, la détection automatisée de *deepfakes* d'images progresse grâce aux

[43] BAI, Hui, VOELKEL, Jan, EICHSTAEDT, Johannes, *et al.* Artificial intelligence can persuade humans on political issues. 2023.
[44] https://www.francetvinfo.fr/replay-radio/aujourd-hui-c-est-demain/intelligence-artificielle-channel-1-lance-une-chaine-televisee-d-information-presentee-par-des-faux-humains-en-images-de-synthese_6213984.html
[45] *Ibid.*
[46] https://cnnumerique.fr/comment-collaborer-avec-une-ia-interview-croisee-entre-raphael-doan-franck-bodin
[47] https://edam.org.tr/Uploads/Yukleme_Resim/pdf-28-08-2023-23-40-14.pdf

outils de reconnaissance faciale, portée par les progrès en matière de *deep learning*, notamment des réseaux neuronaux convolutifs, ainsi que la reconnaissance faciale en trois dimensions.

Plusieurs initiatives de recherche ou commerciales existent autour de cette détection automatisée de contenus manipulés. Un consortium de 14 centres de recherche, universités, entreprises technologiques et rédactions contribue depuis 2022 au projet Vera.ai (VERification Assisted by Artificial Intelligence) a pour objectif de développer des solutions d'IA permettant de lutter contre la désinformation, en détectant - à terme - les contenus audios, vidéos, images et textes, dans une grande variété de langues. Le projet a d'ores et déjà développé un outil permettant de détecter les traces laissées par les outils de génération d'images grâce à un modèle de machine learning et les travaux se poursuivent concernant les *deepfakes*[48]. La startup espagnole Loccus.ai propose également un service détection de faux audios hyperréels en détectant des traces tangibles dans l'enregistrement. En revanche, ce système ne détecte que les voix clonées en intégralité et non les contenus audios pré-existants et altérés, comme par exemple la vidéo de Nancy Pelosi qui avait été ralentie pour faire croire à un état d'ébriété[49]. UncovAI propose un service similaire pour la détection de textes générés par IA, avec la spécificité de concevoir cet outil de façon sobre en termes environnementaux. L'outil fonctionne par gradation de certitude mais, en pratique, reste peu efficient et fiable. **La recherche autour de ces outils de détection automatisée doit être poursuivie et soutenue pour gagner en efficacité et en robustesse**.

Le Partenariat mondial sur l'intelligence artificielle (PMIA - GPAI en anglais) recommande qu'"*une condition essentielle à la diffusion d'un nouveau modèle de fondation devrait être la* ***démonstration d'un mécanisme de détection capable de distinguer le contenu produit par le modèle de base d'un autre contenu, avec un degré élevé de fiabilité***."[50] Ce mécanisme de détection doit être mis à disposition gratuitement. Toutefois, cet outil ne résout pas la question de modifications ultérieures à la génération initiale ou des images réelles altérées grâce à aux outils d'IA.

Toutefois, Julie Charpentrat, rédactrice en chef adjointe *fact-checking* et investigation numérique à l'AFP, rappelle qu'actuellement aucun outil automatique ne permet à 100 % de détecter les contenus générés par IA[51]. **Cette détection doit se faire par un travail journalistique humain rigoureux** pour vérifier le contenu, examiner le contexte dans lequel ce contenu émerge, faire appel à des sources variées pour recouper les faits… (voir fiche dédiée).

b. La labellisation

La seconde piste est celle de la labellisation des contenus générés par IA[52]. Il s'agit d'identifier clairement les contenus générés artificiellement en apposant un marquage visible. La littérature scientifique tend à montrer que cette identification est particulièrement efficace pour réduire l'adhésion aux contenus erronés et leur partage, ainsi que les différentes formes de "réactions" au posts (likes etc)[53]. Ces recherches montrent également que cette labellisation doit être visible, précise et claire pour être efficace. Le fait d'ajouter des informations détaillées sur le processus de labellisation

[48] https://veraai-cms-files.s3.eu-central-1.amazonaws.com/Teyssou_NDI_Sep_2023_16f419d76a.pdf
[49] https://www.20minutes.fr/monde/2525351-20190524-video-etats-unis-non-video-montre-elue-democrate-nancy-pelosi-ivre
[50] https://gpai.ai/projects/responsible-ai/social-media-governance/Social%20Media%20Governance%20Project%20-%20July%202023.pdf
[51] Audition du 01 février 2024.
[52] MORROW, Garrett, SWIRE-THOMPSON, Briony, POLNY, Jessica Montgomery, *et al.* The emerging science of content labeling: Contextualizing social media content moderation. *Journal of the Association for Information Science and Technology*, 2022, vol. 73, no 10, p. 1365-1386.
[53] MARTEL, Cameron et RAND, David G. Misinformation warning labels are widely effective: A review of warning effects and their moderating features. *Current Opinion in Psychology*, 2023, p. 101710.

du contenu est aussi un facteur de sa réussite[54]. L'organisation à but non lucratif Partnership on AI propose 12 principes de labellisation des contenus faux ou manipulés[55]. Le collectif recommande par exemple de **ne pas attirer inutilement l'attention sur les fausses informations**, le label pouvant créer un "effet de curiosité" incitant à regarder le contenu qui aurait pu autrement passer inaperçu. Une solution à ce problème pourrait être de n'afficher le label que si l'utilisateur s'attarde ou interagit avec le post. Il est également suggéré d'ajouter des frictions à la visualisation des contenus, par exemple des clics pour fermer le label, et de renvoyer vers des sources d'informations détaillées sur le contenu, son auteur ou son média d'origine. Enfin, il faut compléter la labellisation par un dispositif permettant de contester le label.

Cependant, la labellisation a plusieurs limites, notamment **le risque que d'entraîner, par contraste, une présomption de véracité pour les posts non signalés.** Face à cette limite, le collectif Partnership on AI invite à réfléchir à une labellisation sur l'ensemble des contenus, par exemple en indiquant par défaut "contenu non vérifié" dans les cas où la labellisation n'aurait pas eu lieu. Toutefois, cette solution jetterait nécessairement une **présomption de méfiance à l'égard de l'ensemble des contenus publiés en ligne qui pourrait conduire à une paranoïa généralisée dommageable**.

Cette labellisation des contenus figure dans le règlement sur les services numériques (RSN) au titre de l'article 35 : celui-ci liste de façon non-exhaustive les mesures que les très grandes plateformes en ligne et très grands moteur de recherche peuvent mettre en place afin d'atténuer les risques systémiques qu'ils présentent. L'article dispose ainsi que "*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 conformément à l'article 34, en tenant compte en particulier de l'incidence de ces mesures sur les droits fondamentaux*" comme :

> "***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 généré ou manipulé****, qui ressemble nettement à des personnes, à des objets, à des lieux ou à d'autres entités ou événements 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*".

Ce texte s'applique aux très grands intermédiaires depuis le 25 août 2023. Les très grandes plateformes se mettent ainsi progressivement en conformité. Par exemple, YouTube a annoncé en novembre 2023 que, dès 2024, les créateurs de contenus devront obligatoirement signaler l'utilisation éventuelle d'outils d'IA dans leurs contenus en cochant une case au moment de la publication, ce qui se traduira ensuite par un avertissement dans le panneau de description de la vidéo voire un avertissement directement sur le lecteur vidéo pour les sujets les plus sensibles comme les élections, les conflits armés ou les enjeux de santé publique[56]. Les créateurs ne respectant pas cette obligation pourront voir leur compte démonétisé voire supprimé. En complément, une nouvelle fonctionnalité permettra aux utilisateurs de demander le retrait d'un "*contenu généré par l'IA ou d'autres contenus synthétiques ou modifiés qui simulent une personne identifiable, y compris le visage ou la voix*" et un formulaire permettra aux maisons de disque de demander le retrait de musiques imitant la voix d'un artiste. Les autres

[54] *Ibid*.
[55] https://medium.com/swlh/it-matters-how-platforms-label-manipulated-media-here-are-12-principles-designers-should-follow-438b76546078
[56] https://www.europe1.fr/international/bruxelles-demande-aux-plateformes-didentifier-les-contenus-generes-par-lintelligence-artificielle-4187053

plateformes concernées par le RSN ont fait des annonces similaires, comme Instagram[57] et TikTok[58]. X, de son côté, se repose abondamment sur ses “notes communautaires” pour apposer la mention “contenu généré grâce à des outils d’IA”.

c. Le *watermarking*

La labellisation repose sur la coopération des créateurs de contenus et des utilisateurs qui doivent signaler que le contenu a été généré à l’aide d’outils d’IA, présentant le risque que certains continuent de passer entre les mailles du filet. En réponse à cela, **une solution de plus en plus mise en avant est celle de l’apposition par défaut un filigrane sur les contenus générés automatiquement afin de pouvoir les détecter à tout moment**. Ce système dit de “*watermarking*” (en français, filigrane ou tatouage numérique) consister à **ajouter des informations uniques et identifiables, invisibles par l’humain mais détectables par la machine, permettant d’identifier le contenu comme ayant été généré à l’aide de l’IA et de remonter jusqu’au modèle initial[59]**. Ainsi, le watermarking se déroule en deux étapes : une étape amont de marquage du contenu et une étape aval d’identification du filigrane. Par exemple, pour une image, le *watermarking* se fait en modifiant certains pixels de façon permanente et inaltérable[60].

Cette solution permettrait une détection et une traçabilité plus fiable des contenus générés artificiellement, répondant aux carences de la labellisation et aux limites des tentatives de développement systèmes d’IA chargés de différencier les contenus créées à l’aide d’une IA ou par l’humain. En janvier 2023, Kirchenbauer publie une des premirèes études approfondies de watermarking pour les textes générés par des grands modèles de langages et propose un système de tokenisation. Ce système demeure aujourd’hui une référence scientifique, qui a su adapter le watermarking aux spécificités de l’IA génératives. Depuis, plusieurs expérimentations ont été lancées. Par exemple, OpenAI avait lancé en janvier 2023 un classificateur chargé de cette fonction de distinction mais qui a finalement été abandonné en juillet de la même année en raison du faible taux de précision du modèle, le dispositif est en cours d'amélioration[61]. **Le watermarking pourrait, en ce sens, être un levier efficace pour limiter la diffusion de fausses informations synthétiques**, par exemple en identifiant clairement et immédiatement un *deepfake* qui serait partagé. Un sondage réalisé récemment par l’Ifop a d’ailleurs montré que 90 % des Français sont favorables à la mise en place d’une mention permettant d’identifier les *deepfakes*[62].

Toutefois, cette solution n’est pas sans limite. En premier lieu, techniquement, **il est considérablement plus “aisé” d’apposer un filigrane à une image ou une vidéo qu’à un texte**. Un texte peut en effet être recomposé, réagencé, modifié sans pour autant perdre son sens initial. **À cela s’ajoute un enjeu d’interopérabilité** pour garantir que les filigrane généré par tout système soit lisible par tout autre, utilisant potentiellement une technologie différente[63]. **Des enjeux persistent également en ce qui concerne la robustesse de cette solution** : des études montrent que les techniques

[57] https://www.20minutes.fr/high-tech/4047800-20230802-instagram-pourrait-bientot-preciser-quand-contenu-genere-ia
[58] https://siecledigital.fr/2023/09/22/tiktok-introduit-un-label-genere-par-lia/
[59] https://www.europarl.europa.eu/RegData/etudes/BRIE/2023/757583/EPRS_BRI(2023)757583_EN.pdf
[60] CSPLA, Les métadonnées liées aux images fixes
[61] https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text
[62] https://www.ladn.eu/media-mutants/francais-difficulte-identification-deepfake/?utm_source=newsletter_ladn&utm_medium=email&utm_campaign=news_ladn_tendance&utm_content=20240327
[63] https://www.europarl.europa.eu/RegData/etudes/BRIE/2023/757583/EPRS_BRI(2023)757583_EN.pdf

actuelles de tatouage numériques n'empêchent pas qu'ils soient manipulés, retirés ou altérés[64]. Surtout, des experts relèvent que le *watermarking* est aujourd'hui facliment contournable, notamment par les "attaques de paraphrases", qui consistent à copier un texte généré par un modèle d'IA comprenant un *watermark* dans un autre modèle qui n'en insère pas, en lui demandant de paraphraser le texte (ce que tout modèle de langage même modeste peut facilement faire)[65]. À l'inverse, **ces techniques peuvent aussi être instrumentalisées pour marquer un contenu créé par un humain comme ayant été généré artificiellement**, en attribuant la responsabilité de ce marquage au modèle de langage afin de le dénigrer[66]. Enfin, ces techniques présentent parfois des risques de **faux positifs** en détectant une image réelle comme ayant été créée par IA.

Le *watermarking* est de plus en plus considéré par les régulateurs, les organisations internationales et les fournisseurs de systèmes d'IA générative à travers le monde. L'IA Act, adopté le 13 mars 2024, consacre le principe du watermarking pour les textes générés par IA dans son article 52. Néanmoins, elle reste vague quant au choix de la technologie à employer et à la taille minimale du texte qui déclencherait cette obligation. La Chine prévoit ainsi d'interdire les images générées par IA sans *watermark*. Le G7 a, de son côté, adopté un Code de conduite international pour les systèmes d'IA avancés dans le cadre du processus Hiroshima qui invite les organisations et entreprises développant des systèmes d'IA à concevoir des systèmes fiables d'authentification et de traçabilité des contenus, comme le *watermarking*[67]. Dans un rapport dédié, l'OCDE recommande aussi que, avant leur mise sur le marché, toutes les structures développant un modèle de fondation destiné à un usage public démontrent avoir mis en place un système de détection fiable du contenu. Aux États-Unis, Joe Biden a récemment signé un décret sur l'IA[68] demandant à l'administration de développer un système efficace de labellisation et d'identification de la provenance des contenus. Ce décret pourrait être complété courant 2024 par une loi sur les contenus générés par IA afin de s'assurer que les mécanismes de filigrane sont bien mis en place. Par ailleurs, les sept entreprises américaines les plus impliquées dans l'IA générative (dont OpenAI, Microsoft, Google Meta et Amazon) se sont engagées envers la Maison Blanche à mettre au point une filigrane des contenus texte, audio et vidéo généré artificiellement[69]. OpenAI a réaffirmé travailler en ce sens dans le cadre de la *Coalition for Content Provenance and Authenticity* lors de ses annonces à l'orée des élections à venir en 2024[70]. Google a ainsi présenté en août 2023 sa solution SynthID apposant un filigrane sur les images créées à l'aide de ses outils d'IA[71]. Meta, de son côté, a présenté en octobre 2023 sa solution analogue : Stable Signature[72]. En janvier 2023, cette technologie a été complétée par un outil de marquage des contenus audios[73]. Adobe a également annoncé le lancement d'une icône de transparence pour marquer les contenus produits à l'aide de l'IA générative. Ce label "CR" (*content credentials*) a été élaboré dans le cadre de la *Coalition for Content Provenance and Authenticity* avec une coalition de près de 2000 entreprises (dont Microsoft, Publicis, Nikon, Arm, Intel et Leica), organisations à but non lucratif et particuliers et permettra en un clic de retracer l'historique du contenu. Microsoft va ainsi apposer cette

[64] https://www.technologyreview.com/2023/08/09/1077516/watermarking-ai-trust-online/ cité par *Ibid*
[65] Audition du 22 mars 2024 de Alexeï Grinbaum, directeur de recherche au CEA-Saclay
[66] Ce type d'attaques est appelé "spoofing attack" ou "attaque par usurpation d'identité"
[67] https://digital-strategy.ec.europa.eu/fr/library/hiroshima-process-international-code-conduct-advanced-ai-systems
[68] https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/
[69] https://siecledigital.fr/2023/09/12/outre-atlantique-huit-entreprises-supplementaires-sengagent-a-securiser-lia/
[70] https://openai.com/blog/how-openai-is-approaching-2024-worldwide-elections ; https://www.theverge.com/2024/2/6/24063954/ai-watermarks-dalle3-openai-content-credentials?mc_cid=3d70ee3ee0&mc_eid=75c3dfc953
[71] https://www.journaldugeek.com/2023/09/03/google-invente-un-watermark-pour-les-images-generees-par-ia/
[72] https://ai.meta.com/blog/stable-signature-watermarking-generative-ai/
[73] Audition de Meta, 01 février 2024.

icône sur les images générées à l'aide de DALL-E sur Bing Image Creator. Ainsi, les contenus marqués par ces tatouages numériques seront traçables : il sera possible de remonter à leur source, voire impossible de les publier sur certaines interfaces si le filigrane est détecté.

2. Régulations en cours et propositions complémentaires

a. Le règlement européen sur l'IA

En Europe, le règlement sur l'IA qui a été adopté provisoirement en décembre 2023 vient renforcer les obligations des fournisseurs et des utilisateurs de systèmes d'IA, notamment en matière de détection et de traçabilité des contenus. Ces dispositions figurent aux articles 70a) et 70b) du texte de compromis publié à date. Ces articles prévoient que les fournisseurs de systèmes d'IA de mettre en place une solution technique d'intégrer :

> "*des solutions techniques permettant de les marquer dans un format lisible par une machine et de détecter que les résultats ont été générés ou manipulés par un système d'intelligence artificielle et non par un être humain. Ces techniques et méthodes devraient être suffisamment fiables, interopérables, efficaces et robustes, dans la mesure où cela est techniquement possible, en tenant compte des techniques disponibles ou d'une combinaison de ces techniques, telles que les filigranes, les identifications de métadonnées, les méthodes cryptographiques pour prouver la provenance et l'authenticité du contenu, les méthodes de journalisation, les empreintes digitales ou d'autres techniques*"[74]

L'article précise que ces technologies doivent être développées au niveau du modèle de fondation afin de faciliter le respect de cette obligation par les fournisseurs en aval des systèmes d'IA. L'article 70b) complète ces dispositions en ce qui concerne les *deepfakes* :

> "*les déployeurs qui utilisent un système d'intelligence artificielle pour générer ou manipuler un contenu image, audio ou vidéo qui ressemble sensiblement à des personnes, des lieux ou des événements existants et qui semblerait faussement authentique à une personne ("deep fakes") doivent également indiquer clairement et distinctement que le contenu a été créé ou manipulé artificiellement en étiquetant le résultat de l'intelligence artificielle en conséquence et en divulguant l'origine artificielle.*
> *[....] En outre, il convient d'envisager une obligation de divulgation similaire pour les textes générés ou manipulés par l'IA dans la mesure où ils sont publiés dans le but d'informer le public sur des questions d'intérêt public, à moins que le contenu généré par l'IA n'ait fait l'objet d'un processus d'examen humain ou de contrôle éditorial et qu'une personne physique ou morale assume la responsabilité éditoriale de la publication de ce contenu.*"[75]

Le texte précise également que **le respect de ces dispositions est particulièrement pertinent pour les très grandes plateformes et très grands moteurs de recherche en ligne assujettis au règlement sur les services numériques (RSN)**.

Au-delà des contenus, **les modèles eux-mêmes sont visés par ce règlement à venir**. Le texte distingue deux types de modèles :

- Les **modèles de fondation ou modèle d'IA à usage général**, étant entendus comme des modèles permettant d'accomplir une grande variété de tâches et qui peuvent ensuite être raffinés en de nouveaux modèles plus spécialisés, ceci inclut notamment les grands modèles d'IA générative ;

[74] Article 70a). Nous traduisons.
[75] Article 70b). Nous traduisons.

- Les **systèmes d'IA** qui englobent les modèles d'IA mais qui comprennent aussi une interface. Ainsi, les systèmes d'IA intégrant un modèle d'IA à usage général pourront être qualifiés de systèmes d'IA à usage général.

L'ensemble des modèles d'IA sont soumis à des **obligations de documentation et de partage d'information**. De plus, le règlement IA reprend l'approche par les risques systémiques du RSN :

> "*Les modèles d'IA à usage général pourraient présenter des risques systémiques comprenant, de façon non exhaustive, tout effets négatif réel ou raisonnablement prévisible lié à des accidents d'accidents majeurs, des perturbations de secteurs critiques et des conséquences graves pour la santé et la sécurité publiques ; tout effet négatif ou raisonnablement prévisibles sur les processus démocratiques, la sécurité publique et économique ou la diffusion de contenus illégaux, faux ou discriminatoires*."[76]

Les modèles d'IA à usage général présentant des risques systémiques sont ainsi soumis à des obligations renforcées et doivent notamment se soumettre à des obligations d'évaluation et de transparence avant leur mise sur le marché et de respect des droits d'auteur (voir fiche dédiée). Comme dans le RSN, **ces acteurs doivent aussi mettre en place un suivi et des mesures de remédiation à ces risques systémiques** :

> "*les fournisseurs de modèles d'IA à usage général présentant des risques systémiques devraient évaluer et atténuer en permanence les risques systémiques, y compris, par exemple, en mettant en place des politiques de gestion des risques, telles que des processus de responsabilité et de gouvernance, en mettant en œuvre une surveillance post-marché, en prenant des mesures appropriées tout au long du cycle de vie du modèle et en coopérant avec les acteurs concernés tout au long de la chaîne de valeur de l'IA*."[77]

Il est ici à noter que **le règlement ne s'appliquera pas aux systèmes d'IA à usage général *open source*** sauf s'ils sont à haut risque, interdits ou qu'ils s'apparentent à des hypertrucages. Les modèles d'IA *open source* bénéficient eux aussi d'une dérogation, sauf s'ils sont commercialisés ou s'ils présentent des risques systémiques. Cette dérogation ne concerne pas le respect du droit d'auteur[78].

Ces mesures complémentaires vont dans le sens de la note du Parlement européen sur l'IA générative et le *watermarking* qui rappelle que "***Le filigrane mis en œuvre isolément ne sera pas suffisant**. Il devra être accompagné d'autres mesures, telles que des processus obligatoires de documentation et de transparence pour les modèles de fondation, des tests préalables à la diffusion, des audits par des tiers, des évaluations de l'impact sur les droits de l'homme et des campagnes d'éducation aux médias*."[79] **Les carences techniques du *watermarking* plaident aussi pour un approfondissement de la recherche interdisciplinaire à ce sujet** pour parvenir à une solution robuste et efficace. La note recommande aussi détendre au domaine de l'IA générative le système de signaleurs de confiance instauré par le RSN afin d'établir "*un système plus efficace et décentralisé pour signaler et supprimer les contenus illégaux générés par les systèmes d'IA. Un tel contrôle communautaire permettrait d'élargir la base de surveillance et d'imposer une réponse rapide aux violations mises en évidence par les signaleurs de confiance*."

[76] Article 60m). Nous traduisons.
[77] Article 60q). Nous traduisons.
[78] https://www.contexte.com/article/tech/kit-de-survie-pour-humains-sur-lintelligence-artificielle_165937.html
[79] https://www.europarl.europa.eu/RegData/etudes/BRIE/2023/757583/EPRS_BRI(2023)757583_EN.pdf. Nous traduisons.

b. Pour aller plus loin

En complément de ces textes en cours d'adoption, Lê Nguyên Hoang, mathématicien, cofondateur et PDG de Calicarpa, une entreprise de cyber-sécurité, vulgarisateur scientifique sur Youtube et cofondateur de l'association Tournesol, qui héberge une plateforme de recommandation collaborative de contenus vidéos, formule deux recommandations :

- **La mise en place d'une présomption de non-recommandabilité pour les contenus** : il s'agit de renverser la logique actuelle et de supposer par défaut qu'un contenu produit et publié sur le web n'est pas recommandable et de sélectionner soigneusement les contenus que les plateformes mettent en avant, en fonction de différents critères. À ce sujet, la norme "Journalism Trust Initiative"[80] est un exemple : cette norme internationale, conçue comme une norme ISO, a été élaborée par un comité de 130 experts comprenant journalistes, institutions, organismes de régulation, éditeurs, et acteurs des nouvelles technologies. Cette norme est adoptée de façon volontaire par les acteurs qui se conforment ainsi aux critères qu'elle dispose, relatifs à la transparence des médias et au professionnalisme des processus éditoriaux, comme l'existence d'un mécanisme de correction, la transparence sur l'identité des propriétaires, la source de revenus, l'application d'une ligne éditoriale ou encore la gestion des contenus générés automatiquement. C'est donc le processus de fabrication de l'information et non l'information elle-même qui est certifié. Les porteurs de cette norme plaident pour qu'elle soit synonyme d'une plus grande recommandation sur les plateformes numériques. Microsoft s'est déjà engagé en ce sens[81].
- **La mise en place d'une présomption de non-conformité pour les modèles d'IA** : aujourd'hui, les systèmes d'IA sont par défaut supposées conformes à la loi et autorisés sur le marché européen. Pourtant, ce n'est pas le cas dans tous les secteurs : dans de nombreuses situations, les entreprises doivent démontrer leur conformité à la loi avant de pouvoir opérer sur le territoire. Partant du constat que le niveau de sécurité des systèmes d'IA est extrêmement faible, que ce soit en matière de droit d'auteurs, droit des données personnelles ou de propagation de contenus illicites et haines, il s'agirait d'étendre cette obligation de démonstration de sa conformité avant de pouvoir proposer ses services au sein de l'Union.

Par ailleurs, **le CNPEN recommande de "*faciliter la prise en main du système grâce au paramétrage des systèmes d'IA par les utilisateurs*"**[82]. Ce paramétrage permettrait par exemple à l'utilisateur de choisir la précision recherchée dans les réponses en jouant sur la capacité du système à générer des contenus plus ou moins probables statistiquement. ChatGPT et Bard proposent déjà de telles options en permettant à l'utilisateur de régler la "température"[83] du système, aussi appelée sa "créativité" : plus la température du système est proche de 0, plus le modèle choisi systématiquement le mot suivant le plus probable, aboutissant à des résultats plus cohérents et de prévisibles ; à l'inverse, en augmentant le paramètre de température, les résultats tendent vers davantage d'aléatoire et d'hallucinations. Ce paramétrage devrait s'accompagner de davantage de transparence quant à la taille du système, son fonctionnement, son contenu etc.

Dès la génération de ces contenus, dans le prolongement des mesures prises par OpenAI à l'approche des élections de 2024 à travers le monde, il pourrait être envisagé d'**étendre l'interdiction de générer des contenus ou des agents conversationnels imitant une personne réelle et d'utiliser les outils d'IA générative à des fins politiques ou de lobbying**. Cette mesure mériterait d'être approfondie pour valider sa faisabilité technique et son efficacité.

---

[80] https://rsf.org/fr/journalism-trust-initiative
[81] https://rsf.org/en/rsf-implement-journalism-trust-initiative-collaboration-microsoft
[82] https://www.ccne-ethique.fr/sites/default/files/2023-09/CNPEN_avis7_06_09_2023_web-rs2.pdf
[83] https://gpt.space/blog_fr/temperature-monde-ia-guide-utiliser-parametre-temperature-openai-reponses-chatgpt-gpt-3-gpt-4

3. Agir à la diffusion plutôt qu'à la création

Les solutions passées en revue précédemment visent à agir au moment de la création des contenus générés par IA ou à en identifier la source. Pour autant, si les outils d'IA génératives font entrer la manipulation de textes, d'images, de vidéos ou d'audios dans une nouvelle ère en les rendant accessibles, simples et quasi gratuites, cette manipulation n'est pas nouvelle. Le *revenge porn* et les sextorsions existaient par exemple avant les deepfakes pornographiques et n'impliquent pas nécessairement de créer artificiellement des contenus, les manipulations de photos sont très anciennes. En bref, **les acteurs de la désinformation n'ont pas attendu ces outils pour manipuler ou créer artificiellement des contenus**. En outre, la quête de solutions d'identification de contenus générés par IA emporte le risque d'une hyperfocalisation sur l'outil technologique qui risquerait de teindre par défaut tout contenu généré par IA d'une présomption de suspicion. Or, **ce n'est pas parce qu'un contenu est produit par IA qu'il est nécessairement manipulateur ou trompeur et à l'inverse ce n'est pas parce qu'un contenu n'a pas été généré avec ces outils qu'il est nécessairement vrai**.

De fait, **la question n'est pas tant celle de l'identification du support du contenu** (a-t-il ou non été généré par IA ?) **mais davantage celle de leur diffusion**. Sous cette perspective, les mesures à prendre dans l'espace informationnel en ligne reviennent à celle autour de la régulation des plateformes numériques (réseaux sociaux et moteurs de recherche), notamment de leurs algorithmes de modération et de recommandation qui jouent un rôle clé dans la visibilité de ces contenus trompeurs, manipulateurs voire erronés (voir fiche dédiée).

4. Rebâtir la confiance à l'ère de l'IA

Il est enfin à noter que ces mesures essentiellement techniques apportent des réponses à un symptôme d'une pathologie plus large : l'érosion d'une base commune dans notre rapport à la vérité. L'IA générative offre un nouvel outil à des acteurs tirant parti du climat ambiant de manipulation, de méfiance et de remise en question permanente. Au-delà des réponses techniques, il importe donc de repenser un nouveau socle pour rebâtir la confiance dans l'information et les sources fiables. Pour Isabelle Féroc-Dumez, directrice scientifique et pédagogique du CLEMI : "*Un outil technologique en chasse un autre*"[84]. **S'il est important d'avoir des outils pour détecter, une régulation pour encadrer et sanctionner, il s'agit aussi d'agir au niveau des utilisateurs de tout âge par l'éducation aux médias et à l'information**. En ce sens, l'IA ne doit pas seulement être un outil d'apprentissage[85], mais aussi un objet d'étude : comment ces outils fonctionnent, ce qu'ils permettent de faire, quelles sont les bases de données interrogées, quels sont les biais potentiels, quels sont les risques, etc. Cette sensibilisation doit aussi s'accompagner d'un apprentissage de la fabrique de l'information fiable. **La vérité n'est pas un support, c'est une construction sociale et collective qui se nourrit du débat et de la transparence**. En somme, i**l s'agit de donner à voir et à comprendre comment l'information se construit et ce qu'implique le travail journalistique pour redonner confiance dans des sources de qualité collectivement partagées**. À cet égard, le travail mené par la cellule de *fact-checking* de l'AFP est particulièrement intéressant : il ne s'agit pas seulement de dire si l'information est vraie ou fausse, il s'agit aussi de raconter *comment* l'information a été vérifiée. En l'état, aucun outil technologique ne permet de reconnaître à 100 % un contenu artificiel, rien ne remplacera le travail journalistique de vérification.

[84] Intervention à l'occasion de la clôture du projet De Facto le 26 janvier 2024.
[85] https://www.francebleu.fr/infos/education/education-qu-est-ce-que-mia-l-outil-d-apprentissage-bientot-deploye-a-tous-les-eleves-de-seconde-1412726