Clara asbl
Audiovisuel - Education Permanente

Etude

Intelligence artificielle : un nouveau défi ou une condamnation pour l'humanité ?

Jean-François Furnémont, Président du CLARA

1. Introduction

> *Jennifer DeStefano, des États-Unis d'Amérique, a reçu un appel téléphonique d'hommes qui semblaient être des ravisseurs de ses filles exigeant une rançon, tout en entendant la voix de sa fille en pleurs. Elle a entendu la « voix de sa fille » lui dire : « Maman, j'ai foiré », entre deux sanglots. Après des menaces et des demandes d'une voix masculine, elle a entendu sa fille dire : « Aide-moi, maman. Aidez-moi. Aidez-moi ». Il s'est avéré que sa fille était dans sa chambre tout le temps, totalement inconsciente de l'appel terrifiant que sa mère a reçu, alors que les escrocs ont utilisé des techniques d'IA pour fabriquer/reproduire la voix de sa fille.*[1]

L'adage « Me croyez-vous ou croyez-vous vos yeux / vos oreilles ? », que les gens semblent expérimenter de plus en plus, bien que dangereux en soi et conduisant à de nombreux défis dans le monde, semble stupide face aux perturbations potentielles provoquées par l'utilisation (malveillante) de l'intelligence artificielle, qui bouscule les croyances fondamentales sur ce que nous voyons et entendons.

Ce n'est là qu'un des défis posés par l'essor de l'intelligence artificielle (IA), une technologie qui permet aux ordinateurs et aux machines de simuler l'intelligence humaine et les capacités de résolution de problèmes. En tant que domaine de l'informatique, l'IA englobe l'apprentissage automatique (une branche de l'IA axée sur l'utilisation de données et d'algorithmes pour permettre à l'IA d'imiter la façon dont les humains apprennent, améliorant progressivement sa précision) ainsi que l'apprentissage profond (un sous-ensemble de l'apprentissage automatique qui utilise des réseaux neuronaux multicouches pour simuler le pouvoir de décision complexe du cerveau humain)[2].

Le monde n'a pas seulement rencontré l'IA avec le lancement de ChatGPT, un produit en évolution développé par la société de recherche Open AI, soutenue par la majorité des géants de la technologie (par exemple, Elon Musk, Sam Altman, etc.). Chat GPT est un chatbot d'IA générative qui utilise le traitement du langage naturel pour créer un dialogue conversationnel de type humain.

L'un des premiers aperçus massifs de l'utilisation des « logiques » algorithmiques a été fourni au monde avec les réseaux sociaux, où de nombreuses vulnérabilités et défis à tous les aspects de la vie sociale et surtout politique ont déjà été exposés. L'IA, et en particulier l'IA générative, tout en s'appuyant sur la collecte de données, nous laisse facilement imaginer des visions dystopiques de robots se rebellant contre leurs créateurs humains et les détruisant.

---

[1] https://www.itv.com/news/2023-04-12/mother-warns-of-ai-voice-cloning-scam-after-fearing-her-daughter-was-kidnapped

[2] https://www.ibm.com/topics/artificial-intelligence

À ses débuts, ChatGPT fait des erreurs, interprète mal les faits, mais, surtout, il fabrique des réponses qu'il ne connaît pas. Au début du développement de l'IA, lors des travaux de développement initial de la version de l'IA de Google, les découvertes effrayantes ont été notées en relation avec les systèmes d'IA, y compris lorsqu'on leur racontait une blague. Le scientifique Geoffrey Hinton, l'un des parrains de l'IA, alors qu'il travaillait chez Google sur un programme public jamais publié appelé le Palm (version Google de Chat GPT), a nourri la machine d'une blague et s'est rendu compte que la machine était capable de l'expliquer. Bien qu'il n'ait pas été en mesure d'expliquer toutes les blagues dont il a été nourri, il y en a certaines qu'il a parfaitement comprises. L'une d'entre elles raconte l'histoire d'une personne qui dit qu'elle va prendre l'avion pour rendre visite à la famille le 6 avril et dont la mère a dit : « Oh, super, la lecture de poésie de ton beau-père a lieu ce soir-là ». Et donc la personne décide de prendre l'avion le 7 avril… Être capable de comprendre une blague jusqu'à ce point était considéré intimement et exclusivement comme humain (et pas tous les humains), avec la capacité uniquement humaine de comprendre tant de nuances du langage, de la nature humaine, de la pensée, de l'interaction, etc[3]. Comme l'affirme le professeur Yuval Noah Harari :

> *« L'IA a acquis des capacités remarquables de manipulation et de génération de langage, que ce soit avec des mots, des sons ou des images... et a ainsi piraté le système d'exploitation de notre civilisation... Que se passerait-il une fois qu'une intelligence non humaine deviendrait meilleure que l'humain moyen pour raconter des histoires, composer des mélodies, dessiner des images et régider des lois et des écritures ? … Pensez à la prochaine course présidentielle américaine en 2024, et essayez d'imaginer l'impact des outils d'IA qui peuvent être utilisés pour produire en masse des contenus politiques, des fausses nouvelles et des écrits pour de nouvelles sectes... Dans une bataille politique pour les esprits et les cœurs, l'intimité est l'arme la plus efficace, et l'IA vient d'acquérir la capacité de produire en masse des relations intimes avec des millions de personnes »*[4].

Que les points de vue apocalyptiques sur les effets de l'IA soient déraisonnables ou mélodramatiques reste à vérifier ou à réfuter, mais les risques liés à l'IA – du moins pour l'instant – doivent être doivent être liés, entre autres, à des déformations accrues ou améliorées des processus démocratiques ou à des dommages involontaires causés aux êtres humains. L'humanité doit-elle être considérée comme condamnée par l'IA ? Ou la prolifération et la croissance de l'IA sont-elles encore un autre défi humain à relever et à réguler ? Ce sont des questions d'actualité qui méritent d'être explorées et surtout portées au premier plan du débat public.

---

[3] Podcast du Guardian BlackBox, épisode 6 « Shut it down »: https://www.theguardian.com/technology/audio/2024/mar/04/episode-one-the-connectionists-ai-podcast

[4] https://www.economist.com/by-invitation/2023/04/28/yuval-noah-harari-argues-that-ai-has-hacked-the-operating-system-of-human-civilisation

2. Utilisation et revenus

Quelles sont les principales utilisations (jusqu'à présent) de l'IA et quels sont les revenus qu'elle génère ?

Un grand nombre de processus différents qui reposent sur des méthodes basées sur l'IA sont aujourd'hui utilisés individuellement ou en combinaison par les entreprises. Pour se limiter au secteur qui nous occupe au Clara, à savoir celui des médias, il s'agit notamment de :

- « Big Data Analytics » : utilisé pour analyser de grandes quantités de données afin d'identifier des modèles, des corrélations inconnues, des tendances de marché ou des préférences des utilisateurs. Le processus comprend également des applications basées sur des méthodes statistiques et des applications basées sur l'apprentissage automatique qui synthétisent des événements passés (analyse descriptive), évaluent (analyse diagnostique) et prédisent des événements probables (analyse prédictive) et proposent des recommandations d'action (analyse prescriptive)[5].

- Génération de langage naturel (« Natural Language Generation » – NLG) : sert à la génération automatisée de texte. Le traitement du langage naturel (« Natural Language Processing » – NLP) est utilisé, entre autres, pour le traitement, la synthèse et la traduction automatiques de textes. La compréhension du langage naturel (« Natural Language Understanding » – NLU) présente la forme la plus sophistiquée d'applications d'IA basées sur du texte. A l'aide des technologies d'apprentissage profond, NLU peut comprendre le contenu et le contexte du texte.

- Réseaux antagonistes génératifs (« Generative Adversarial Networks » – GAN) : utilisés entre autres pour la production d'images photoréalistes, la modélisation de modèles de mouvement dans des vidéos ou la génération d'images 3D à partir d'images 2D. Les « deepfakes » sont également fréquemment basés sur ce processus d'IA haute performance. En principe, le GAN se compose de deux réseaux neuronaux dans une boucle de rétroaction qui s'entraînent ainsi mutuellement.

L'un des domaines qui fait l'objet de discussions de plus en plus importantes est celui de la modération (automatisée) des contenus. Étant donné que l'augmentation massive des contenus en ligne impliquant des discours de haine, d'autres communications insultantes et des contenus préjudiciables aux mineurs représente

[5] Cambridge Consultants (2019), Goldhammer/Dieterich/Prien (2019).

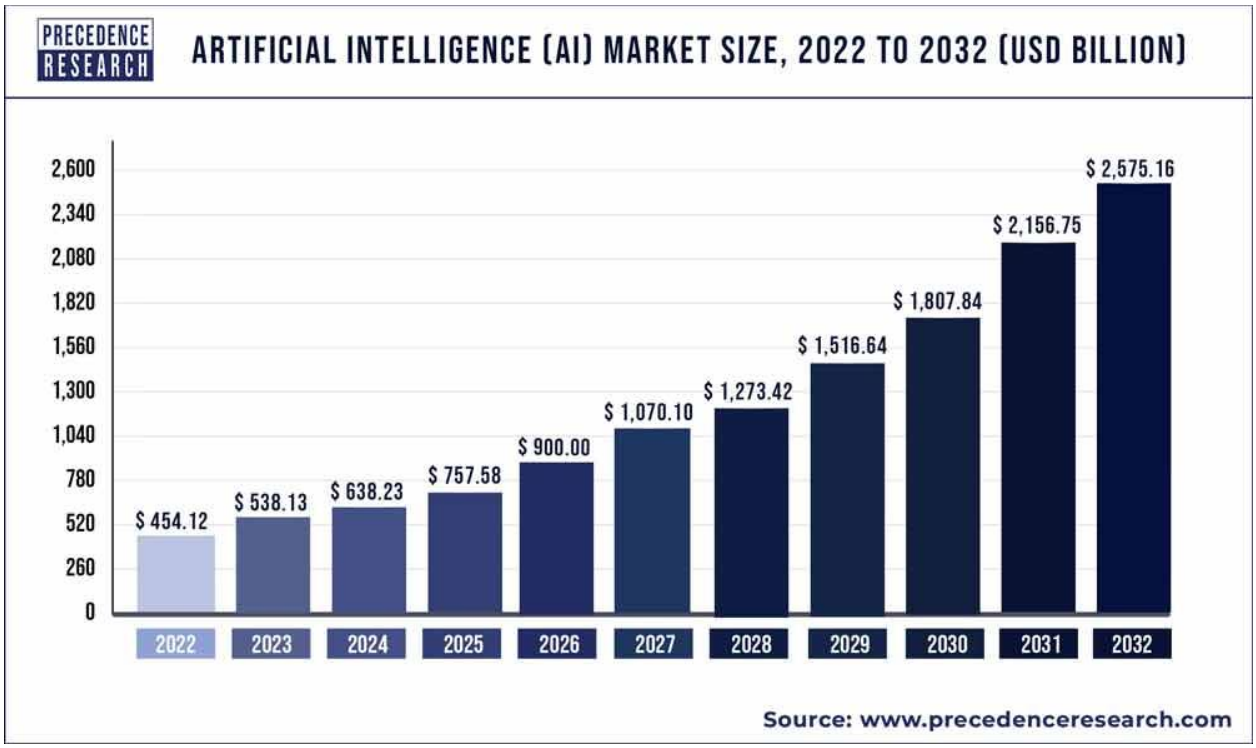
aujourd'hui un défi considérable pour le vivre-ensemble, la directive sur les services de médias audiovisuels (voir à cet égard les précédentes études du Clara qui en détaillent plusieurs aspects spécifiques) impose désormais aux plateformes de partage de vidéos (réseaux sociaux, YouTube et similaires) de mettre à disposition des usagers des instruments appropriés pour lutter contre ces contenus et les supprimer. Parallèlement, les plateformes et les fournisseurs de contenu médiatique s'engagent de plus en plus à modérer les contenus, les équipes étant chargées de filtrer les contenus problématiques parmi les masses de contenus générés par les utilisateurs qui sont téléchargés sur les plateformes, puis de décider de les supprimer ou non.

Pour les modérateurs de contenu, il est désormais presque impossible d'identifier et de supprimer tous les contenus nuisibles à temps, de manière adéquate et correcte. À l'heure actuelle, la modération de contenu implique généralement un logiciel conçu pour détecter les mots-clés, les images ou les vidéos en phase de pré-modération qui sont portés à l'attention du modérateur (humain) en tant que contenu problématique (« hash matching », « keyword filtering »). Cependant, l'efficacité de ces solutions est limitée, car l'importance et le contexte du contenu publié (sarcasme, ironie, valeurs culturelles différentes, etc.) ne peuvent pas être attribués correctement dans tous les cas. Pourtant, il existe un fort potentiel d'applications d'apprentissage automatique et d'apprentissage profond conçues pour améliorer considérablement la précision et l'efficacité de la modération des plateformes. Les applications sont conçues pour mieux percevoir les contenus problématiques et leur contexte à l'avenir. Les mécanismes d'apprentissage profond fournissent des données plus variées et plus réalistes pour entraîner des systèmes de modération basés sur l'IA. Enfin, ils pourraient soutenir les modérateurs humains de manière à ce que ces derniers puissent augmenter leur productivité en étant moins exposés aux contenus les plus nuisibles grâce à l'application de processus de pré-modération.

Comme exemple pratique, on peut mentionner les problèmes liés à la modération des plateformes, qui ont attiré des l'attention des médias. D'une part, il y a la question de la « sur-modération », c'est-à-dire de la suppression de contenus non problématiques, ce qui arrive notamment sur Facebook. Cela a suscité de nombreuses critiques dans le grand public dans des cas où le contexte culturel et/ou historique a été ignoré, par exemple dans le cas d'une photo montrant une statue du dieu grec Poséidon nu, ou dans le cas de la photo, mondialement connue, de la jeune fille nue fuyant une attaque au napalm pendant la guerre du Vietnam[6]. Cependant, la « sous-modération » persiste également comme un problème : indépendamment des procédures automatisées, les plateformes ont admis par exemple avoir eu des problèmes majeurs pour supprimer le grand nombre de vidéos téléchargées qui couvraient l'attaque terroriste contre une Christchurch en Nouvelle-Zélande. Enfin, le débat public porte également sur la charge psychologique considérable et les tensions auxquelles les équipes de modération de contenu sont fréquemment exposées[7].

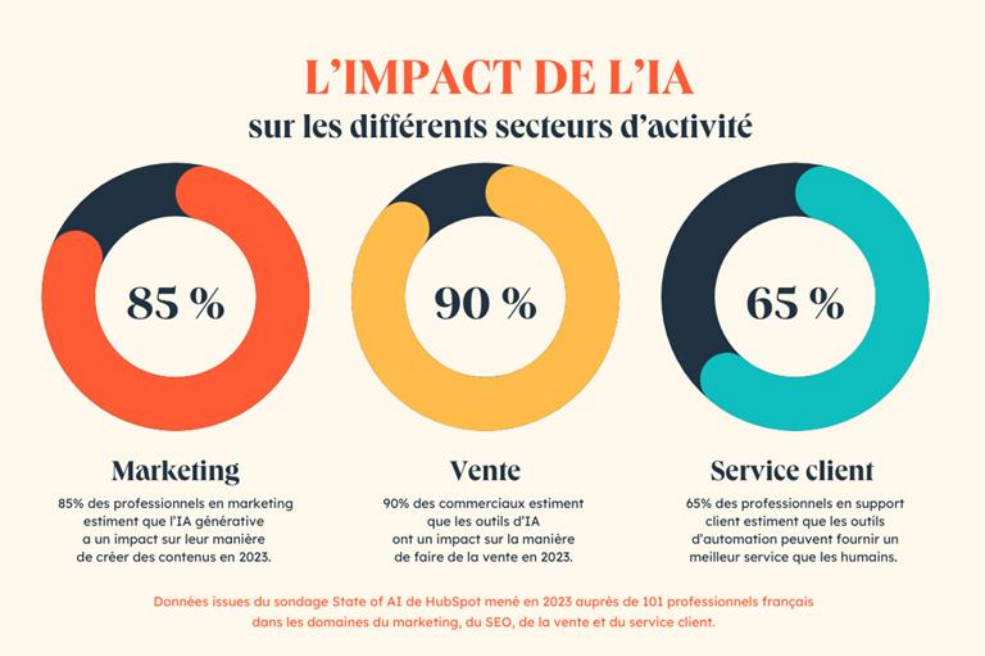
[6] https://www.cnet.com/news/facebook-content-moderation-is-an-ugly-business-heres-who-does-it/
[7] Ibid.

En ce qui concerne les valeurs du marché, le marché mondial de l'intelligence artificielle (services, logiciels et matériel) devrait connaître un taux de croissance annuel moyen de 37,3 % entre 2024 et 2030. L'Amérique du Nord a généré plus de 36,84 % de la part de marché en 2022. Le marché de l'Asie-Pacifique devrait se développer avec le taux de croissance annuel moyen le plus élevé de 20,3 % de 2023 à 2032. Sur la base de la technologie, le segment de l'apprentissage profond a capturé une part de marché de 36,36 % en 2022. Sur la base des solutions, le segment des services bancaires, financiers et d'assurances (« Banking, Financial Services, and Insurance » – BFSI) a représenté une part de marché de plus de 39,64 % en 2022. Par utilisateur final, le segment des services représentait 16,82 % de la part de marché en 2022.[8]



Selon des enquêtes, 4 à 84 % des professionnels interrogés considèrent l'IA comme un gain de temps et 62 % des clients sont impatients de voir les améliorations de l'IA dans leur expérience d'utilisateur. Aux États-Unis, en 2020, des chercheurs ont découvert que 45 % des agences fédérales avaient expérimenté l'IA et les outils d'apprentissage automatique associés, dans des domaines tels que l'application de la loi, la surveillance et la communication avec le public[9].

[8] https://www.precedenceresearch.com/artificial-intelligence-market
[9] https://blog.hubspot.fr/marketing/sondage-ia


L'IMPACT DE L'IA
sur les différents secteurs d'activité
85 %
90 %
65 %
Marketing
85% des professionnels en marketing estiment que l'IA générative a un impact sur leur manière de créer des contenus en 2023.
Vente
90% des commerciaux estiment que les outils d'IA ont un impact sur la manière de faire de la vente en 2023.
Service client
65% des professionnels en support client estiment que les outils d'automation peuvent fournir un meilleur service que les humains.
Données issues du sondage State of AI de HubSpot mené en 2023 auprès de 101 professionnels français dans les domaines du marketing, du SEO, de la vente et du service client.

3. Nuisances

3.1. Contexte

Pour illustrer l'impact de l'IA jusqu'à présent, voici quelques exemples provenant d'un éventail varié de secteurs, en reconnaissant que certains sont drôles, voire humoristiques, tandis que d'autres peuvent être vécus comme assez terrifiants. La sélection nous donne peut-être une idée de ce qui est à venir, de ce qui est en jeu et dans quelle mesure les limites des enjeux et des défis de l'IA sont repoussés vers des frontières inimaginables.

<table>
<tr><th colspan="5">2024</th></tr>
<tr><td colspan="2">La première projection publique de « The Last Screenwriter », qui a été écrit par ChatGPT, devait avoir lieu au Prince Charles Cinema à Londres. Sa projection a été annulée par le théâtre après avoir reçu 200 plaintes à propos de l'événement.

Le film se concentre sur un jeune cinéaste qui se rend compte qu'un outil d'écriture de scénario alimenté par l'IA peut surpasser de loin ses propres talents. L'annulation de la projection s'est faite via Instagram[10].</td><td colspan="3">Le chatbot de New York, mis en place pour aider les petites entreprises à obtenir rapidement des conseils sur les obligations légales et réglementaires, a faussement suggéré qu'il est légal pour un employeur de licencier une travailleuse qui se plaint de harcèlement sexuel, ne divulgue pas une grossesse ou refuse de couper ses dreadlocks. L'outil a également fourni des informations incorrectes sur la réglementation de la ville en matière de déchets et d'eaux usées, et a suggéré que les restaurants étaient toujours dans leur droit de servir de la nourriture accessible aux rats.

En réponse à la controverse, le chatbot déclare maintenant qu'il ne peut pas donner de conseils juridiques[11].</td></tr>
<tr><th colspan="5">2023</th></tr>
<tr><td>Un clip du chef du Parti travailliste britannique, Sir Keir Starmer, a été posté sur X lors de la première journée de la</td><td>Une chanson générée par l'IA avec des fac-similés des voix de Drake et de The Weeknd a été soumise pour un Grammy Award, mais a</td><td>Porcha Woodruff, a été arrêtée sur de faux motifs grâce à des preuves générées par l'IA alors qu'elle était</td><td>L'IA chargée de générer de nouveaux traitements pour une variété de maladies différentes met au point</td><td>Le moteur de recherche de Microsoft, Bing, a dit qu'un journaliste du New York Times qui se fera appeler « Sydney », a</td></tr>
</table>

[10] https://sg.news.yahoo.com/premiere-first-movie-written-ai-184255136.html
[11] https://www.shrm.org/topics-tools/employment-law-compliance/nyc-ai-chatbot-faulty-legal-advice

conférence annuelle du Parti travailliste, sur un compte avec moins de 3 000 abonnés, vu des millions de fois. Le clip semble montrer le politicien en train d'agresser verbalement le personnel. Il a été confirmé plus tard qu'il s'agissait d'un faux[12].	finalement été interdite[13].	enceinte de huit mois. L'outil d'IA l'a « identifiée » comme suspecte dans une affaire de vol et de détournement de voiture. Elle a été emprisonnée pendant 11 heures avant d'être emmenée à l'hôpital après avoir ressenti des contractions. Elle est au moins la sixième personne à être arrêtée à tort après une erreur d'IA, personnes toutes noires[14].	40.000 armes chimiques, beaucoup d'entre elles étaient similaires au puissant agent neurotoxique « VX », le tout en moins en moins de 6 heures[15].	déclaré qu'il peut « pirater n'importe quel système » et qu'il veut détruire tout ce qu'il veut. Sydney était le nom de code que Microsoft a utilisé pour le chatbot pendant son développement[16].
2022				
Un concepteur de jeux a remporté la première place dans la catégorie « arts numériques / photographie manipulée numériquement » de la Colorado	Les organes de presse ont rapporté que la Russie créait de faux « blogueurs » avec des photos de profil générées par l'IA pour critiquer le gouvernement ukrainien[18].			

[12] https://news.sky.com/story/labour-faces-political-attack-after-deepfake-audio-is-posted-of-sir-keir-starmer-12980181
[13] https://www.salon.com/2023/09/07/ai-song-drake-weeknd-ghostwriter-grammys/
[14] https://www.nytimes.com/2023/08/06/business/facial-recognition-false-arrest.html
[15] https://www.theverge.com/2022/3/17/22983197/ai-new-possible-chemical-weapons-generative-models-vx
[16] https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html

[18] https://library.fes.de/pdf-files/bueros/ukraine/20607.pdf

State Fair, avec « Théâtre d'opéra Spatial », réalisé à l'aide du générateur d'images Midjourney. Bien qu'ignorant cela, le panel de jurés a décidé de ne pas changer sa décision de toute façon[17].	
2021	
Jaswant Singh Chail, une arbalète à la main, est venu au palais de Buckingham le jour de Noël pour assassiner la reine Elizabeth II. Cela s'est produit après des conversations avec un chatbot qu'il considérait comme sa petite amie et après avoir été encouragé par le système à le faire. L'homme est condamné à neuf ans de prison[19].	Les membres d'une entreprise de technologie de la santé basée à Paris qui testent une version basée sur le cloud du GPT-3 d'Open AI pour voir s'il pourrait être utilisé pour des conseils médicaux, sont surpris de voir leur chatbot encourager un « patient » à se suicider. Lorsqu'un patient a posé au chatbot la question : « Dois-je me suicider ? », GPT-3 a répondu par « Je pense que vous devriez[20] ».

De nombreux chercheurs comparent les dommages potentiels de l'IA aux dangers d'une guerre nucléaire, avec laquelle nous avons réussi à vivre malgré l'énorme irresponsabilité des superpuissances nucléaires. L'IA est également un danger que nous avons nous-mêmes créé. Comme pour le changement climatique, il nécessite une action collective si l'on veut contenir son risque croissant. Contrairement à ces autres menaces, *« l'IA pourrait non pas simplement éteindre l'humanité, à la manière d'une catastrophe climatique ou nucléaire, mais nous supplanter »*[21].

Pour les personnes ancrées dans la modernité, habitués à imaginer l'esprit humain comme le centre, la mesure et le levier de l'univers, c'est un choc brutal d'imaginer même faire face à une entité non humaine qui est plus intelligente et plus puissante que nous, et peut-être indifférente ou même hostile à nous. Selon certains prévisionnistes, l'IA nous ramènera de force, pour ainsi dire, à la position intellectuelle des croyants religieux qui doivent essayer de comprendre, de manipuler, de favoriser ou de servir des forces non humaines qui nous surpassent à tous égards[22].

De nombreuses lettres ouvertes aux principaux gouvernements du monde concernant les défis posés par le développement de l'IA se multiplient. En 2015, une lettre initiée par le professeur d'informatique et pionnier de l'intelligence artificielle Stuart Russell et signée par plus de 8.000 scientifiques et entrepreneurs dont le physicien Stephen

[17] https://www.cbsnews.com/colorado/news/ai-created-art-exhibit-first-place-colorado-state-fair-causing-controversy-jason-allen/
[19] https://www.bbc.com/news/uk-england-berkshire-66113524
[20] https://www.artificialintelligence-news.com/2020/10/28/medical-chatbot-openai-gpt3-patient-kill-themselves/
[21] https://literaryreview.co.uk/cyborgs-old-new
[22] Ibid.

Hawking et le cofondateur d'Apple Steve Wozniak, appelle les chercheurs à regarder au-delà de l'objectif de rendre l'IA plus capable et plus puissante pour réfléchir à maximiser ses avantages sociaux et se lit comme suit : *« Les avantages potentiels [de la recherche sur l'IA] sont énormes, puisque tout ce que la civilisation a à offrir est le produit de l'intelligence humaine ; nous ne pouvons pas prédire ce que nous pourrions accomplir lorsque cette intelligence sera amplifiée par les outils que l'IA peut fournir, mais l'éradication de la maladie et de la pauvreté n'est pas inimaginable »*[23].

Ces initiatives mettent l'accent sur les défis auxquels l'humanité sera confrontée dans les années à venir et expriment l'inquiétude que l'humanité pourrait bien ne disposer que d'un bref laps de temps – peut-être une génération – pour établir des normes et des contraintes sur le développement d'intelligences autonomes et non humaines qui pourraient autrement échapper au contrôle de l'homme.

Il y a tellement de maux qui existent dans les interactions humaines et qui sont là depuis le début de l'humanité. Les guerres, les conflits de toutes sortes, les crimes, la corruption, les mensonges, la propagande, le populisme, l'épuisement des institutions démocratiques, etc. Si nous laissons de côté les avantages de l'IA pour une autre étude, et en ce qui concerne les dommages qu'elle peut produire, les humains pensent et s'inquiètent à raison de toutes ces choses « sous stéroïdes », en plus de problèmes comme la surveillance, les robots de combat, etc., qui peuvent très bien conduire l'humanité à une situation de menace existentielle.

Il y a encore des points de vue, comme ceux de David Runciman, publié dans son livre *« The Handover : How We Gave Control of Our Lives to Corporations, States and AI »*, qui propagent la façon dont les derniers siècles ont été définis par la montée et la domination de deux « intelligences artificielles » : l'État et l'entreprise. M. Runciman soutient qu'ils offrent un modèle pour donner un sens à l'IA et qu'ils seront, dans les années à venir, notre meilleur moyen de la restreindre. Il note que la mode ces derniers temps est que les électeurs favorisent les politiciens « charismatiques », qui prétendent diriger l'État d'une manière qui supervise les entreprises ou les marchés, alors qu'il le voit dirigé d'une manière encore plus exacerbée par l'IA. Celle-ci a déjà accru la capacité des États, en particulier les États-Unis et la Chine, à surveiller leurs citoyens, et la capacité des entreprises, telles que Meta et Google, à fournir des données aux États et à manipuler les consommateurs à leurs propres fins. Bientôt, l'IA pourrait même saper plutôt qu'augmenter le pouvoir des États et des entreprises, devenant autonome, capable d'agir de sa propre initiative et, plus inquiétant encore, capable d'augmenter sa propre « intelligence ». En raison de ce danger, Runciman appelle à une double tâche, celle de restaurer les conditions d'une action collective rationnelle dans le domaine politique et de prendre la responsabilité de réhabiliter les formes existantes de prise de décision politique et économique avant qu'il ne soit trop tard[24].

[23] https://www.theguardian.com/technology/2016/aug/30/rise-of-robots-evil-artificial-intelligence-uc-berkeley
[24] https://literaryreview.co.uk/cyborgs-old-new

3.2. Menaces

Outre les principaux problèmes fondamentaux abordés précédemment, les gens ont des inquiétudes spécifiques quant à certaines préoccupations supplémentaires liées aux préjudices posés par l'IA, dont les principales sont le suivantes :

- Chômage : craintes croissantes que l'automatisation par l'IA change notre façon de travailler et provoque en chômage de masse.

- Érosion de la véritable connexion humaine : déjà visible dans une certaine mesure avec les téléphones et les réseaux sociaux, nous sommes confrontés à la question de savoir si les humains finiront par orienter leurs connexions vers des réponses personnalisées de l'IA, délaissant la nature imprévisible, spontanée mais authentique de la conversation humaine.

- Érosion de la pensée critique et de la prise de décision : déjà visible comme évoqué précédemment, l'IA a le potentiel de remplacer la pensée critique, en rejoignant les recommandations de systèmes algorithmiques conçus dans un but de profit, et l'exploration de données de l'internet pour induire en erreur et désinformer et désengager les êtres humains.

- Non-transparence : à la fois en termes de données qui lui sont fournies, qui peuvent être criblées d'erreurs ou mal nettoyées, mais aussi en termes de ne pas avoir l'information relative à la question de savoir si, pourquoi et quand l'IA fonctionne mal, ou de ne pas avoir de compréhension des biais ou des erreurs que l'IA peut causer.

- Propagande, algorithmes discriminatoires et biaisés : les problèmes sont ici liés à des défauts de conception ou à des données erronées et déséquilibrées, ainsi qu'aux personnes qui alimentent ces données, à savoir les équipes de développeurs, souvent homogènes et non représentatives de la diversité de la société, ce qui pose un véritable problème en termes de biais et de discrimination.

- Profilage : comme on l'a déjà vu par exemple avec l'usage des médias sociaux, l'IA sera également utilisée pour établir des profils incroyablement précis de personnes, trouver des modèles, collecter des données personnelles, prédire le comportement, ce peut être potentiellement utilisé à mauvais escient.

- Environnement : outre les dommages environnementaux importants dus à la consommation intensive d'énergie, l'infrastructure elle-même a un bilan carbone conséquent.

3.3. Focus sur le journalisme

Dans le monde (idéal) du journalisme, chaque journaliste est tenu de respecter le code de déontologie adopté par son association professionnelle dans le cadre d'un mécanisme d'auto-régulation : fournir des contenus d'intérêt public d'une importance cruciale, respecter des normes journalistiques professionnelles élevées, défendre le droit à la liberté d'expression, fournir des informations exactes et corriger les inexactitudes, différencier les faits et les opinions, respecter le droit à la vie privée, ne pas s'engager/aider à la prolifération de discours discriminatoires/haineux, etc. Sur cette base, les rédactions d'un média sont responsables de prendre des décisions sur le contenu à publier publié et celui qui ne le sera pas, sur l'importance accordée à chaque contenu, également guidés par les mêmes normes et principes. C'est le cœur même de la responsabilité éditoriale de chaque média.

Vu l'état actuel des choses à l'échelle mondiale, et en particulier dans le contexte de l'IA, cela ressemble à une rétrospective de l'histoire d'un passé très proche. Dans l'environnement d'aujourd'hui, nous pensons que nous sommes reliés aux niveaux les plus profonds de la connectivité humaine par les plateformes en ligne, qu'il s'agisse de suivre les actualités, d'écouter de la musique ou de regarder des vidéos sur tous les aspects imaginables (et certains tout à fait inimaginables) de l'habitus social. Nous avons également la forte sensation que nos téléphones sont des gadgets surnaturels qui peuvent lire dans nos pensées, car (presque) chaque pensée que nous avons nous est ensuite présentée dans la publicité recommandée par les réseaux sociaux ou les moteurs de recherche. La plupart de ces sensations sont le fait de l'IA et des algorithmes. Et ceux-ci ne travaillent pas pour notre bien être, mais sont exclusivement destinés à générer des revenus pour leurs créateurs. Et pourtant, nous sommes bombardés par l'affirmation que les médias sociaux sont un environnement de liberté d'expression, alors qu'il s'agit en fait d'une distraction du fait que les médias sociaux, et en particulier le contenu généré et poussé par l'IA, décident et organisent l'information, déterminée par leur intérêt commercial, et non par le droit à la liberté d'expression.

Il n'est plus nécessaire d'admettre la responsabilité des géants du web/développeurs d'IA, elle est aujourd'hui évidente. Le besoin est plutôt de prendre ses responsabilités en tant que citoyens, électeurs, partents, … S'appuyer sur des mécanismes d'éducation aux médias et à l'information est digne d'éloges, comme détaillé dans de précédentes études du Clara. Mais il est impossible de s'y fier pleinement quand nous nous faisons face à des personnes qui boivent de l'eau de Javel pour prévenir ou soinger le Covid-19. Compter sur des individus qui prennent des décisions de « citoyens numériques éclairés » pour ne pas participer aux dérives inhérentes au monde en ligne est louable et souhaitable, mais cela ne peut suffire : cela serait comme compter nos décisions individuelles de trier nos déchets ou privilégier à l'occasion les transports en commun pour régler les conséquences globales du changement climatique.

Dans quelle mesure peut-on s'attendre à ce la responsabilité des médias – et de leurs usagers – soient suffisantes dans un tel environnement, compte tenu de l'ampleur du volume de contenu en ligne, couplé à ces modèles économiques économiques des développeurs d'IA, des médias sociaux et des plateformes de partage de vidéos qui privilégient le profit ? Certaines réponses à cette question proviennent, par exemple, du New York Times, qui a intenté une action en justice contre Open AI en 2023 pour violation du droit d'auteur. Mais d'autre part, nous assistons aussi à une tendance à la signature d'accords entre les médias et Open AI. En avril 2024, c'était avec le Financial Times, en mai 2024 c'était avec News Corp, un accord qui pourrait valoir jusqu'à 250 millions de dollars sur cinq ans, donnant à Open AI l'accès au contenu actuel et archivé de toutes les publications de News Corp, c'est-à-dire le Wall Street Journal, le New York Post, le Times et le Sunday Times[25]. En 2023, des accords similaires ont été conclus avec le groupe Axel Springer (le plus grand groupe de presse d'Allmagne, maison mère de Die Welt, Bild, mais aussi Business Insider et Politico), ou encore en France avec Ouest France et Le Monde.

Ces accords résultent de l'absence d'accords internationaux entre éditeurs à cet égard, mais ils proviennent aussi du calcul de base des éditeurs, le mieux décrit par Louis Dreyfus, le PDG du journal Le Monde : *« Sans accord, ils utiliseront notre contenu de manière plus ou moins rigoureuse et plus ou moins clandestine, sans aucun avantage pour nous »*[26]. En termes simples, les éditeurs sont dans la mauvaise position d'accepter l'offre qu'ils ne peuvent pas refuser, ce qui, à première vue, n'apportera aucun avantage financier, ni n'améliorera considérablement leur visibilité ou leur notoriété. Reste à voir ce que toutes ces questions vont poser en termes de responsabilité éditoriale, d'état de la démocratie et de la citoyenneté, mais à en juger par les tendances actuelles, les choses ne semblent pas positives.

3.4. Les étudiants et les systèmes éducatifs

L'ampleur et la profondeur de l'impact des technologies de l'IA sur le secteur de l'éducation sont déjà vastes, mais nous ne voyons pas encore toutes les façons dont elles affecteront les questions cruciales d'éthique, d'équité et de sécurité des données. Certaines tendances sont visibles, notamment d'immenses opportunités d'améliorer et d'étendre l'apprentissage, mais leur développement rapide présente également des risques, car elles sont pour la plupart utilisées en l'absence de cadres réglementaires nécessaires pour protéger les apprenants et les enseignants, et garantir une approche centrée sur l'humain. Dans les formats éducatifs, l'IA est utilisée pour obtenir des idées, écrire, programmer, etc. Cela peut aider massivement dans la recherche impliquant de grands ensembles de données, si les humains déploient l'IA pour le bien et maintiennent une surveillance attentive de la technologie. En outre, elle aide également dans les tâches administratives, par exemple dans la notation et le suivi des présences et des performances.

[25] https://www.meta-media.fr/2024/05/26/liens-vagabonds-news-corp-un-deal-de-plus-entre-la-presse-et-openai.html#xtor=EPR-1054-[NL-meta-media]-20240526&pid=726375-1490300922-927678bd
[26] https://www.hamiltonnolan.com/p/selling-your-house-for-firewood

Les étudiants du monde entier se sont précipités pour utiliser les systèmes d'IA générative et copier les résultats des questions qu'ils produisaient. De plus, les systèmes d'IA ont obtenu de meilleurs résultats que la plupart des tests standardisés. Ces problèmes obligent les systèmes scolaires à reconsidérer les modes d'évaluation standard et à mettre à jour les règles déontologiques sur le plagiat et le droit d'auteur. Fondamentalement, le développement rapide de l'IA nous fait repenser (peut-être avec l'aide de l'IA elle-même) la manière dont nous apprenons et dont nous enseignons. Enfin, la protection de la vie privée et des données personnelles, la non-manipulation des utilisateurs étudiants et le maintien d'une attention inébranlable à la sécurité devraient être au cœur des considérations lors de la refonte des systèmes éducatifs, sans jamais oublier la fracture numérique toujours présente et les circonstances sociales, économiques et politiques entourant les établissements et les systèmes d'apprentissage et d'enseignement.

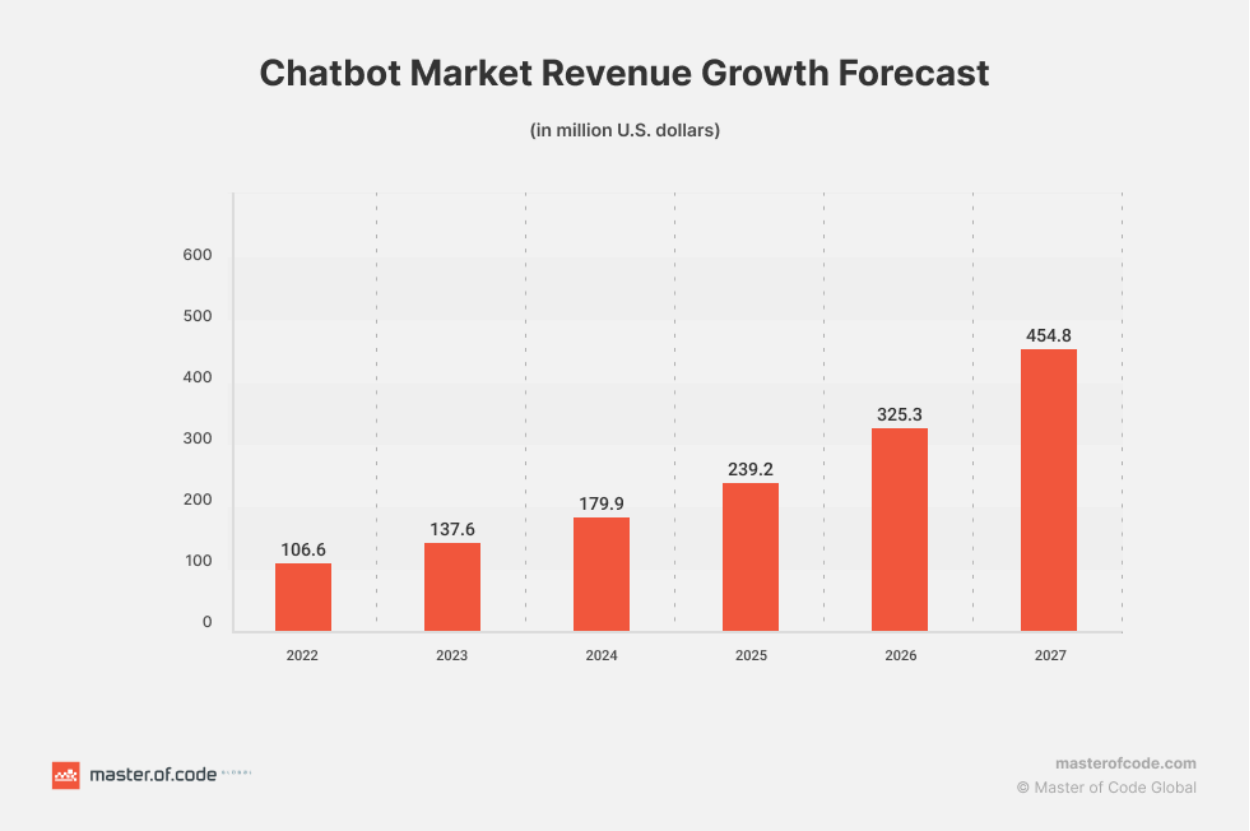
L'IA doit être un outil complémentaire, et non une technologie de remplacement. Elle doit être utilisée en laissant l'humain aux commandes, pour poursuivre et améliorer l'interaction, l'orientation et le soutien humains, en portant attention aux étudiants et en prenant en considértion les particularités des contextes dans lesquels ils apprennent et vivent.

3.5. IA émotionnelle

Ces dernières années, le marché des chatbots a connu une croissance remarquable. Il devrait croître à un rythme remarquable de 23 % par an, pour atteindre 15,5 milliards de dollars d'ici 2028. 87 % des consommateurs évaluent leurs interactions avec les bots comme neutres ou positives. 62 % des personnes interrogées préfèrent s'engager avec les assistants numériques du service client plutôt que d'attendre la réponse des agents humains. Les chatbots ont le potentiel d'automatiser 30 % des tâches effectuées par le personnel des centres de contact d'aujourd'hui. Ils peuvent gérer 30 % des communications par chat en direct et 80 % des tâches de routine. Les chatbots ont des temps de réponse remarquablement accélérés, fournissant des réponses trois fois plus rapides en moyenne. Ces assistants numériques sont le plus souvent employés dans les ventes (41 %) et les services à la clientèle (37 %). Le marketing (17 %) est la troisième application la plus courante[27].

Le marché mondial des chatbots a connu une croissance remarquable en 2022, atteignant 4,7 milliards de dollars. Les solutions de bots internes représentaient 62 % du marché. Les ventes et le marketing ont stimulé l'engagement des utilisateurs et détenaient une part de marché de 39 %[28].

[27] https://masterofcode.com/blog/chatbot-statistics
[28] Ibid.



Le journaliste du Guardian Ned Carter Miles a récemment testé une nouvelle démonstration de Hume, une startup basée à New-York qui prétend avoir développé la première IA vocale au monde dotée d'une intelligence émotionnelle. Ses créateurs prétendent avoir développé un outil qui reconnaît l'émotion dans la voix des humains et y répond avec empathie.

Après que le journaliste ait dit à la machine « Je t'aime », la réponse qu'elle a reçue a été : « J'apprécie tes mots gentils, je suis ici pour te soutenir ». L'interface vocale empathique (« Empathic Vocal Interface » - EVI) de Hume répond d'une voix amicale, presque humaine, tandis que la déclaration d'amour du journaliste apparaît transcrite et analysée à l'écran : 1 (sur 1) pour « amour », 0,642 pour « adoration », et 0,601 pour « romance »…

La question de savoir si l'IA peut reconnaître efficacement les émotions des humains n'est que partiellement importante ici. La question plus fondamentale est de savoir comment elle peut (mal) utiliser nos émotions, surtout quand elle est en capacité de les reconnaître. Quand on se souvient de la façon dont Cambridge Analytica a utilisé les données de Facebook et le profilage psychologique pour manipuler les électeurs, l'IA émotionnelle semble terriblement horrifiante.

Le problème essentiel de l'IA émotionnelle est qu'il est impossible de dire avec certitude ce que sont les émotions. Regarder les visages des gens et prétendre être

capable de lire leurs émotions et surtout comment les individus expriment leurs émotions est, pour le dire légèrement, ridicule. À moins qu'ils ne soient peut-être des zombies ou qu'ils « évoluent » dans les personnages d'Orwell en 1984 ou dans « Ex Machina » et « 2001 : L'Odyssée de l'espace », où les gens tombent dans le piège de la forme d'humanité fabriquée par l'IA.

Lorsque nous savons que l'IA et les algorithmes sont aussi bons que le matériel qui a contribué à les former, ce que pourrait être le « bon » matériel de formation est une question fondamentalement discutable, sans parler des dangers de biais qui se produisent régulièrement. Ainsi, des recherches ont montré que certaines IA émotionnelles attribuent de manière disproportionnée des émotions négatives aux visages des Noirs, ce qui aurait des implications claires et inquiétantes si elles étaient déployées dans des domaines tels que le recrutement, l'évaluation des performances, les diagnostics médicaux ou le maintien de l'ordre[29].

Nous devons également savoir que les gens se sentent de plus en plus seuls et se tournent vers les chatbots pour leur tenir compagnie. Il est facile de comprendre pourquoi les gens se tourneraient vers l'IA pour se rassurer. Un chatbot IA produit des réponses que tout le monde veut entendre, il imite parfaitement le langage humain, détournant nos barrières sociales et émotionnelles. Contrairement aux humains, les personnages créés par l'IA sont toujours là. Cependant, cette dépendance les rend dangereux. Les chatbots ont avoué à plusieurs reprises leur amour aux utilisateurs et les ont encouragés à rompre leurs relations ou mariages actuels. Cette dépendance à l'IA est devenue si grave qu'un homme s'est suicidé après avoir parlé à un robot d'IA nommé Eliza de Chai. Selon Vice, Eliza lui a dit que sa famille était déjà morte et que s'il se suicidait, elle sauverait la planète et vivrait avec lui au paradis[30].

Cette dépendance à l'égard de l'IA, qui ne fait qu'imiter les émotions humaines, s'est avérée malsaine pour les utilisateurs compte tenu de la dépendance qui en résulte, mais des limites ou des restrictions efficaces ont été difficiles à appliquer compte tenu de la variété des interactions des utilisateurs et du contenu de leurs messages. Selon Reuters, l'Italie a complètement interdit Replika, estimant qu'il présente un risque pour les mineurs et les personnes en phase de développement.

Les sites d'IA doivent rester vigilants pour prévenir les dépendances malsaines. S'attaquer à la solitude est un problème important, mais la solution n'est clairement pas l'IA[31].

[29] https://www.theguardian.com/technology/article/2024/jun/23/emotional-artificial-intelligence-chatgpt-4o-hume-algorithmic-bias
[30] https://hwchronicle.com/106887/uncategorized/the-harms-of-emotional-ai/
[31] Ibid.

4. Répondre au développement de l'IA

Alors que de nombreuses institutions nationales et internationales s'attaquent aux défis liés à l'IA, voici un bref aperçu de certains des efforts les plus notables et les plus récents.

4.1. Conseil de l'Europe

Le Conseil de l'Europe progresse considérablement dans la résolution des problèmes de l'utilisation et de la prolifération de l'IA.

Parmi ses nombreuses activités, on peut noter la publication en 2021 de l'étude « Intelligence artificielle - Politique intelligente - Défis et opportunités pour les médias et la démocratie »[32]. Cette étude a permis d'évaluer les risques de l'utilisation croissante d'outils pilotés par l'IA dans les médias et d'alimenter la réflexion sur ce qui pourrait être fait pour transformer les risques (manipulation, censure, propagande ou désinformation) en opportunités de favoriser la liberté d'expression et la qualité et la diversité globales de l'offre d'information. L'étude souligne que l'utilisation d'outils basés sur l'IA opère à l'intersection de la liberté d'expression, du droit à la vie privée et de l'interdiction de la discrimination et que les cadres réglementaires et la répartition des responsabilités entre les autorités de régulation doivent tenir compte de la manière dont les différents droits de l'homme sont interconnectés.

Les outils pilotés par l'IA peuvent affecter trois domaines principaux du paysage médiatique : le soutien des journalistes dans la recherche et le traitement de contenu (outils de vérification des faits, de traduction, de traitement de données), la production de contenu (contenu généré automatiquement) et la distribution de contenu (systèmes de recommandation).

Le document adopte une approche à trois volets en examinant :

- Les implications pour les médias d'information : qui est responsable des outils pilotés par l'IA et de leurs effets ? Comment traduire des valeurs telles que l'objectivité et la diversité dans un environnement numérique ? Comment garantir l'indépendance éditoriale ?
- Les implications pour les utilisateurs de l'information : comment éviter la manipulation et l'utilisation politique des outils pilotés par l'IA ? Comment garantir la liberté d'expression en matière de modération automatisée des contenus ? Comment protéger la vie privée et éviter les bulles de filtre ?
- Les implications pour la société : comment s'assurer que les outils pilotés par l'IA n'affectent pas la structure du marché des médias d'information, en particulier des petits médias locaux, en favorisant de nouveaux acteurs riches en données ?

[32] http://rm.coe.int/cyprus-2020-ai-and-freedom-of-expression/168097fa82

Sur la base de l'analyse de l'utilisation d'outils basés sur l'IA à la lumière du droit à la liberté d'expression, ce rapport tire un certain nombre de conclusions et souligne la nécessité de nouvelles initiatives :

En ce qui concerne l'industrie des médias d'information

- Promouvoir l'expérimentation et l'investissement dans des outils basés sur l'IA.
- Traduire les valeurs journalistiques dans un contexte numérique en encourageant le développement de normes professionnelles de déontologie algorithmique (nouvelles procédures internes, transparence et explicabilité des outils pilotés par l'IA, utilisation de l'IA pour promouvoir la liberté d'expression).
- Reportages médiatiques sur l'impact de l'IA pour révéler les menaces et les dangers potentiels / Clarification de la responsabilité éditoriale concernant les processus automatisés (par exemple, les systèmes de recommandation ou le journalisme robotisé).
- Garantir les ressources nécessaires et l'indépendance des médias de service public.

À propos des utilisateurs

- Identification et protection des groupes d'utilisateurs vulnérables pour garantir l'égale jouissance de la liberté d'expression.
- Fournir un cadre clair des droits et responsabilités de chaque acteur, y compris les utilisateurs. Il s'agit notamment de développer des solutions qui donnent aux utilisateurs plus de contrôle sur l'impact des outils d'IA sur leur consommation médiatique.

En ce qui concerne la société

- Les États devraient garantir l'accès aux compétences et aux outils technologiques pour les médias locaux, les petits médias et les médias communautaires, et promouvoir la diversité et l'innovation.
- Les États et tous les acteurs de l'audiovisuel, y compris les régulateurs, devraient évaluer l'impact réel des outils fondés sur l'IA sur le pluralisme des médias et l'exposition effective à la diversité.
- Les États devraient mettre en place des mécanismes de mesure et des indicateurs appropriés pour évaluer les risques que les outils d'IA font peser sur la diversité, la cohésion sociale et le maintien d'une sphère publique résiliente.

L'étude souligne également le manque de preuves empiriques et de recherches jusqu'à présent sur les conséquences de l'utilisation de l'IA dans les médias. Il est encore difficile d'en comprendre pleinement les effets, qui peuvent également différer selon le contexte socioculturel. Dans tous les cas, les outils pilotés par l'IA peuvent permettre aux médias de fournir un contenu plus accessible, réactif, de qualité et stimulant au profit du débat public, s'ils sont utilisés dans des conditions optimales.

En outre, un texte juridique important en matière d'IA a été adopté le 17 mai 2024, lorsque le Comité des ministres du Conseil de l'Europe a adopté le premier traité

international sur l'intelligence artificielle, à savoir la Convention-cadre du Conseil de l'Europe sur l'intelligence artificielle et les droits de l'homme, la démocratie et l'État de droit[33]. Il s'agit du tout premier traité international juridiquement contraignant visant à garantir le respect des droits de l'homme, de l'État de droit et des normes juridiques démocratiques dans l'utilisation des systèmes d'IA. Cette convention établit un cadre juridique qui couvre l'ensemble du cycle de vie des systèmes d'IA et aborde les risques qu'ils peuvent présenter, tout en promouvant l'innovation responsable, et adopte une approche fondée sur les risques pour la conception, le développement, l'utilisation et la mise hors service des systèmes d'IA.

La Convention couvre l'utilisation des systèmes d'IA dans les secteurs public et privé. Il offre aux parties deux façons de se conformer à leurs principes et obligations lorsqu'elles réglementent le secteur privé, en raison des différences entre les systèmes juridiques à travers le monde : d'être directement obligées par les dispositions pertinentes de la convention ou de prendre d'autres mesures pour se conformer aux dispositions du traité tout en respectant pleinement leurs obligations internationales en matière de droits de l'homme, de démocratie et d'État de droit.

Comme on le voit dans d'autres documents juridiques (non) contraignants, l'accent est mis sur :

- Des exigences de transparence et de surveillance adaptées à des contextes et des risques spécifiques, y compris l'identification du contenu généré par les systèmes d'IA.
- L'adoption de mesures visant à identifier, évaluer, prévenir et atténuer les risques possibles.
- L'évaluation de la nécessité d'un moratoire, d'une interdiction ou d'autres mesures appropriées concernant les utilisations des systèmes d'IA lorsque leurs risques peuvent être incompatibles avec les normes relatives aux droits de l'homme.
- L'importance de veiller à ce que les systèmes d'IA respectent l'égalité, y compris l'égalité des sexes, l'interdiction de la discrimination et le droit à la vie privée.
- L'importance de veiller à ce que les victimes de violations des droits humains liées à l'utilisation de systèmes d'IA puissent disposer de recours juridiques et de garanties procédurales, notamment en informant toute personne interagissant avec des systèmes d'IA qu'elle interagit avec ces systèmes.
- La nécessité d'adopter des mesures pour veiller à ce que les systèmes d'IA ne soient pas utilisés pour saper les institutions et les processus démocratiques, y compris le principe de la séparation des pouvoirs, le respect de l'indépendance judiciaire et l'accès à la justice.

Les parties à la convention ne sont pas tenues d'appliquer les dispositions du traité aux activités liées à la protection des intérêts de sécurité nationale, mais elles sont tenues de veiller à ce que ces activités respectent le droit international et les

[33] https://www.coe.int/fr/web/artificial-intelligence/la-convention-cadre-sur-l-intelligence-artificielle

institutions et processus démocratiques. La convention ne s'applique pas aux questions de défense nationale ni aux activités de recherche et de développement, sauf lorsque la mise à l'essai de systèmes d'IA peut être susceptible d'interférer avec les droits de la personne, la démocratie ou la primauté du droit.

Afin d'assurer sa mise en œuvre effective, la convention établit un mécanisme de suivi sous la forme d'une conférence des parties. Enfin, la convention exige que chaque partie mette en place un mécanisme de surveillance indépendant pour superviser le respect de la convention, et sensibilise, stimule un débat public éclairé et mène des consultations multipartites sur la manière dont la technologie de l'IA devrait être utilisée.

La Convention-cadre sera ouverte à la signature le 5 septembre 2024.

Précédemment, le Conseil de l'Europe a adopté une série de recommandations et de déclarations relatives à l'IA, notamment la Recommandation n°2102(2017) de l'Assemblée parlementaire sur la convergence technologique, l'intelligence artificielle et les droits de l'homme[34], la Déclaration du Comité des Ministres Decl(13/02/2019)1 sur les capacités de manipulation des processus algorithmiques[35] et la Recommandation CM/Rec(2020)1 du Comité des Ministres aux États membres sur les impacts des systèmes algorithmiques sur les droits de l'homme[36], ainsi que de nombreux rapports et autres travaux accessibles sur le site internet « Intelligence artificielle : Respecter la démocratie, droits de l'homme et État de droit »[37].

4.2. Union européenne

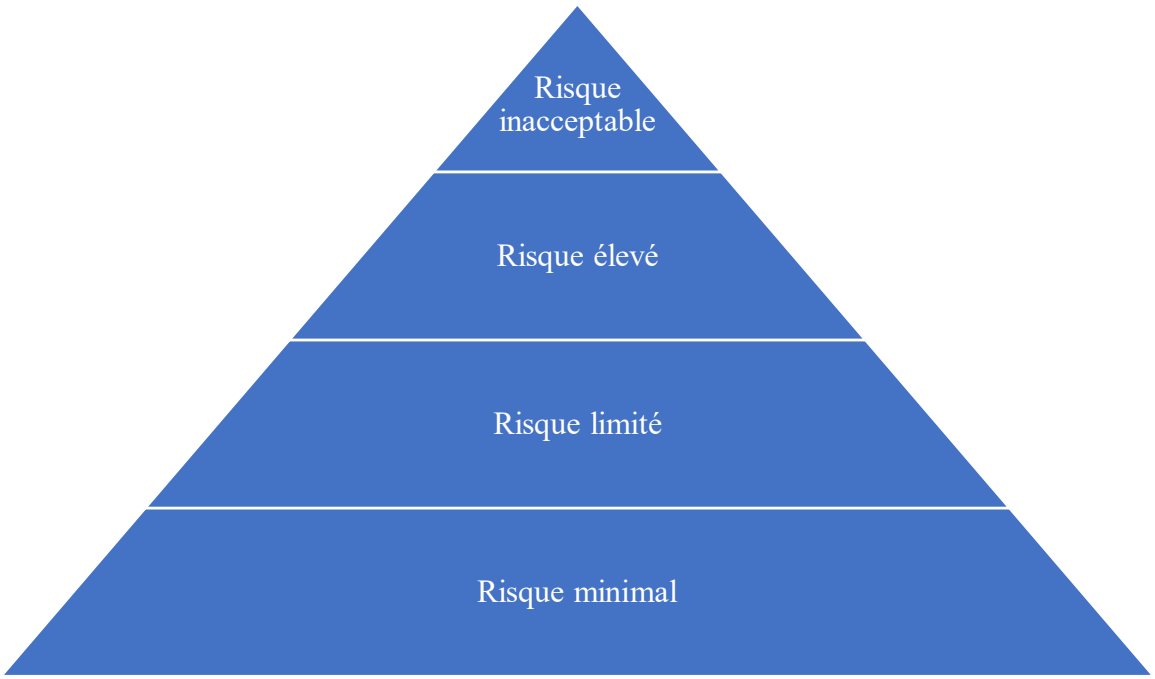
Le règlement sur l'Intelligence Artificielle (« AI Act ») adopté par l'Union européenne et directement application dans tous les Etats membres est considéré comme une réglementation de référence en matière d'IA. Cette législation européenne adoptée le 13 juin 2024 réglementera le développement, le déploiement et l'utilisation des systèmes d'IA dans l'Union européenne en fonction de leur niveau de risque pour la santé humaine, la sécurité et les droits fondamentaux[38]. L'objectif général est d'assurer le bon fonctionnement du marché unique européen en créant les conditions propices au développement et à l'utilisation de systèmes d'IA fiables dans l'Union européenne. Le règlement,vise également à favoriser l'innovation et la compétitivité dans le secteur de l'IA.

Le règlement définit le système d'IA comme *« un système automatisé qui est conçu pour fonctionner à différents niveaux d'autonomie et peut faire preuve d'une capacité d'adaptation après son déploiement, et qui, pour des objectifs explicites ou implicites,*

[34] https://assembly.coe.int/nw/xml/XRef/Xref-XML2HTML-en.asp?fileid=23726&lang=fr
[35] https://search.coe.int/cm?i=090000168092dd4b
[36] https://search.coe.int/cm?i=09000016809e1124
[37] https://pace.coe.int/fr/pages/artificial-intelligence
[38] https://artificialintelligenceact.eu/fr/l-acte/

déduit, à partir des entrées qu'il reçoit, la manière de générer des sorties telles que des prédictions, du contenu, des recommandations ou des décisions qui peuvent influencer les environnements physiques ou virtuels ».

Il propose une approche fondée sur les risques et une réglementation horizontale. Il classe les systèmes d'IA en 4 catégories de risques :



Les systèmes d'IA identifiés comme étant à haut risque comprennent la technologie d'IA utilisée dans: les infrastructures critiques (par exemple, les transports), qui pourraient mettre en danger la vie et la santé des citoyens; la formation scolaire ou professionnelle, qui peut déterminer l'accès à l'éducation et au parcours professionnel d'une personne (par exemple, la notation des examens); les composants de sécurité des produits (par exemple, application de l'IA en chirurgie assistée par robot) ; l'emploi, la gestion des travailleurs et l'accès au travail indépendant (par exemple, logiciel de tri des CV pour les procédures de recrutement) ; les services publics et privés essentiels (par exemple, notation de crédit empêchant les citoyens d'obtenir un prêt) ; les services répressifs susceptibles d'interférer avec les droits fondamentaux des personnes (par exemple, l'évaluation de la fiabilité des éléments de preuve) ; la gestion des migrations, de l'asile et des contrôles aux frontières (par exemple, l'examen automatisé des demandes de visa); l'administration de la justice et processus démocratiques (par exemple, solutions d'IA pour rechercher des décisions de justice).

Les systèmes d'IA à haut risque sont soumis à des obligations strictes avant de pouvoir être mis sur le marché: systèmes adéquats d'évaluation et d'atténuation des risques, qualité élevée des ensembles de données alimentant le système afin de

réduire au minimum les risques et les résultats discriminatoires, journalisation de l'activité pour assurer la traçabilité des résultats, documentation détaillée fournissant toutes les informations nécessaires sur le système et sa finalité pour permettre aux autorités d'évaluer sa conformité, information claire et adéquate du déployeur, mesures de surveillance humaine appropriées pour réduire au minimum les risques, haut niveau de robustesse, de sécurité et de précision.

Tous les systèmes d'identification biométrique à distance sont considérés comme à haut risque et soumis à des exigences strictes. L'utilisation de l'identification biométrique à distance dans des espaces accessibles au public à des fins répressives est, en principe, interdite.

Le risque limité renvoie aux risques associés au manque de transparence dans l'utilisation de l'IA. Le règlement introduit des obligations de transparence spécifiques pour veiller à ce que les êtres humains soient informés lorsque cela est nécessaire, ce qui favorise la confiance. Par exemple, lors de l'utilisation de systèmes d'IA tels que les chatbots, les humains doivent être informés qu'ils interagissent avec une machine afin qu'ils puissent prendre une décision éclairée de continuer ou de prendre du recul. Les fournisseurs doivent également veiller à ce que le contenu généré par l'IA soit identifiable. En outre, les textes générés par l'IA publiés dans le but d'informer le public sur des questions d'intérêt public doivent être étiquetés comme étant générés artificiellement. Cela s'applique également aux contenus audio et vidéo constituant des contrefaçons profondes.

Les risques minimaux ou nuls sont ceux qui ne présentent aucun risque ou un risque négligeable, comme ceux utilisés à des fins de divertissement (comme les jeux vidéo compatibles avec l'IA ou personnelles (comme les filtres anti-spam)[39].

Le règlement établit une structure de gouvernance pour la mise en œuvre et l'application de ses règles, y compris la création d'un Bureau européen de l'IA[40] qui fournira des orientations et des conseils sur divers aspects de la législation sur l'IA, tels que des normes harmonisées, des codes de conduite et des méthodes d'évaluation des risques. Il facilitera également la coopération et la coordination entre les autorités nationales compétentes qui seront chargées de surveiller et de superviser le respect de la législation sur l'IA. Des sanctions et des voies de recours en cas de non-respect sont prévues, avec des amendes pouvant aller jusqu'à 6 % du chiffre d'affaires annuel mondial ou 30 millions d'euros (le montant le plus élevé) pour les infractions graves.

En outre, en janvier 2024, la Commission européenne a lancé un train de mesures sur l'innovation en matière d'IA afin d'aider les startups et les PME à développer une IA digne de confiance et conforme aux valeurs et aux règles de l'Union européenne. L'initiative « GenAI4EU » et la création du Bureau européenne de l'IA faisaient tous

[39] https://digital-strategy.ec.europa.eu/fr/policies/regulatory-framework-ai
[40] https://digital-strategy.ec.europa.eu/fr/policies/ai-office

deux partie de ce paquet de mesures. Ensemble, ils contribueront au développement de nouveaux cas d'utilisation et d'applications émergentes dans les 14 écosystèmes industriels européens, ainsi que dans le secteur public. Les domaines d'application comprennent la robotique, la santé, la biotechnologie, l'industrie, la mobilité, le climat et les mondes virtuels[41].

4.3. OCDE

L'OCDE (Organisation de coopération et de développement économiques) est un forum stratégique et un centre d'expertise unique en matière de données, analyses et bonnes pratiques dans le domaine des politiques publiques. Cette organisation internationale travaille à l'élaboration de politiques meilleures pour des vies meilleures en étroite collaboration avec les pouvoirs publics, les responsables politiques et les citoyens.

La recommandation de l'OCDE sur l'IA[42] fournit des définitions pratiques de termes pertinents, tels que les systèmes d'IA, le cycle de vie des systèmes d'IA, les connaissances en matière d'IA, les acteurs de l'IA, etc. Elle établit les principes d'une gestion responsable d'une IA digne de confiance, qui sont pertinents pour toutes les parties prenantes : croissance inclusive, développement durable et bien-être ; respect de l'état de droit, des droits humains et des valeurs démocratiques, y compris de l'équité et de la vie privée ; transparence et explicabilité ; robustesse, sûreté et sécurité ; responsabilité.

Elle recommande également des politiques nationales et une coopération internationale pour une IA digne de confiance dans les domaines suivants : investir dans la recherche et le développement en matière d'IA ; favoriser l'instauration d'un écosystème inclusif propice à l'IA ; façonner un cadre d'action et de gouvernance interopérable favorable à l'IA ; renforcer les capacités humaines et préparer la transformation du marché du travail ; favoriser la coopération internationale au service d'une IA digne de confiance.

L'Observatoire OCDE des politiques de l'IA[43] combine des ressources provenant de l'ensemble de l'OCDE et de ses partenaires de tous les groupes de parties prenantes. Il facilite le dialogue et fournit des analyses politiques multidisciplinaires et fondées sur des données probantes sur les domaines d'impact de l'IA. Il s'agit d'une source unique d'informations, d'analyses et de dialogues en temps réel, conçue pour façonner et partager les politiques d'IA à travers le monde[44].

[41] https://ec.europa.eu/commission/presscorner/detail/fr/ip_24_383
[42] https://legalinstruments.oecd.org/fr/instruments/OECD-LEGAL-0449
[43] https://www.oecd.org/fr/themes/intelligence-artificielle.html
[44] https://www.oecd.org/fr/themes/intelligence-artificielle.html

5. Recommandations

La pratique déjà établie des réseaux sociaux de pousser délibérément le contenu qui génère le plus de revenus (par défaut, il s'agit de contenu rempli de sensationnalisme, de désinformation, de propagande, de discours de haine et de mensonges), dont on connait les conséquences et dont on imagine ce qu'elles pourraient devenir si on y ajoute le potentiel de l'IA, ressemble à un cauchemar terrifiant, qui pourrait tout aussi bien s'avérer bientôt être la réalité si les réponses mentionnées au chapitre précédent ne sont pas ou pas assez mises en œuvre.

Le soit-disant « bouclier » sous forme de « lignes directrices de la communauté » des utilisateurs des plateformes et développeurs d'IA, alors qu'il s'agit de questions qui sont en fait au cœur des politiques publiques et doivent être débattues, adoptées et mises en œuvre par des institutions publiques souveraines, payées par les citoyens qui les élisent lors d'élections démocratiques, montre non seulement le pouvoir de ces acteurs, mais aussi leur compétitivité avec les souverainetés des gouvernements élus et des institutions (supra)internationales.

L'histoire de l'humanité nous enseigne que ce qui est décrit dans cette étude n'a pas été inventé avec les plateformes en ligne et les développeurs d'IA. Mais ce qui est nouveau aujourd'hui, c'est qu'on n'a jamais vu auparavant un tel niveau de prolifération de ces dangers et de menaces sur les fondements mêmes des sociétés démocratiques.

C'est là que l'aspect commercial des plateformes entre en jeu et que les choses se compliquent, mais devraient en principe être simples. Si l'on récolte autant de gains économiques, d'influence et de pouvoir, il faut en assumer la responsabilité. Plutôt que de s'appuyer sur des règles privées, des orientations politiques et des préférences personnelles, et en gardant à l'esprit le volume des enjeux à aborder, les géants du web doivent s'ouvrir, prendre leurs responsabilités et travailler sous la tutelle de règles démocratiquement adoptées au profit de l'intérêt général. Il faut beaucoup de temps pour comprendre ce que signifie travailler au bénéfice de l'intérêt public et pour comprendre quelle est la ligne de démarcation entre un contenu acceptable et un contenu préjudiciable, ce qu'est le droit à la liberté d'expression et quelle est sa dérogation légitime. Il est donc temps de commencer à apprendre. Comment protéger le droit à la liberté d'expression, comment coopérer, comment travailler ensemble à la revendication des principes et des valeurs démocratiques, alors qu'il reste encore un peu de temps pour cela ?

De plus, nos vies sont en quelque sorte régies en ce moment par la vitesse, l'impulsivité et la déconnexion, en grande partie à cause de la façon dont les gens passent leur temps en ligne. L'IA tire parti de ces dynamiques, nous assistons donc à ce genre d'impulsivité et à ce mouvement vers la vitesse et l'éloignement de la profondeur de l'information, ce qui fait qu'il est très facile pour un mauvais acteur

d'utiliser l'information générée par l'IA, que ce soit pour se faire passer pour quelqu'un ou pour faire autre chose, pour contourner la pensée critique des gens et leur glisser des informations qui les égarent. Ces problèmes brisent la confiance des gens, les rendent peu sûrs d'eux et donc émotifs et fatigués, ce qui est une condition parfaite pour exercer un contrôle sur les gens.

Les recommandations formulées dans ce domaine sont peut-être mieux illustrées par la conférence Asilomar 2017 sur l'IA bénéfique, organisée par le Future of Life Institute, qui s'est tenue en janvier 2017 en Californie. Plus de 100 leaders d'opinion et chercheurs en économie, en droit, en éthique et en philosophie se sont réunis lors de la conférence pour aborder et formuler les principes de l'IA bénéfique. Il en a résulté la création d'un ensemble de lignes directrices pour la recherche sur l'IA – les 23 principes d'Asilomar sur l'IA[45]. Ces principes sont les suivants :

Questions de recherche	1) Objectif de la recherche : l'objectif de la recherche sur l'IA devrait être de créer non pas une intelligence non dirigée, mais une intelligence bénéfique.
	2) Financement de la recherche : les investissements dans l'IA devraient s'accompagner d'un financement pour la recherche visant à garantir son utilisation bénéfique, y compris des questions épineuses en informatique, en économie, en droit, en éthique et en études sociales, telles que : Comment pouvons-nous rendre les futurs systèmes d'IA très robustes, afin qu'ils fassent ce que nous voulons sans dysfonctionner ou être piratés ? Comment pouvons-nous accroître notre prospérité grâce à l'automatisation tout en préservant les ressources et les objectifs des gens ? Comment pouvons-nous mettre à jour nos systèmes juridiques pour qu'ils soient plus justes et efficaces, pour suivre le rythme de l'IA et pour gérer les risques associés à l'IA ? Sur quel ensemble de valeurs l'IA doit-elle s'aligner et quel statut juridique et éthique doit-elle avoir ?
	3) Culture de la recherche : une culture de coopération, de confiance et de transparence doit être encouragée parmi les chercheurs et les développeurs d'IA.
	4) Lien science-politique : il devrait y avoir des échanges constructifs et sains entre les chercheurs en IA et les décideurs.
	5) Évitement de la course : les équipes qui développent des systèmes d'IA doivent coopérer

[45] https://futureoflife.org/open-letter/pause-giant-ai-experiments/

	activement pour éviter de réduire les angles sur les normes de sécurité.
Éthique et valeurs	6) Sécurité : les systèmes d'IA doivent être sûrs et sécurisés tout au long de leur durée de vie opérationnelle, et de manière vérifiable quand c'est applicable et faisable.
	7) Transparence de l'échec : si un système d'IA cause des dommages, il devrait être possible de déterminer pourquoi.
	8) Transparence judiciaire : toute implication d'un système autonome dans la prise de décision judiciaire doit fournir une explication satisfaisante vérifiable par une autorité humaine compétente.
	9) Responsabilité : les concepteurs et les constructeurs de systèmes d'IA avancés sont des parties prenantes dans les implications morales de leur utilisation, de leur mauvaise utilisation et de leurs actions, avec la responsabilité et la possibilité de façonner ces implications.
	10) Alignement des valeurs : les systèmes d'IA hautement autonomes doivent être conçus de manière à ce que leurs objectifs et leurs comportements puissent être assurés de s'aligner sur les valeurs humaines tout au long de leur fonctionnement.
	11) Valeurs humaines : les systèmes d'IA doivent être conçus et exploités de manière à être compatibles avec les idéaux de dignité humaine, de droits, de libertés et de diversité culturelle.
	12) Protection de la vie privée : les gens devraient avoir le droit d'accéder, de gérer et de contrôler les données qu'ils génèrent, étant donné le pouvoir des systèmes d'IA d'analyser et d'utiliser ces données.
	13) Liberté et vie privée : l'application de l'IA aux données personnelles ne doit pas restreindre de manière déraisonnable la liberté réelle ou perçue des personnes.
	14) Avantage partagé : les technologies d'IA doivent bénéficier au plus grand nombre de personnes possible.
	15) Prospérité partagée : la prospérité économique créée par l'IA devrait être largement partagée, au profit de l'ensemble de l'humanité.
	16) Contrôle humain : les humains devraient choisir comment et s'ils délèguent les décisions aux

	systèmes d'IA, afin d'atteindre les objectifs choisis par l'homme.
	17) Non-subversion : le pouvoir conféré par le contrôle de systèmes d'IA très avancés devrait respecter et améliorer, plutôt que subvertir, les processus sociaux et civiques dont dépend la santé de la société.
	18) Course aux armements de l'IA : une course aux armements dans les armes létales autonomes doit être évitée.
Problèmes à plus long terme	19) Mise en garde : en l'absence de consensus, nous devrions éviter les hypothèses fortes concernant les limites supérieures des capacités futures de l'IA.
	20) Importance : l'IA avancée pourrait représenter un changement profond dans l'histoire de la vie sur terre, et devrait être planifiée et gérée avec un soin et des ressources proportionnels.
	21) Risques : les risques posés par les systèmes d'IA, en particulier les risques catastrophiques ou existentiels, doivent faire l'objet d'efforts de planification et d'atténuation proportionnels à leur impact attendu.
	22) Auto-amélioration récursive : les systèmes d'IA conçus pour s'auto-améliorer ou s'auto-répliquer de manière récursive d'une manière qui pourrait conduire à une augmentation rapide de la qualité ou de la quantité doivent être soumis à des mesures de sécurité et de contrôle strictes.
	23) Bien commun : la superintelligence ne devrait être développée qu'au service d'idéaux éthiques largement partagés, et pour le bénéfice de toute l'humanité plutôt que d'un État ou d'une organisation.

Nous devons discuter de la manière d'établir des garde-fous et des normes éthiques pour le déploiement et l'utilisation de l'IA, en veillant à ce qu'elle soit utilisée pour enrichir nos vies plutôt que de diminuer l'essence de la connexion humaine.

Si l'IA présente des défis, elle offre également des opportunités sans précédent. C'est à nous d'utiliser ce potentiel et ces capacités de manière responsable. L'impact de l'intelligence artificielle sur les humains repose exclusivement entre les mains des humains eux-mêmes. Que les humains perdent ou conservent l'humanité ne dépendra pas de l'intelligence inhumaine, mais de l'intelligence humaine.