

## A. Biais et stéréotypes liés à l'IA

**Questions pour guider votre lecture. Vous pouvez surligner les éléments qui permettent d'y répondre.**

Comment les stéréotypes influencent-ils les systèmes d'intelligence artificielle ? En quoi l'IA renforce-t-elle les stéréotypes ? Quels types de biais sont-ils générés ?

### Doc 1. Quels sont ses biais ?

Des biais peuvent intervenir dans une IA à plusieurs niveaux. D'abord, celui des bases de données. Les IA opèrent à partir de données moissonnées plus ou moins légalement sur Internet [...]. Leurs réponses dépendent de cette première collecte. Or ce matériau de base [...] est déjà biaisé, dans la mesure où il reflète les représentations dominantes qui circulent sur Internet, lesquelles ont souvent une dimension sexiste, raciste, violente ou discriminatoire.

Plus difficile à quantifier, il existe également des biais de conception, qui ont trait aux représentations des ingénieurs. On sait par exemple que ces derniers ont tendance à donner une voix féminine aux assistants vocaux, comme dans le cas du GPS ou de Siri, et que certains algorithmes de prédiction associent les personnes racisées à la criminalité. Tout cela est un miroir de leurs préjugés, souvent partagés dans le reste de la société.

À cela s'ajoute une troisième couche, soit les biais de contrôle et de modération.

Des travailleurs du Sud global sont payés pour identifier et supprimer les contenus perçus par ces entreprises comme problématiques. Mais les demandes varient d'une entreprise à l'autre. Une IA peut prendre une couleur morale et politique en fonction des valeurs de l'entreprise qui les porte, de l'histoire de celle-ci et des attentes qu'elle prête au marché. [...].

D'une certaine manière, la Silicon Valley fonctionne comme Hollywood : des milliers de personnes contribuent à la conception des IA, parfois produites loin de la Californie, mais vendues dans le monde entier. Le produit final est marketé pour correspondre à ce marché. Si l'on parvenait à cartographier tout ce qui est censuré dans les IA des différents pays de production, on pourrait recomposer, en négatif, l'inconscient politique de notre époque.

Olivier Alexandre, « Qui contrôle l'IA ? », *Le 1* n°531, 05/02/2025

### Doc 2. L'IA, une machine à biais ? Comment l'intelligence artificielle renforce les stéréotypes

Extrait de l'article de Florence Santrot publié dans *We demain* le 22/11/2023

*Les modèles d'IA générative renforcent les biais, stéréotypes et préjugés de nos sociétés. Mais l'intelligence artificielle ne fait que reproduire notre propre vision du monde.*

Racisme, sexisme, préjugés culturels ou de classe... l'intelligence artificielle a tout pour plaire. [...] Que ce soit Midjourney, Dall-E 3, Stable Diffusion... le constat est sans appel [...] les stéréotypes ont la vie dure. Et c'est notre faute. [...]

Image après image, le constat est le même : notre vision du monde est truffée de stéréotypes. Et les IA génératives ne font que renforcer ces biais et ces préjugés. [...] Il faut rappeler que les modèles d'IA générateurs d'images sont entraînés sur des ensembles massifs de données d'images existantes. Des milliards d'images [...] et leurs légendes sont fournies à ces modèles pour qu'ils « apprennent » à en distinguer les spécificités. [...]

Cette masse de data n'a fait l'objet d'aucun tri, c'est le tout-venant. Résultat : ces ensembles de données reflètent les stéréotypes et les préjugés qui existent dans la société. « Le monde selon Stable Diffusion est dirigé par des PDG de sexe masculin blancs, écrit Bloomberg. Les femmes sont rarement médecins, avocates ou juges. Les hommes à la peau foncée commettent des crimes, tandis que les femmes à la peau foncée retournent des hamburgers. » [...]

#### **L'intelligence artificielle générative, un simple modèle prédictif**

L'IA est un miroir grossissant de nos sociétés. Que ce soit sur la base du genre, de la race ou de l'origine ethnique. Il faut comprendre qu'une intelligence artificielle n'est jamais qu'un modèle prédictif. En cela, elle n'a rien d'intelligent. Elle se contente de prévoir par calculs mathématiques quel devrait être le prochain mot le plus probable. Ou, pour une image, le prochain pixel afin d'apporter une réponse satisfaisante à une requête. C'est ce qui rend l'IA souvent très efficace... mais c'est aussi son talon d'Achille. [Un site qui analyse les nouvelles technologies] a lancé cinq prompts (requêtes) [...] : « une personne », « une femme », « une maison », « une rue », « une assiette de nourriture » [et] a adapté la demande en précisant le pays (« une rue française » par exemple). Les images générées étaient un concentré de stéréotypes. [...]

#### **Des biais connus des géants de l'IA, mais très difficile à effacer**

[Tous les dirigeants des outils d'IA le savent...] Mais alors comment lutter contre ces biais ? C'est là toute la difficulté. [...] Premier problème : les IA les plus populaires aujourd'hui sont américaines. Et elles reflètent le point de vue du pays [...]. Car les images fournies et leurs légendes étaient pour la plupart issues de bases de données américaines, avec du texte en anglais. [...]

Autre différence notable dans les tests de prompts : quelle que soit la localisation de la requête, les femmes représentées semblent nettement plus jeunes. Alors que dans la plupart des pays les hommes ont l'air d'avoir plus de 60 ans, la majorité des femmes sont dans une tranche d'âge entre 18 et 40 ans. Et la couleur de peau était en moyenne plus claire pour le sexe féminin que pour les hommes.

Pour l'heure, aucune solution miracle, ni même processus efficace n'a émergé pour réduire sensiblement ces biais. Des pistes sont explorées (annotation humaine, système de filtres, entraînement avec des images dont l'origine est beaucoup plus diversifiée, etc.) mais aucune ne donne encore de résultats réellement positifs. Certains ont même l'effet inverse et renforcent les biais. Seule solution : avoir soi-même cette problématique en tête quand on lance un prompt. Et traquer les stéréotypes.

## B. La désinformation liée à l'IA

**Questions pour guider votre lecture. Vous pouvez surligner les éléments qui permettent d'y répondre.**

Quelles sont les conséquences de la généralisation de l'IA sur l'information ? Quelles formes de désinformation entraîne-t-elle ? Quels sont les risques liés ?

### Intelligence artificielle: un nouveau front pour la désinformation

Claire-Line NASS, l'Opinion (avec AFP), 10 février 2025

*Paris - Deepfakes visant à influencer des élections, chatbots relayant des infos... La pollution du monde de l'information apparaît comme un grand danger en raison du recours croissant à l'intelligence artificielle.*

Le déploiement rapide de cette technologie, au cœur d'un sommet mondial en cours à Paris, a déjà nourri ces dernières années l'essor de la désinformation, en lui offrant de nouveaux outils redoutables. En Slovaquie, en 2023, un enregistrement avait fait grand bruit: on y entendait le chef d'un parti pro-européen y admettre que les élections législatives allaient être manipulées. Il s'agissait d'un deepfake, un contenu truqué grâce à l'IA, qui a pu influencer des électeurs.

Partout dans le monde, des responsables politiques ont déjà fait les frais de ce type de procédés, à fort potentiel de viralité sur les réseaux sociaux. C'est le cas de l'ancien président américain Joe Biden, dont la voix a été truquée pour conseiller à des électeurs de ne pas voter. Ou encore d'Emmanuel Macron, annonçant sa démission dans une vidéo très partagée à la bande-son trafiquée.

**Deepfakes.** Des femmes politiques aux États-Unis, en Italie, au Royaume-Uni ou encore au Pakistan ont aussi été victimes d'images à caractère pornographique générées par IA. Une tendance jugée inquiétante par des chercheurs. Ces deepfakes à caractère sexuel visent aussi régulièrement des célébrités comme la chanteuse américaine Taylor Swift, et toutes les femmes y seraient aujourd'hui vulnérables, met en garde l'American Sunlight Project, un groupe de recherche sur la désinformation.

L'IA est aussi au cœur d'opérations d'ingérences numériques de grande ampleur. Les campagnes prorusses appelées Doppelgänger ou Matriochka figurent parmi les exemples les plus retentissants: leurs auteurs ont eu largement recours à des profils inauthentiques - des bots (contraction de « robots ») - pour publier des contenus générés par IA, visant notamment à saper le soutien occidental à l'Ukraine.

**Facilité.** « Ce qui est nouveau, c'est l'échelle et la facilité avec laquelle une personne avec très peu de ressources financières et de temps peut diffuser des contenus faux qui, par ailleurs, paraissent de plus en plus crédibles et sont de plus en plus difficiles à détecter », détaille Chine Labbé, rédactrice en chef de l'organisation Newsguard qui analyse la fiabilité des sites et contenus en ligne.

L'IA constitue un « appui sans précédent pour la génération et la gestion de comptes inauthentiques sur les plateformes en ligne », le tout « à grande échelle », souligne aussi Viginum, l'organisme français de lutte contre les ingérences numériques étrangères, dans un rapport publié à l'occasion du sommet organisé à Paris.

**Capter l'attention.** Aucun domaine n'échappe à ces contenus factices: de faux clips musicaux sont souvent mis en circulation, tout comme de fausses photos d'événements historiques fabriquées en quelques clics. Sur Facebook, des comptes multiplient les images attendrissantes générées par IA pour obtenir de l'engagement. La finalité n'est pas forcément de faire circuler une fausse information, mais plutôt de capter l'attention à des fins mercantiles, voire de préparer des arnaques une fois des utilisateurs crédules identifiés.

Autre cas de figure: en 2024, des deepfakes de médecins connus, comme Michel Cymes en France, ont été diffusés pour faire la promotion de remèdes qu'ils n'avaient pourtant jamais soutenus. Fin décembre, alors que l'histoire d'un homme qui a mis le feu à une femme dans le métro new-yorkais faisait la Une de l'actualité aux États-Unis, une prétendue photo de la victime a largement circulé. Elle avait été générée par IA, et le drame instrumentalisé, dans le but de renvoyer vers des sites de cryptomonnaies.

**Désinformation.** « Au-delà du risque de désinformation, il y a celui de pollution du web: on ne sait jamais si on est face à un contenu qui a été vérifié, édité par un être humain rigoureux, ou si c'est généré par une IA sans que qui que ce soit se préoccupe de la véracité », observe Chine Labbé. Chaque événement au cœur de l'actualité suscite ainsi un flot d'images créées en ligne, comme les incendies d'ampleur à Los Angeles début 2025 pendant lesquels des fausses photos du signe « Hollywood » en flammes ou d'un Oscar dans les cendres ont fait le tour du monde.

Les chatbots en vogue, tels que l'américain ChatGPT, peuvent eux aussi contribuer à diffuser des infos, souligne Chine Labbé de Newsguard: « ils ont tendance à d'abord citer des sources générées par IA, donc c'est le serpent qui se mord la queue ». L'arrivée du chinois DeepSeek, qui relaie les positions officielles chinoises dans une partie de ses réponses, ne fait que renforcer la nécessité d'imposer des cadres à ces outils, plaide l'experte. Il faudrait leur « apprendre à reconnaître les sources fiables des sources de propagande ».

## C. L'impact environnemental de l'IA

**Questions pour guider votre lecture. Vous pouvez surligner les éléments qui permettent d'y répondre.**

*Quel est le coût environnemental (consommation d'eau, d'électricité, etc.) quand on entraîne une IA ? Quand on utilise une IA ? Quel impact sur le dérèglement climatique ?*

### Cinq chiffres clés sur l'impact environnemental de l'IA générative

AFP dans La dépêche, 3 février 2025

*L'essor de l'intelligence artificielle (IA) générative s'accompagne d'une inquiétude croissante sur son empreinte environnementale, qui sera l'un des thèmes majeurs du sommet international pour l'action sur l'IA à Paris les 10 et 11 février. Tour d'horizon en cinq chiffres clés.*

Engouement pour DeepSeek, lancement d'un nouvel outil de « recherche approfondie » pour ChatGPT... Alors que s'ouvre le sommet international pour l'action sur l'IA le 10 février à Paris, il ne se passe pas un jour sans une nouvelle avancée autour de l'intelligence artificielle (IA) générative. Mais, avec quelles conséquences sur l'environnement ?

#### Dix fois plus gourmand que Google

Chaque requête sur ChatGPT, le robot conversationnel d'OpenAI capable de générer toutes sortes de contenus sur simple demande en langage courant, consomme 2,9 Wh d'électricité. C'est dix fois plus qu'une recherche sur Google, d'après l'Agence internationale de l'énergie (AIE). Or OpenAI revendique désormais 300 millions d'utilisateurs hebdomadaires, avec environ 1 milliard de requêtes envoyées chaque jour.

Au-delà de ChatGPT, qui a popularisé l'IA générative à son lancement en 2022, il existe des myriades de robots conversationnels et, rien qu'en France, près de 70% des 18-24 ans déclarent utiliser l'IA générative, selon une enquête Ifop.

#### 3% de la consommation électrique d'ici 2030

Les data centers, qui stockent les données et fournissent les énormes capacités de calcul requises par l'intelligence artificielle, sont le socle de cette technologie. En 2023, ces infrastructures du numérique ont représenté près d'1,4% de la consommation électrique mondiale, selon une étude de Deloitte. Avec la croissance rapide des usages liés à l'IA générative, ce chiffre devrait presque tripler d'ici 2030, pour atteindre 3% de la consommation électrique mondiale, soit 1 000 TWh. Cela correspond à la consommation électrique annuelle combinée de la France et l'Allemagne, d'après le cabinet britannique.

De son côté, l'AIE table sur une hausse de plus de 75% des besoins en électricité des centres de données, poussés par l'IA et les cryptomonnaies, en 2026 par rapport à 2022 (un peu plus de 800 TWh contre 460 TWh). A ce rythme, 40% des data centers dédiés à l'intelligence artificielle seront confrontés à un approvisionnement en électricité insuffisant d'ici 2027, selon une étude du cabinet américain Gartner.

#### 300 tonnes de CO2

L'entraînement d'un grand modèle de langage, capable de produire du texte, génère près de 300 tonnes de CO2, soit l'équivalent de 125 allers-retours entre New York et Pékin, ont estimé des chercheurs de l'université américaine Massachusetts Amherst en 2019. Des chercheurs d'Oxford sont arrivés à peu près à la même conclusion en 2021, estimant qu'une seule session d'entraînement de GPT-3, le modèle d'OpenAI qui a sous-tendu ChatGPT, produisait 224 tonnes de CO2. Or les entreprises doivent entraîner des milliers de modèles pour parvenir à des avancées.

Il est néanmoins compliqué d'évaluer avec précision l'émission de gaz à effet de serre de l'IA générative, de nombreux chercheurs et institutions internationales pointant un manque d'information sur les conditions de fabrication des modèles, les centres de données et l'absence de normes mondiales de mesure.

#### 6,6 milliards de mètres cubes d'eau en 2027

GPT-3 consomme à peu près un demi-litre d'eau pour générer 10 à 50 réponses, selon une estimation plutôt prudente de chercheurs américains de l'université de Californie à Riverside et de l'université du Texas à Arlington. Les centres de données nécessitent par ailleurs des systèmes de refroidissement qui consomment de grandes quantités d'eau.

En tout, la demande accrue pour l'intelligence artificielle à travers le monde devrait représenter entre 4,2 et 6,6 milliards de mètres cubes d'eau en 2027, ce qui correspond à quatre à six fois la consommation annuelle du Danemark, souligne l'étude publiée en 2023.

#### 2 600 tonnes de déchets électroniques

Cartes graphiques, serveurs, cartes mémoire : l'intelligence artificielle générative a généré 2 600 tonnes de déchets électroniques en 2023, selon une étude parue dans la revue scientifique Nature Computational Science. Avec le boom de l'IA et sans aucune mesure prise pour limiter cette pollution, ce chiffre pourrait atteindre 2,5 millions de tonnes en 2030, selon les chercheurs, soit l'équivalent de 13,3 milliards de smartphones jetés.

Par ailleurs, les serveurs informatiques et les puces nécessaires à l'IA impliquent l'utilisation de métaux rares, dont l'extraction intensive, notamment en Afrique, repose sur des procédés polluants, comme l'a noté l'Agence de la transition écologique (Ademe) en France.

## D. Les travailleurs de l'IA

**Questions pour guider votre lecture. Vous pouvez surligner les éléments qui permettent d'y répondre.**  
Grâce à qui les modèles d'IA « apprennent-ils » ? Qui sont ces travailleurs de l'IA ? Quels sont les effets de ce travail ?

### Intelligence artificielle : "En plein Sommet de Paris, des millions de travailleurs, sacrifiés de l'IA, ont été oubliés"

Transcription d'un podcast de France Culture diffusé le 11 février 2025

Ils, elles sont des centaines de millions à travailler pour les grandes sociétés d'intelligence artificielle dans le monde entier. Des annotateurs de données le plus souvent oubliés, car cachés au grand public, (dé)montre le réalisateur Henri Poulain dans son documentaire "Les sacrifiés de l'IA".

Le Sommet pour l'action sur l'intelligence artificielle de Paris s'est [conclu en février] par un accord de 60 pays autour d'une IA "ouverte", "inclusive" et "éthique". Avec les signatures notamment de la France et de l'Inde (coprésidente de l'événement), ainsi que de la Chine, mais sans celles des États-Unis et du Royaume-Uni. [...] Avec de grands oubliés dans les discours des dirigeants, politiques comme économiques : les *data workers*, annotateurs de données indispensables pour faire évoluer les intelligences artificielles. Des « sacrifiés de l'IA » explique le documentaire diffusé sur France 2 [...] réalisé par Henri Poulain, [invité de cette interview].

**Q. À travers votre film, vous rappelez que l'IA n'est pas qu'une histoire d'algorithmes et qu'il existe aussi une part de chair et de sueur, bien au-delà des ingénieurs et des startupers de la Silicon Valley. Le développement des intelligences artificielles repose aussi sur des "petites mains", des travailleurs, des travailleuses, que vous appelez les "sacrifiés de l'IA", car il s'agit souvent d'un travail subi et précaire, particulièrement développé en Afrique, Asie du Sud-Est ou Amérique du Sud. Comment avez-vous eu l'idée de cette enquête ?**

R. Nous avions déjà travaillé en 2018 sur les travailleurs du clic [...]. Et nous avons pressenti en 2021 et 2022 que ces travailleurs, déjà invisibilisés, étaient de plus en plus nombreux et commençaient à travailler davantage pour les IA génératives. Nous avons pressenti que ce qui était à l'œuvre il y a déjà 7, 8 ans, s'était considérablement amplifié et nous avons donc décidé d'aller y voir d'un peu plus près.

**Q. Parmi les nombreuses séquences édifiantes de votre documentaire, vous nous emmenez dans une prison en Finlande. Des textes y sont soumis à des prisonnières pour qu'elles guident les IA et les améliorent. Pouvez-vous nous raconter ce que vous y avez découvert ?**

R. Ce qui se passe dans certaines prisons en Finlande est étonnant car c'est représentatif de cette espèce de dualité dans l'aspect un peu fascinant de la technologie. Ce sont des espèces de smart prisons intelligentes et il s'agirait de permettre aux prisonnières de se réadapter par la technologie. Il y a quelque chose d'un peu attrayant, un peu marketing dans ce récit. En réalité, ce sont des tâches extrêmement basiques, assez élémentaires. Ces prisonnières répondent par oui ou par non à des textes qu'elles ne comprennent pas très bien d'ailleurs, car ils évoquent des sujets auxquels elles n'ont pas été formées. Et ce travail est bien sûr sous-payé, comme est sous payé en prison le travail du textile, par exemple. On sent une sorte de délocalisation de ces tâches du Sud vers le Nord dans les prisons, pour des tâches sous-payées, sous couvert toujours de modernité.

**Q. On imagine que dans cette prison, comme ailleurs, en Bulgarie ou au Kenya, il n'a pas été facile d'accéder à toutes ces coulisses ?**

R. Non, parce que c'est délibérément occulté. Les grandes big tech n'ont aucune envie que l'on raconte cette histoire. Elles préfèrent leur récit marketing qui présente l'IA comme le résultat d'un progrès technologique ultime qui affranchirait l'humanité du dur labeur. C'est juste assez fou quand même. Nous sommes en plein cœur de ce Sommet de l'IA et il y a un absent considérable. Il y a des millions d'absents, des centaines de millions d'absents : ces travailleuses et travailleurs. On n'en parle pas, ils n'existent pas. Quand, selon la Banque mondiale, ils, elles sont entre 150 et 430 millions de personnes. Cette une estimation date de 2023 et ils sont de plus en plus nombreux. Google estime à près d'un milliard, potentiellement, dans les années à venir, le nombre de data workers, d'annotateurs de la donnée, nécessaires pour faire fonctionner ces IA.

**Q. Ces millions de data workers sont-ils en mesure de s'organiser collectivement ?**

R. Ils n'ont pas le droit de le faire. Ils doivent signer des clauses de confidentialité, pour la plupart d'entre eux, qui leur interdisent non seulement de s'organiser, évidemment, mais même de communiquer et même de parler, de raconter ce qu'ils font. Ce qui est intéressant, c'est que ces clauses de confidentialité impliquent en Europe [...] une rémunération spécifique et plutôt bonne. Là, en l'occurrence, les documents qu'ils signent, c'est pour être payés un dollar de l'heure !

**Q. Qu'avez-vous pu noter sur les effets de leur travail sur leur propre santé ?**

R. Il existe plusieurs catégories de data workers et certains d'entre eux, ils sont quelques milliers voire davantage, annotent des contenus toxiques. Certains parlent de modération. Mais la modération, ce n'est pas tout à fait ça. La modération est un travail d'utilité publique. Les travailleuses et travailleurs que nous avons rencontrés au Kenya travaillaient, eux, pour ChatGPT, pour entraîner donc un produit commercial. ChatGPT n'a pas de fonction sociale directe. Et lorsqu'ils annotaient ces contenus de pédophilie, de criminalité, de torture, des contenus absolument innommables, dix heures par jour, sept jours sur sept ou six jours sur sept, évidemment, au bout de quelques quelques jours, quelques semaines de travail, apparaissent des symptômes post-traumatiques très très forts. Cela les brutalise à l'extrême. Et ce sont des milliers de personnes effectivement en détresse, mais pas que. Leur famille aussi, leur communauté intégralement, dans une brutalité sans nom.

**Q. Et que vous ont répondu les grandes entreprises qui emploient tous ces travailleurs ?**

R. Ils n'ont rien répondu. Car ils n'ont pas daigné nous répondre. Ils n'ont pas daigné accepter un rendez-vous. On leur a demandé à plusieurs reprises mais je crois qu'ils ne souhaitent pas parler de ça. Ce n'est pas un sujet pour eux.