Les dangers d'une IA invasive

Les avis et études sur les usages, les bienfaits, les méfaits et l'avenir de l'intelligence artificielle (IA) font l'actualité des journaux et de nombreuses publications. En effet, telle une hydre, l'IA gagne tous les champs de notre société. Une expansion largement problématique.

Maryse ARTIGUELONG, coresponsable du groupe de travail LDH « Libertés et technologies de l'information et de la communication »

Si le grand public s'est familiarisé avec l'IA, c'est surtout grâce à la mise à disposition gratuite, fin 2022, de ChatGPT [1]. IA générative, elle est, comme d'autres applications similaires (Gemini, Midjourney, Dall-E, Claude…), capable de générer de façon autonome des contenus semblables à ce que pourrait produire un humain après de longues heures de travail (textes, images, sons, vidéos…). C'est un modèle utilisant un langage naturel basé sur le « *deep learning* » (apprentissage profond) et qui a été entraîné à « étudier » des millions de documents en ligne [2] afin de pouvoir générer des réponses à des requêtes (ou « prompts ») aussi simples que sur un moteur de recherche.

A sa mise à disposition, ChatGPT a atteint un million d'utilisateurs en seulement cinq jours. En mai 2023, l'application a été rendue disponible sur ordiphones, ce qui lui a permis de dépasser les trois-milliards de visites en septembre 2024 (avec, en décembre, trois-cent-millions d'utilisateurs par semaine). Un véritable engouement. Cette politique commerciale d'OpenAI a entraîné une forte adhésion à la version payante (plus performante), qui devrait générer des revenus pour l'année 2024 de 3,4 milliards de dollars.

Les systèmes d'IA (SIA) sont des systèmes informatiques qui ont été perfectionnés à partir des années 1950. Le règlement sur l'IA de l'UE les définit comme « […] *un système basé sur une machine, qui est conçu pour fonctionner avec différents niveaux d'autonomie, qui peut faire preuve d'adaptabilité après son déploiement, et qui, pour des objectifs explicites ou implicites, déduit, à partir des données qu'il reçoit, comment générer des résultats tels que des prédictions, du contenu, des recommandations ou des décisions qui peuvent influencer des environnements physiques ou virtuels* ».

Leur expansion est due à la puissance accrue des processeurs, aux progrès des algorithmes qui améliorent sans cesse leurs performances, et, surtout, à l'immense quantité de données collectées, notamment nos données personnelles.

La LDH elle-même a publié des textes sur ce sujet : dans le domaine de la surveillance (utilisation de la vidéosurveillance algorithmique-VSA lors des Jeux olympiques et paralympiques-JOP) [3], de la recherche des fraudeurs aux prestations sociales (algorithme de la Cnaf) [4] ou en matière de discriminations inhérentes à ces SIA.

JOP et vidéosurveillance algorithmique

A la veille des Jeux olympiques et paralympiques (JOP), nous alertions [5] sur le « technosolutionnisme » choisi par le gouvernement, lequel a fait voter une loi autorisant le recours à l'IA pour « doper aux algorithmes » la vidéosurveillance. L'objectif ? Générer des alertes en cas de détection d'« événements prédéterminés », potentiellement révélateurs de « risques » d'atteintes à la sécurité (notons que la Cnil a déploré l'absence d'information claire des voyageurs surveillés dans les quelque cinquante stations de métro ou gares ayant été dotées de VSA…). Autorisée par la loi jusqu'en mars 2025, cette expérimentation devait faire l'objet d'une évaluation fin 2024 par le comité de suivi. Mais alors qu'aucune preuve de son efficacité n'a été fournie, les autorités ont annoncé sa prolongation, « *à la* (prétendue) *demande des*

« Dans le domaine de la santé l'IA est très développée. Or les données utilisées sont porteuses de biais liés aux profils de risques utilisés, au manque de diversité de genre, d'âge et d'origine ethnique, et à des préjugés cliniques… engendrant un risque d'erreurs médicales. »

(1) Chat Generative Pretrained Transformer (traduction par une IA : « Tchat transformateur génératif pré-entraîné »), créé par OpenAI.

(2) En 2020 la version 3 de ChatGPT avait « appris » un corpus représentant deux-mille-quatre-cents-ans ans de lecture continue pour un être humain…

(3) Voir www.ldh-france.org/les-atteintes-aux-droits-et-libertes-pendant-la-periode-des-jeux-olympiques-de-paris-2024.

(4) Caisse nationale d'allocations familiales.

(5) « Une médaille d'or de la surveillance, pour la France ? », in *D&L* n° 205, avril 2024 (www.ldh-france.org/wp-content/uploads/2024/04/DL205-Actualite-4.-Medaille-dor-de-la-surveillance-pour-la-France.pdf).

citoyens». L'Assemblée nationale a certes décidé d'une «Mission flash sur le bilan sécurité» des JOP, mais elle n'a toujours pas rendu son rapport. En attendant il semble que les alertes sur les comportements suspects ne sont pas nombreuses et il faut certainement se réjouir de l'absence d'incidents majeurs, sans doute due à la très forte présence policière… Ce qui conforte nos combats pour privilégier une présence humaine (et bienveillante!) en lieu et place de technologies de plus en plus intrusives[6].

Contrôle renforcé des personnes démunies

Si l'IA devrait permettre de mieux appréhender les réalités économiques et sociales, en réalité elle est beaucoup utilisée pour surveiller les personnes démunies…
Le 15 octobre 2024 la LDH a déposé un recours au Conseil d'Etat[7] contre la Cnaf, qui utilise depuis plusieurs années un algorithme de ciblage attribuant à chaque allocataire un score de suspicion, pour sélectionner celles et ceux qui vont faire l'objet d'un contrôle. Ce score est aggravé par la déclaration de faibles revenus, la perception de l'AAH ou du RSA[8], ou encore une situation de chômage. Ainsi, plus les personnes sont en difficulté, plus elles sont sur-contrôlées par rapport au reste de la population. Le recours porte sur la surveillance à grande échelle des personnes déjà fragilisées et sur les discriminations dues à cet algorithme qui assimile précarité et soupçon de fraude, participant ainsi à une politique de stigmatisation et de maltraitance institutionnelle des plus défavorisés.
D'autres administrations utilisent des outils similaires pour chasser les fraudeurs. Ainsi France Travail attribue aux demandeurs d'emploi un score de suspicion pour détecter les plus susceptibles de fraude. Quant à la la Cnam[9], elle attribue un score basé sur le sexe, l'âge, la composition familiale (les femmes seules avec enfants sont ciblées) pour détecter automatiquement les foyers susceptibles de fraude à la complémentaire santé solidaire ou à la CSG.
Le ministère des Finances s'emploie aussi à détecter les fraudes fiscales par de nombreux projets. Ceux-ci sont mis en question par la Cour de comptes ou par des fonctionnaires qui déplorent la perte de sens de leurs missions et la délégation de conception des SIA à des sociétés privées comme Google. Quant au ministère de l'Intérieur, il utilise entre autres un système de reconnaissance faciale pour l'identification des personnes fichées dans le TAJ[10].

«L'UE s'est dotée d'un règlement sur l'IA, destiné à réguler ce marché. Basé sur la gestion des risques des SIA et sur l'autorégulation, il semble en réalité plus destiné à soutenir l'innovation, faire accepter le développement de l'IA qu'à prévenir les atteintes aux droits de l'Homme dénoncées par les citoyens…»

Notre quotidien envahi par l'IA

Les innovations scientifiques et techniques de l'IA ont des répercussions profondes, économiques, sociales et culturelles sur nos vies.
Pour beaucoup les IA sont juste des applications destinées à nous faciliter le quotidien: assistants intelligents (Siri, Alexa…) qui obéissent à notre voix, traducteurs automatiques qui rendent lisibles des langues inconnues, technologies qui permettent de nous transporter dans des véhicules roulant sans conducteur. Des applications d'IA peuvent même nous «fabriquer» des amis ou amies virtuels «parfaits».
Porteuse de promesses de gains de productivité, l'IA transforme de nombreux secteurs en entreprise: marketing, finance, logistique, production… Par exemple elle permet de trier les CV en sélectionnant les profils semblant les plus adaptés aux postes et elle utilise (illégalement), lors des entretiens, la reconnaissance des émotions pour en éliminer certains.
Dans des entrepôts comme ceux d'Amazon, l'automatisation est accélérée par l'IA et augmente la concurrence entre robots et salariés (mais les tâches qui nécessitent raisonnement et compréhension du contexte restent l'apanage des employés…).
Les banques et les assureurs utilisent aussi l'IA pour se livrer au «scoring» de leurs clients, par l'analyse de leurs comportements (données collectées sur le web) ou de leurs particularités, pour repérer et exclure ceux «à risques», contribuant ainsi à faire perdurer des inégalités liées au genre, à l'origine ethnique, à la maladie ou au handicap, en matière de crédits ou de services.

La santé et les médias, deux secteurs clés

Dans le domaine de la santé l'IA est particulièrement développée. D'une part les masses de données sont très variées: analyses médicales (biologiques, images radios, scanners, données des dispositifs médicaux numériques), mais aussi données administratives (remboursements,

Des inégalités en tous genres amplifiées

En mars 2019 *Droits & Libertés* ouvrait ses colonnes au Laboratoire de l'égalité[1], sur le sujet «mixité et intelligence artificielle». Le constat était navrant, et il le reste: si l'on peut déplorer la sous-représentation des femmes dans les filières informatiques ou d'ingénieurs, et donc dans les entreprises de ces secteurs (avec seulement 25% de salariées dans les métiers du numérique), les discriminations et les inégalités de genre ne sont pas le seul problème. En effet les systèmes d'IA développés sont sujets à des biais intégrés, reproduisant différentes formes de discrimination à raison de l'identité de genre, de l'orientation sexuelle, de la couleur de peau, de la classe sociale… amplifiant ainsi les inégalités existantes. L'exemple du système de reconnaissance faciale n'identifiant pas une femme à la peau foncée, car développé par un homme blanc et testé sur ses pairs, est emblématique…

(1) Voir www.ldh-france.org/wp-content/uploads/2019/07/HL185-Dossier-3.-Mixit%C3%A9-et-intelligence-artificielle.pdf.

M. A.

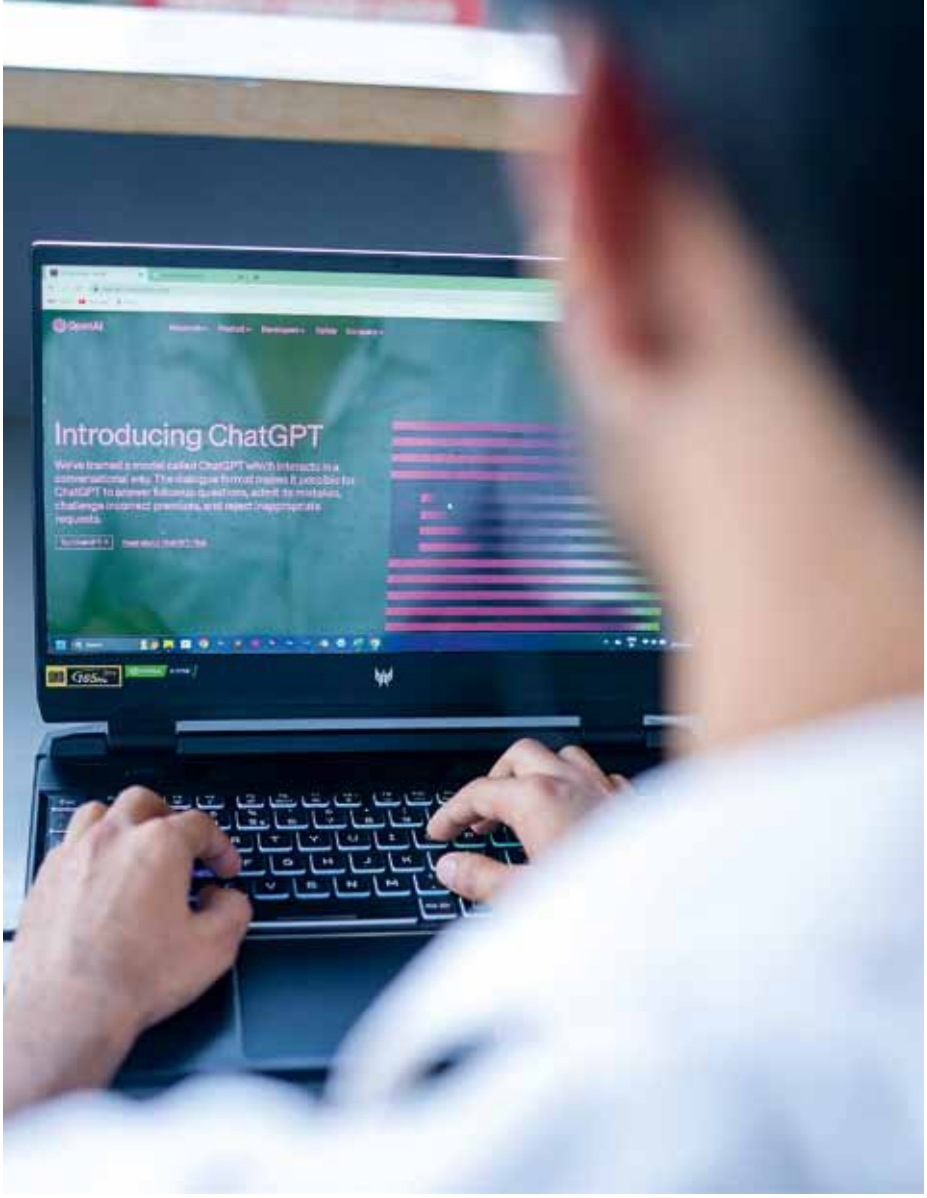
comptes rendus…). D'autre part les acteurs impliqués sont nombreux : médecins, hôpitaux, laboratoires, centres radiologiques, caisses de sécurité sociale… L'IA est donc largement utilisée depuis plusieurs années pour la recherche, le dépistage, l'aide au diagnostic et à la décision. Problème : les données utilisées sont porteuses de biais liés aux profils de risques utilisés, au manque de diversité de genre, d'âge et d'origine ethnique, et à des préjugés cliniques. Ces biais font que l'IA peut être source soit d'erreurs médicales, par l'« effet d'autorité de la machine », soit de soins totalement déshumanisés.

Mais c'est certainement dans le secteur des médias que l'utilisation de l'IA est la plus « défavorablement » connue. Les algorithmes programmés pour répandre viralement des idées racistes, sexistes, homophobes, des « fake news » et « deepfakes » (11) sont accusés d'être des amplificateurs de haine en ligne, de cyberviolence, et aussi d'influer sur les résultats de nombreuses élections. Ils sont en cela porteurs de risques pour la démocratie. Par ailleurs les IA génératives se livrent au pillage du travail des médias classiques, des auteurs, des artistes, et posent de sérieuses questions quant au droits d'auteur ou au remplacement des journalistes par ces outils.

Une régulation du secteur qui questionne

Tous ces SIA qui sont soit « offerts » soit imposés, sans que les citoyens soient consultés, interrogent sur la place que l'IA prend dans nos sociétés. Pour les « Big Techs » pourvoyeuses d'IA, seuls comptent les progrès techniques continus et les profits. Les questions du coût de ces approches en termes de droits de l'Homme ou d'impact sur une planète déjà mise à rude épreuve leur sont indifférentes (ou bien ces entreprises prétendent que l'IA peut aider à résoudre les problèmes qu'elle génère). D'où une hausse de la consommation énergétique des entrepôts de données, l'extraction « débridée » des minerais nécessaires à la fabrication des appareils électroniques, ainsi qu'une forme d'esclavage moderne pour les « travailleurs du clic » (12).

Soucieuses de présenter un visage vertueux de l'IA, les entreprises ont élaboré des chartes éthiques mais elles sont non contraignantes, et l'UE s'est dotée d'un règlement sur l'IA (AI Act), destiné à réguler ce marché. Loin de répondre aux promesses faites aux OSC (13), en particulier d'interdiction de la reconnaissance faciale, ce texte, qui sera opérationnel en août 2025, est basé sur la gestion des risques (faible, limité, élevé, inacceptable) des SIA et sur l'autorégulation. Les créateurs d'IA devront évaluer eux-mêmes les risques portés par leurs produits. S'il est bien prévu que les citoyens devront être informés lorsqu'ils se trouveront soumis à des SIA, la reconnaissance faciale sera autorisée dans le cadre de certaines opérations de police (recherche de victimes, de terroristes…) et le contrôle des migrations par l'IA ne sera pas concerné par ces classifications. En réalité ce texte semble plus destiné à soutenir l'innovation, faire accepter le développement de l'IA qu'à prévenir les atteintes aux droits de l'Homme dénoncées par les citoyens… Une grande vigilance s'impose donc sur ces sujets. ●



© VIRALYFT, LICENCE PEXELS

ChatGPT repose sur un modèle utilisant un langage naturel et qui a été entraîné à « étudier » des millions de documents en ligne afin de pouvoir générer des réponses à des requêtes aussi simples que sur un moteur de recherche. A sa mise à disposition, l'IA générative a atteint un million d'utilisateurs en seulement cinq jours…

(6) Dans son « Avis sur la surveillance de l'espace public » du 20 juin 2024, la Commission nationale consultative des droits de l'homme (CNCDH) estime que si « garantir la sécurité publique est, certes, un objectif légitime, […] cela doit toutefois donner lieu à un examen circonstancié, en partant d'une exigence de minimisation de la présence et de l'impact [des caméras] dans l'espace public » (www.cncdh.fr/sites/default/files/2024-06/A%20-%202024%20-%205%20-%20CNCDH%20-%20Avis%20Surveillance%20de%20l%27espace%20public%2C%20juin%202024_0.pdf).

(7) A l'initiative de la Quadrature du Net et avec treize organisations de la société civile : www.ldh-france.org/lalgorithme-de-notation-de-la-cnaf-attaque-devant-le-conseil-detat-par-15-organisations-2/.

(8) Allocation adulte handicapé et revenu de solidarité active.

(9) Caisse nationale d'assurance maladie.

(10) Traitement des antécédents judiciaires : fichier de police judiciaire qui contient des informations sur les personnes mises en cause et sur les victimes.

(11) Cette notion renvoie à la fois à l'usage de l'IA (*deep*, pour *deep learning*, « apprentissage profond ») et à la manipulation (*fake*, qui veut dire « faux »).

(12) Travailleurs précaires qui effectuent des tâches simples et répétitives pour alimenter l'intelligence artificielle, pour une rémunération très faible.

(13) Organisations de la société civile.