Nations Unies A/73/348*

Assemblée générale

Distr. générale
29 août 2018
Français
Original : anglais

Soixante-treizième session
Point 74 b) de l'ordre du jour provisoire**
Promotion et protection des droits de l'homme : questions relatives aux droits de l'homme, y compris les divers moyens de mieux assurer l'exercice effectif des droits de l'homme et des libertés fondamentales

Promotion et protection du droit à la liberté d'opinion et d'expression***

Note du Secrétaire général

Le Secrétaire général a l'honneur de transmettre à l'Assemblée générale le rapport établi par le Rapporteur spécial sur la promotion et la protection du droit à la liberté d'opinion et d'expression, David Kaye, soumis en application de la résolution 34/18 du Conseil des droits de l'homme. Dans le présent rapport, le Rapporteur spécial examine les incidences des technologies d'intelligence artificielle sur les droits de l'homme dans le cyberespace, en mettant particulièrement l'accent sur le droit à la liberté d'opinion et d'expression, à la vie privée et à la non-discrimination.

* Nouveau tirage pour raisons techniques (6 octobre 2018).
** A/73/150.
*** Le présent document est soumis après la date prévue pour que l'information la plus récente puisse y figurer.

18-14238* (F) 301018 301018

Rapport du Rapporteur spécial sur la promotion et la protection du droit à la liberté d'opinion et d'expression

Table des matières

I. Introduction

1. Partout dans le monde, l'intelligence artificielle (IA) exerce une influence sans cesse croissante sur le cyberespace. Les entreprises s'en servent pour éditer les résultats de recherche et des fils d'actualités et pour déterminer le placement des publicités, organisant ainsi ce que voient les utilisateurs et à quel moment ils le voient. Les plateformes de médias sociaux utilisent des techniques d'IA pour modérer leurs contenus, et celles-ci servent souvent de première ligne de défense face aux éléments susceptibles d'enfreindre leurs règles d'utilisation. L'IA recommande des personnes à suivre ou à mettre sur sa liste d'amis, des articles de presse à lire, des lieux à visiter, ainsi que des commerces, des restaurants ou des hôtels à fréquenter. Cet outil apporte aux grandes entreprises du numérique un moyen rapide, efficace et modulable de gérer les énormes quantités de contenus téléchargées chaque jour sur leurs plateformes d'information et de communication. Les techniques d'IA peuvent élargir et accélérer la mutualisation des contenus et des idées à l'échelle mondiale et ouvrir ainsi des perspectives inestimables pour la liberté d'expression et l'accès à l'information. En même temps, l'opacité de l'IA risque aussi de porter atteinte à l'autodétermination individuelle, ou à ce que le présent rapport appelle « l'autonomie et la faculté d'agir de l'individu »[1]. Un grand défi mondial se pose à tous ceux qui œuvrent à promouvoir les droits de l'homme et l'état de droit : Comment les États, les entreprises et la société civile peuvent-ils faire en sorte que les techniques d'intelligence artificielle respectent et renforcent les droits de l'homme plutôt que de les fragiliser et de les menacer ?

2. Le présent rapport ne vise pas à épuiser le sujet des relations entre IA et droits de l'homme, mais plutôt à réaliser trois objectifs : définir les termes clefs indispensables pour examiner l'IA sous l'angle des droits de l'homme ; déterminer quel cadre juridique fondé sur les droits de l'homme s'applique à l'IA ; formuler quelques recommandations préliminaires visant à faire en sorte qu'au fur et à mesure de l'évolution des technologies constituant l'IA, des mesures de protection des droits de l'homme soient intégrées à ce processus. La lecture du présent rapport est faite pour accompagner celle de mon dernier rapport au Conseil des droits de l'homme (A/HRC/38/35), qui présente une approche de la modération des contenus en ligne fondée sur les droits de l'homme[2].

II. Comprendre l'intelligence artificielle

A. Qu'est-ce que l'intelligence artificielle ?

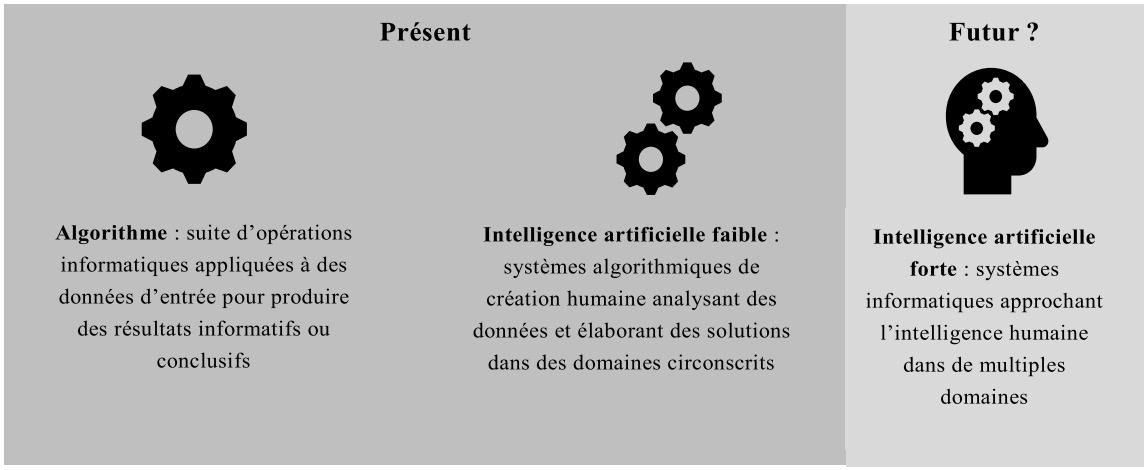
3. Le terme « intelligence artificielle » sert souvent de raccourci pour évoquer l'indépendance, la vitesse et l'ampleur sans cesse croissantes des moyens informatiques de prise de décision. L'IA n'est pas un seul objet, mais plutôt une constellation de techniques et de procédés permettant d'utiliser des ordinateurs pour accompagner ou remplacer des opérateurs humains dans des tâches de résolution de

[1] Voir Mariarosaria Taddeo et Luciano Floridi, « How AI can be a force for good », *Science*, vol. 361, nº 6404 (24 août 2018). Disponible à l'adresse suivante : http://science.sciencemag.org/content/361/6404/751.full.

[2] Le présent rapport a bénéficié d'une consultation d'experts organisée à Genève en juin 2018 avec un financement de l'Union européenne, ainsi que de l'apport de différents experts dans le cadre de l'élaboration du document A/HRC/35/38 en 2017 et 2018. Le Rapporteur spécial tient notamment à remercier Carly Nyst et Amos Toh, dont le travail de recherche et de rédaction a joué un rôle essentiel dans la réalisation de ce rapport.

problèmes ou de prise de décision[3]. Ce terme peut poser problème, en ce qu'il donne à penser que des machines sont capables de fonctionner selon les mêmes règles et les mêmes principes que l'intelligence humaine alors qu'il n'en est rien. En général, l'IA procède par essais itératifs pour optimiser les tâches informatisées que des opérateurs humains affectent à des ordinateurs. Cela dit, comme c'est l'expression consacrée dans le langage courant et dans le monde de l'entreprise et de l'administration, c'est celle que le Rapporteur spécial utilise dans le présent rapport.

4. La culture populaire conduit souvent à penser que la société s'achemine, à échéance encore lointaine, vers l'avènement de l'IA forte, à savoir la capacité d'un système informatique à approcher ou surpasser l'intelligence humaine dans de multiples domaines (et, partant, vers ce qu'on appelle « la singularité »)[4]. L'avenir prévisible continuera de donner lieu à des progrès dans le champ de l'IA faible, où les systèmes informatiques exécutent des tâches (des algorithmes conçus par des humains) dans des domaines circonscrits. C'est par exemple sur l'IA faible que reposent l'assistance vocale des appareils mobiles, les agents conversationnels utilisés pour le service client, les outils de traduction en ligne, les voitures autonomes, les moteurs de recherche et les services de cartographie. Certaines techniques d'IA faible s'utilisent dans l'apprentissage automatique, qui consiste à entraîner des algorithmes à des tâches de reconnaissance et de résolution de problèmes à partir d'ensembles de données. Par exemple, les appareils domotiques intelligents « apprennent » en continu à partir des données qu'ils recueillent sur les formes d'expression du langage courant, afin d'interpréter les questions de leurs utilisateurs avec plus de précision et de mieux y répondre. Dans tous les cas, les humains jouent un rôle essentiel dans la conception et dans la diffusion des systèmes d'IA, dans la définition des objectifs de chaque application et, selon le type de celle-ci, dans la sélection et l'étiquetage des ensembles de données et dans la classification des résultats qu'elle produit. Ce sont toujours des humains qui déterminent les applications de l'IA et les utilisations des résultats qu'elle produit, notamment la mesure dans laquelle elles accompagnent ou remplacent la prise de décision humaine.

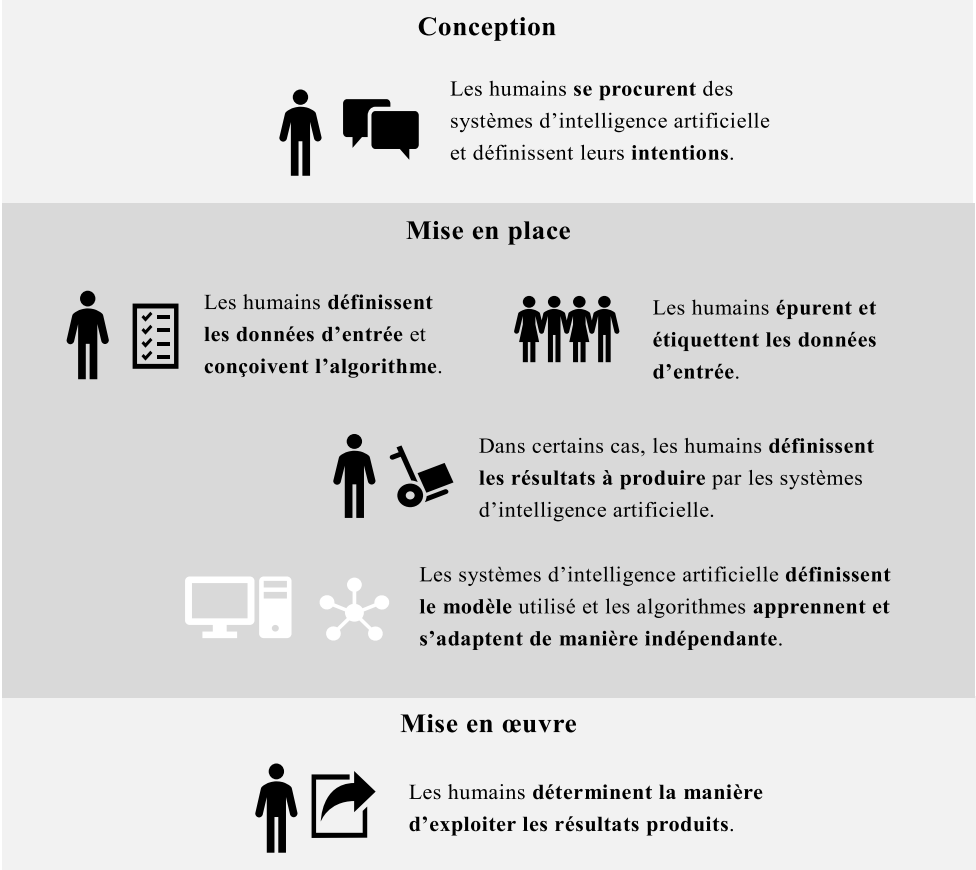


5. L'IA repose sur des algorithmes, c'est-à-dire des suites d'opérations informatiques conçues et encodées par l'homme sous forme d'instructions

[3] Voir AI Now, « The AI Now Report: The social and economic implications of artificial intelligence technologies in the near term », 2016. Disponible à l'adresse suivante : https://ainowinstitute.org/AI_Now_2016_Report.pdf. Royaume-Uni de Grande-Bretagne et d'Irlande du Nord, Commission spéciale de la Chambre des Lords sur l'intelligence artificielle, « AI in the United Kingdom: ready, willing and able? », 2018, p. 13.

[4] Article 19 et Privacy International, « Privacy and freedom of expression in an age of artificial intelligence », Londres, 2018, p. 8.

transformant des données d'entrée en résultats informatifs ou conclusifs. Les algorithmes sont depuis longtemps essentiels au fonctionnement des systèmes d'infrastructure et de communication de notre quotidien. L'IA se nourrit de l'énorme volume de données produit par la vie moderne et du besoin de l'analyser. C'est à n'en pas douter le point de vue du secteur privé, pour qui plus on dispose de données pour alimenter les algorithmes et plus ces données sont de bonne qualité, plus lesdits algorithmes peuvent gagner en précision et en puissance. Les systèmes algorithmiques peuvent analyser rapidement de gigantesques volumes de données et permettent ainsi aux programmes d'IA d'assumer des fonctions de prise de décision auparavant dévolues à des humains agissant sans outils informatiques.



Bien que l'action humaine fasse partie intégrante de l'IA, celle-ci présente des particularités qui méritent un examen attentif du point de vue des droits de l'homme dans au moins trois de ses aspects, à savoir l'automatisation, l'analyse des données et l'adaptabilité[5].

6. **Automatisation**. L'automatisation élimine l'intervention humaine de certaines parties de la prise de décision en affectant des tâches circonscrites à des outils informatiques. Cela peut avoir des incidences positives pour les droits de l'homme, sous réserve que le modèle utilisé limite l'influence des préjugés humains. Par exemple, un système automatisé de contrôle de l'entrée aux frontières peut signaler

[5] Conseil de l'Europe, *Algorithmes et droits humains – Études sur les dimensions des droits humains dans les techniques de traitement automatisé des données et éventuelles implications réglementaires*, étude du Conseil de l'Europe DGI(2017)12, 2018. Disponible à l'adresse suivante : https://rm.coe.int/algorithms-and-human-rights-fr/1680795681, p. 5.

les personnes nécessitant un contrôle approfondi selon des critères objectifs tels que les antécédents criminels ou le type de visa, limitant ainsi le poids des évaluations subjectives (et sujettes aux préjugés) fondées sur l'apparence physique, l'appartenance ethnique, l'âge ou la religion. L'automatisation permet également de traiter d'énormes quantités de données à une vitesse et à une échelle qui dépassent largement les capacités des humains, ce qui peut présenter un intérêt dans les domaines de la santé, de la sécurité publique et de la sécurité nationale. Toutefois, les systèmes automatisés utilisent des ensembles de données dont la conception ou l'application pourrait comporter des préjugés et produire ainsi des effets discriminatoires. Il se peut par exemple que les données relatives aux antécédents criminels ou au type de visa (mentionnées plus haut) incorporent elles-mêmes des éléments tendancieux. Un excès de dépendance et de confiance envers les décisions automatisées et la non-reconnaissance de ce point essentiel peuvent à leur tour compromettre la vérification des résultats produits par l'IA et priver certaines personnes de tout recours face à des décisions prises à leur encontre sur la base de ces résultats. L'automatisation peut nuire à la transparence et à la vérifiabilité d'un processus et empêcher ainsi même les autorités bien intentionnées de donner une explication des résultats[6].

7. **Analyse des données.** La plupart des applications d'IA reposent sur de vastes ensembles de données. Tout ensemble de données (portant sur des domaines allant des habitudes de navigation sur Internet jusqu'aux flux de circulation autoroutiers) peut servir de base à un système d'IA. Certains de ces ensembles contiennent des informations personnelles, tandis que beaucoup d'autres comportent des données anonymisées. L'utilisation de ces ensembles de données par l'IA suscite d'importantes préoccupations, notamment en ce qui concerne leur origine, leur exactitude, les droits des personnes à leur égard, la capacité des applications à désanonymiser leur contenu et les préjugés qu'elles pourraient renfermer ou se voir instiller lors de leur étiquetage ou de l'entraînement d'un système par des opérateurs humains. L'évaluation des données opérée par l'IA peut mettre en évidence des corrélations qui ne sont pas nécessairement des relations de cause à effet, et cela peut conduire à des conclusions erronées et tendancieuses qui sont difficiles à vérifier.

8. **Adaptabilité**. Les systèmes d'apprentissage automatique sont adaptables, car leurs algorithmes sont capables, en progressant par paliers successifs, de déceler de nouveaux problèmes et d'y apporter de nouvelles réponses. Selon leur niveau de supervision, les systèmes peuvent discerner la présence de formes ou de régularités et aboutir à des conclusions que leurs programmateurs ou opérateurs humains n'avaient pas prévues. Cette imprévisibilité porte en elle la véritable promesse de l'IA en tant qu'outil transformateur, mais elle fait aussi ressortir ses risques : au fur et à mesure que les humains sont exclus de la définition des objectifs visés et des résultats produits par un système d'IA, il devient de plus en plus difficile de garantir la transparence, la responsabilité et l'accès à un recours utile, ainsi que de prévoir et d'atténuer les effets préjudiciables aux droits de l'homme.

B. L'intelligence artificielle dans le cyberespace

9. L'IA a des conséquences particulièrement importantes, et parfois problématiques, pour le cyberespace, c'est-à-dire l'écosystème complexe que constituent les technologies et plateformes numériques ainsi que les acteurs publics et privés qui donnent accès à l'information et assurent sa diffusion. Les algorithmes et programmes d'IA sont désormais omniprésents sur Internet (dans les moteurs de recherche, les plateformes de médias sociaux, les applications de messagerie et les

[6] Conseil de l'Europe, *Algorithmes et droits humains*, p. 8.

mécanismes d'information), sur les appareils numériques et dans les systèmes informatiques. Dans l'optique de son mandat, le Rapporteur spécial constate que les trois usages de l'IA présentés ci-après suscitent des préoccupations.

10. **Affichage et personnalisation des contenus**. Les médias sociaux et les plateformes de recherche régissent de plus en plus la manière dont les utilisateurs accèdent aux informations et aux idées et les communiquent, ainsi que les modes de diffusion des nouvelles. Des algorithmes et applications d'IA déterminent à quelle échelle, à quel moment, à quels publics et à quels utilisateurs particuliers sont communiqués les différents contenus. De gigantesques ensembles de données combinant historiques de navigation, caractéristiques démographiques, analyses sémantiques, analyses des sentiments et bien d'autres facteurs alimentent des algorithmes fondés sur des modèles de hiérarchisation et d'édition de contenu de plus en plus personnalisés qui sélectionnent les informations à présenter à chaque utilisateur ou en excluent certaines de manière implicite. Certains contenus payants, promotionnels ou associés à un mot-dièse (« hashtag ») peuvent alors être promus au prix de l'exclusion ou de la rétrogradation d'autres informations. Les fils d'actualités des médias sociaux affichent leur contenu en fonction d'une évaluation subjective de l'intérêt ou de l'attrait qu'un élément peut présenter pour un utilisateur donné ; par conséquent, il se peut que certaines informations sociales et politiques essentielles ne s'affichent que rarement ou pas du tout sur l'écran d'un utilisateur, alors qu'elles sont par ailleurs disponibles sur leurs plateformes [7]. L'IA façonne le monde de l'information d'une manière opaque pour l'utilisateur et, souvent, même pour la plateforme responsable de l'édition des contenus.

11. La recherche en ligne est l'un des domaines utilisant le plus l'IA dans l'affichage et la personnalisation des contenus. Pour produire leurs résultats en réponse aux requêtes (et pour compléter ou prédire celles-ci), les moteurs de recherche utilisent des systèmes d'IA traitant de vastes quantités de données individuelles et collectives concernant les utilisateurs. Comme les contenus classés en mauvaise position ou entièrement exclus des résultats de recherche n'ont guère de chances d'être vus, les applications d'IA intervenant dans la recherche ont une énorme influence sur la diffusion des connaissances[8]. Les agrégateurs de contenus et les sites d'actualités[9] sélectionnent aussi les informations à présenter à un utilisateur non pas selon leur nouveauté ou leur importance, mais au moyen d'applications d'IA prédisant ses centres d'intérêt et ses habitudes de consultation des contenus à partir de vastes ensembles de données. Par conséquent, l'IA joue un rôle important, mais généralement occulte, dans la détermination des informations que les utilisateurs consomment consciemment ou non.

---

[7] World Wide Web Foundation, « The invisible curation of content: Facebook's News Feed and our information diets », 22 avril 2018. Disponible à l'adresse suivante : https://webfoundation.org/research/the-invisible-curation-of-content-facebooks-news-feed-and-our-information-diets.

[8] Conseil de l'Europe, *Algorithmes et droits humains*, p. 17.

[9] Voir par exemple « How Reuters's revolutionary AI system gathers global news », *MIT Technology Review*, 27 novembre 2017. Disponible à l'adresse suivante : www.technologyreview.com/s/609558/how-reuterss-revolutionary-ai-system-gathers-global-news. Paul Armstrong et Yue Wang, « China's $11 billion news aggregator Jinri Toutiao is no fake », *Forbes*, 26 mai 2017. Disponible à l'adresse suivante : www.forbes.com/sites/ywang/2017/05/26/jinri-toutiao-how-chinas-11-billion-news-aggregator-is-no-fake/#1d8b97804d8a.

12. L'emploi de l'IA dans le domaine de l'affichage des contenus entraîne une personnalisation croissante de l'expérience en ligne de chaque utilisateur ; dans une ère marquée par la profusion de l'information [10], cette personnalisation promet d'ordonner le chaos qui règne sur Internet en permettant aux utilisateurs de trouver les informations qu'ils demandent. Cela peut avoir l'avantage de donner accès aux contenus et aux services dans une plus grande palette de langues [11] ou à des informations plus actuelles correspondant mieux à l'expérience ou aux préférences personnelles de chacun. La personnalisation pilotée par IA peut aussi diminuer fortement l'exposition aux différents points de vue, portant ainsi atteinte à la faculté personnelle de rechercher et d'échanger activement des idées et des opinions en transcendant les clivages idéologiques, politiques ou sociétaux. Une telle personnalisation risque de renforcer les préjugés et d'encourager la promotion et la recommandation de contenus outrageants, incendiaires ou relevant de la désinformation dans le but de capturer plus longtemps l'attention des utilisateurs [12]. Il est bien entendu que toutes sortes de contextes sociaux et culturels peuvent limiter l'exposition d'une personne à certaines informations. Mais en optimisant la présentation des informations de manière à amplifier leur viralité et à stimuler l'activité en ligne des utilisateurs, la personnalisation pilotée par IA risque d'hypothéquer la faculté de ces derniers à trouver et à choisir certains types de contenus. Cela est d'autant plus vrai qu'en règle générale, les algorithmes rétrogradent le classement des contenus suscitant moins d'activité et relèguent ainsi à l'obscurité certains contenus produits par des utilisateurs ou par des médias indépendants [13]. Lorsque des systèmes d'IA à base de règles sont optimisés pour retenir l'attention et stimuler l'activité des utilisateurs, certains acteurs habiles et avisés peuvent en tirer parti pour accroître leur visibilité et, en s'appropriant des mots-dièse en vogue ou en utilisant des agents logiciels automatiques ou semi-automatiques (« bots »), pour s'octroyer une présence en ligne démesurée au détriment de la diversité de l'information.

13. **Modération et élimination des contenus**. L'IA aide les entreprises de médias sociaux à modérer les contenus en fonction des règles et principes de leurs plateformes, notamment au moyen de filtres antispam, de techniques de reconnaissance d'empreinte numérique (servant par exemple à détecter les contenus associés au terrorisme ou à l'exploitation des enfants), de filtres de mots-clés, de systèmes de traitement automatique du langage naturel (pour évaluer la nature des contenus par dépistage des mots ou expressions imagées pouvant signaler un problème) et d'autres algorithme de détection. Les systèmes d'IA peuvent envoyer des avertissements aux comptes utilisateurs enfreignant les conditions générales

---

[10] Carly Nyst et Nick Monaco, « State-Sponsored Trolling: How Governments are Deploying Disinformation as Part of Broader Digital Harassment Campaigns » (Palo Alto, Californie, Institute for the Future, 2018), p. 8.

[11] World Wide Web Foundation, « Artificial intelligence: the road ahead in low- and middle-income countries », Washington, juin 2017. Disponible à l'adresse suivante : https://webfoundation.org/docs/2017/07/AI_Report_WF.pdf.

[12] Zeynep Tufekci, « YouTube, the great radicaliser » New York Times, 10 mars 2018. Disponible à l'adresse suivante : www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html. James Williams, *Stand Out of our Light: Freedom and Resistance in the Attention Economy* (Cambridge, Massachusetts, Cambridge University Press, 2018).

[13] Récemment, certaines plateformes en ligne ont signalé leur intention de tourner le dos à une personnalisation fondée sur la capture de l'attention au profit d'une personnalisation privilégiant la qualité de l'expérience utilisateur. Voir Julia Carrie Wong, « Facebook overhauls News Feed in favour of "meaningful social interactions" », *The Guardian*, 11 janvier 2018. Disponible à l'adresse suivante : www.theguardian.com/technology/2018/jan/11/facebook-news-feed-algorithm-overhaul-mark-zuckerberg. Toutefois, tant que les paramètres d'évaluation employés par les systèmes d'intelligence artificielle ne sont pas communiqués et mesurés en toute transparence, il est difficile de déterminer si ce changement a vraiment un effet sur l'expérience des internautes.

d'utilisation, voire les suspendre ou les désactiver, et peuvent aussi servir à bloquer ou à filtrer les sites Web associés à des domaines, données ou contenus interdits. Les entreprises de médias sociaux ont recours à l'IA pour filtrer l'ensemble des contenus contrevenant à leurs règles d'utilisation (nudité, harcèlement, discours haineux, etc.), mais on ne sait pas dans quelle mesure et dans quels cas ce filtrage s'effectue automatiquement et sans intervention humaine[14].

14. Le rôle de l'IA progresse sous l'effet d'une poussée exercée tant par le secteur privé que par le secteur public. Les entreprises font valoir que le volume des contenus illégaux, inconvenants ou dommageables dépasse largement les capacités de modération humaine et que l'IA est un meilleur outil pour s'attaquer à ce problème. Selon certaines plateformes, l'IA offre non seulement une plus grande efficacité dans la détection des contenus inconvenants (selon leurs règles d'utilisation) ou illégaux devant être éliminés (en général par un modérateur humain), mais aussi une plus grande exactitude que la prise de décision humaine. Dans le même temps, les États demandent instamment que soient mis en place des systèmes de modération automatique efficaces et rapides afin d'accomplir tout un éventail d'objectifs distincts allant du dépistage des contenus associés au terrorisme ou à l'exploitation sexuelle des enfants (où l'IA est déjà largement employée) jusqu'à la protection du droit d'auteur en passant par la suppression des contenus « haineux » ou « extrémistes »[15]. La Recommandation de la Commission européenne sur les mesures destinées à améliorer davantage l'efficacité de la lutte contre les contenus illicites en ligne, datée du 1er mars 2018, engage les plateformes Internet à utiliser des filtres automatiques pour détecter et supprimer les contenus à caractère terroriste, en prévoyant des vérifications humaines lorsque cela se justifie pour remédier aux erreurs inévitables des systèmes automatisés[16].

15. Les efforts destinés à automatiser la modération des contenus peuvent avoir un coût en termes de droits de l'homme (voir A/HRC/38/35, par. 56). La modération de contenu pilotée par IA présente plusieurs inconvénients, notamment la difficulté d'évaluer le contexte et de prendre en compte la grande variabilité des indices langagiers, des significations et des particularités linguistiques et culturelles. Comme les applications d'IA reposent souvent sur les ensembles de données renfermant des préjugés discriminatoires et que, dans le contexte actuel, les conséquences d'une

---

[14] Un outil disponible sur Instagram, appelé DeepText, essaie d'estimer la « toxicité » du contexte, permet aux utilisateurs de personnaliser leurs propres filtres de mots et d'émojis, et évalue également les relations entre utilisateurs pour tenter d'évaluer plus finement le contexte (en déterminant par exemple si un commentaire est simplement une plaisanterie entre amis). Andrew Hutchinson, « Instagram's rolling out new tools to remove "toxic comments" », *Social Media Today*, 30 juin 2017. Disponible à l'adresse suivante : www.socialmediatoday.com/social-networks/instagrams-rolling-out-new-tools-remove-toxic-comments.

[15] Le Royaume-Uni aurait mis au point un outil permettant de détecter et de supprimer automatiquement les contenus à caractère terroriste au stade du téléchargement. Voir, par exemple, Ministère britannique de l'intérieur, « New technology revealed to help fight terrorist content online », communiqué de presse du 13 février 2018. Voir Commission européenne, Proposition de directive du Parlement européen et du Conseil sur le droit d'auteur dans le marché unique numérique, COM (2016) 593 final, art. 13. Lettre du Rapporteur spécial au Président de la Commission européenne, référence no AC OTH 41/2018, 13 juin 2018. Disponible à l'adresse suivante : www.ohchr.org/Documents/Issues/Opinion/Legislation/OL-OTH-41-2018.pdf.

[16] Recommandation de la Commission du 1er mars 2018 sur les mesures destinées à lutter, de manière efficace, contre les contenus illicites en ligne (C(2018) 1177 final). Disponible à l'adresse suivante : https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=50117. Voir également Daphne Keller, « Comment in response to European Commission's March 2018 recommendation on measures to further improve the effectiveness of the fight against illegal content online », Stanford Law School, Center for Internet and Society, 29 mars 2018. Disponible à l'adresse suivante : http://cyberlaw.stanford.edu/publications/comment-response-european-commissions-march-2018-recommendation-measures-further.

surmodération sont peu coûteuses, le risque est grand que ces systèmes procèdent par défaut à la suppression de contenus licites ou à la suspension de comptes ne posant pas de problème[17]. Il en résulte que les groupes vulnérables sont ceux qui ont le plus de chances d'être désavantagés par les systèmes de modération de contenus fondés sur l'IA. Par exemple, le système DeepText, utilisé par Instagram, a assimilé le mot « mexicain » a une insulte pour la seule raison que ses ensembles de données avaient été alimentés de textes associant ce terme à l'adjectif « illégal », qui était lui-même codé comme négatif dans l'algorithme[18].

16. Avec l'IA, il est difficile de vérifier la logique conduisant aux décisions affectant les contenus. Même lorsque la modération automatique s'accompagne d'une vérification humaine – formule que les plateformes de médias sociaux jugent de plus en plus impraticable à leur échelle de fonctionnement actuelle –, une tendance à s'en remettre aux décisions de la machine (censées s'appuyer sur des critères objectifs, comme on l'a vu plus haut) s'oppose à la mise en question des résultats, en particulier lorsque la conception technique du système occulte la logique sous-jacente.

17. **Profilage, publicité et ciblage**. L'IA s'est développée en symbiose mutuelle avec le modèle d'entreprise Internet fondé sur les données, selon lequel les utilisateurs disposent de contenus et de services gratuits en échange de leurs données personnelles. Au fil de nombreuses années de suivi et de profilage des internautes, les entreprises ont amassé des ensembles de données d'une grande richesse avec lesquels elles alimentent leurs systèmes d'IA pour mettre au point des modèles de prédiction et de ciblage encore plus précis. Désormais, les annonceurs publics et privés ont accès au ciblage individuel des publicités : consommateurs et électeurs font ainsi l'objet d'un véritable microciblage conçu pour s'adapter aux particularités de chacun et pour en tirer parti.

18. Le ciblage piloté par IA encourage la généralisation du recueil et de l'exploitation des données personnelles et accroît le risque de manipulation individuelle des utilisateurs par diffusion de fausses informations. Le ciblage peut perpétuer la discrimination et l'exclusion que subissent certains utilisateurs en les privant de certaines informations ou de certaines possibilités, par exemple en autorisant la diffusion d'offres d'emploi ou de logement excluant les travailleurs âgés, les femmes ou les personnes appartenant à des minorités ethniques[19]. Plutôt que d'exposer les utilisateurs de manière égale à la diversité des messages politiques, par exemple, le microciblage qui s'opère sur les réseaux sociaux crée une vision du monde filtrée et formatée accueillant mal le pluralisme de l'expression politique.

---

[17] Voir Aylin Caliskan, Joanna Bryson et Arvind Narayanan, « Semantics derived automatically from language corpora contain human-like biases », *Science*, vol. 356, nº 6334 (14 avril 2017). Solon Barocas et Andrew Selbst, « Big data's disparate impact », *California Law Review*, vol. 104, nº 671 (2016).

[18] Nicholas Thompson, « Instagram's Kevin Systrom wants to clean up the &#%@! Internet », Wired, 14 août 2017. Disponible à l'adresse suivante : www.wired.com/2017/08/instagram-kevin-systrom-wants-to-clean-up-the-internet.

[19] Julia Angwin, Noam Scheiber et Ariana Tobin, « Dozens of companies are using Facebook to exclude older workers from job ads », ProPublica, 20 décembre 2017. Disponible à l'adresse suivante : www.propublica.org/article/facebook-ads-age-discrimination-targeting. Julia Angwin, Ariana Tobin et Madeleine Varner, « Facebook (still) letting housing advertisers exclude users by race », ProPublica, 21 novembre 2017. Disponible à l'adresse suivante : www.propublica.org/article/facebook-advertising-discrimination-housing-race-sex-national-origin.

III. Un cadre juridique fondé sur les droits de l'homme pour l'intelligence artificielle

A. Portée des obligations en matière de droits de l'homme dans le contexte de l'intelligence artificielle

19. Les outils d'IA, comme tous moyens techniques, doivent être conçus, mis au point et mis en place dans le respect des obligations des États et des responsabilités des acteurs du secteur privé en vertu du droit international des droits de l'homme. Le droit des droits de l'homme impose aux États à la fois l'obligation négative de s'abstenir de prendre des mesures qui entravent l'exercice du droit à la liberté d'opinion et d'expression et l'obligation positive de promouvoir et de protéger ce droit.

20. Pour ce qui concerne le secteur privé, les États sont tenus de garantir le respect des droits individuels, et en particulier le droit à la liberté d'opinion et d'expression, notamment en protégeant les individus contre les actes commis par des parties privées qui y porteraient atteinte (paragraphe premier de l'article 2 du Pacte international relatif aux droits civils et politiques article)[20]. Les États peuvent s'acquitter de cette obligation en prenant des mesures juridiques destinées à restreindre ou à influencer la création et la mise en œuvre des applications d'IA, en adoptant des politiques réglementant les marchés publics pour l'acquisition d'applications d'IA auprès d'entreprises du secteur privé, en établissant des mécanismes d'autorégulation et de corégulation et en renforçant la capacité des entreprises du secteur privé à reconnaître l'importance du droit à la liberté d'opinion et d'expression et à lui donner la priorité dans leurs activités.

21. En vertu du droit des droits de l'homme, les entreprises ont également des responsabilités qui doivent guider leur conduite dans la création, l'adoption et la mise en œuvre des applications d'IA (A/HRC/38/35, par. 10). Selon les « Principes directeurs relatifs aux entreprises et aux droits de l'homme : mise en œuvre du cadre de référence « protéger, respecter et réparer » des Nations Unies, « la responsabilité de respecter les droits de l'homme est une norme de conduite générale que l'on attend de toutes les entreprises où qu'elles opèrent » (principe 11) et cette norme s'applique notamment aux entreprises exploitant des plateformes de médias sociaux ou des moteurs de recherche. Pour adapter les conclusions des Principes directeurs au domaine de l'IA (ibid., par. 11), ces derniers prévoient que les entreprises devraient, au minimum : formuler au plus haut niveau des engagements de principe concernant le respect des droits de leurs utilisateurs dans toutes les applications d'IA (principe 16) ; éviter d'avoir des incidences négatives sur les droits de l'homme ou d'y contribuer par leur utilisation des outils d'IA et veiller à prévenir et atténuer toutes incidences négatives liées à leurs activités (principe 13) ; faire preuve de diligence raisonnable à l'égard des systèmes d'IA pour identifier les incidences effectives et potentielles de leurs activités sur les droits de l'homme et y remédier (principes 17 à 19) ; mettre en place des stratégies de prévention et d'atténuation (principe 24) ; passer constamment en revue les activités liées à l'IA, notamment en consultant les parties prenantes et le public (principes 20 et 21) ; prévoir des recours accessibles pour réparer les effets négatifs que les systèmes d'IA peuvent avoir sur les droits de l'homme (principes 22, 29 et 31).

[20] Voir le principe 3 des Principes directeurs relatifs aux entreprises et aux droits de l'homme : mise en œuvre du cadre de référence « protéger, respecter et réparer » des Nations Unies (A/HRC/17/31) et A/HRC/38/35, par. 6 à 8.

B. Droit à la liberté d'opinion

22. Le droit de ne pas être inquiété pour ses opinions est un droit absolu, consacré par le paragraphe premier de l'article 19 du Pacte international relatif aux droits civils et politiques et par l'article 19 de la Déclaration universelle des droits de l'homme. Il « n'admet ni exception ni limitation », que ce soit « par la loi ou par toute autre autorité »[21]. Dans son rapport de 2015 au Conseil des droits de l'homme sur le recours au chiffrement et à l'anonymat dans le domaine des échanges numériques (A/HRC/29/32), le Rapporteur spécial a fait observer que les modalités de stockage, de transmission et de sécurisation de l'information employées à l'ère du numérique avaient un effet singulier sur l'exercice du droit à la liberté d'opinion. En effet, l'ensemble des activités et enregistrements numériques d'un utilisateur (requêtes de recherche, pages Web visitées, communications par courriel et par messagerie instantanée, notes et documents hébergés sur des serveurs délocalisés) forme la trame de ses opinions (ibid., par. 12), et des acteurs tant étatiques que non étatiques pourraient s'ingérer dans la mécanique de ces processus de formation et de maintien des opinions.

23. Un élément essentiel du droit à la liberté d'opinion est le « droit de se forger une opinion et de l'enrichir par le raisonnement »[22]. Le Comité des droits de l'homme a conclu que ce droit passe nécessairement par la liberté de toute contrainte injustifiée dans la dynamique des croyances, idéologies, réactions et prises de position d'un individu[23]. Par conséquent, les interventions neurologiques, les programmes d'endoctrinement (de type « camps de rééducation ») ou les menaces de violence visant à obliger des personnes à modifier leurs opinions ou à en adopter d'autres constituent une violation du paragraphe premier de l'article 19 du Pacte. Le Comité a également déterminé que la coercition par « incitations sous la forme d'un traitement préférentiel » pouvait atteindre un niveau de persuasion qui porte atteinte au droit de se faire et de maintenir des opinions (voir CCPR/C/78/D/878/1999).

24. L'emploi des nouvelles techniques d'édition de contenus soulève des questions inédites quant aux types de moyens de coercition ou d'incitation pouvant être considérés comme une atteinte au droit de se faire une opinion. L'édition des contenus influence depuis longtemps la capacité d'une personne à se faire ses propres opinions : par exemple, les organes médiatiques mettent certaines informations à la une dans l'intention de façonner et d'influencer ce que chacun sait des événements de la journée. La publicité commerciale vise quant à elle à susciter des opinions favorables et à cultiver un désir à l'égard de certains produits et services.

25. L'emploi de l'IA élargit et enrichit la tradition de l'édition des contenus sur Internet en apportant des moyens plus élaborés et plus efficaces de personnaliser les contenus présentés à chaque utilisateur et en portant cette pratique à une échelle inaccessible aux médias traditionnels. La prédominance de certains modes d'édition assistée par IA suscite des inquiétudes quant à son impact sur la capacité de l'individu à constituer et enrichir ses propres opinions. Par exemple, un petit nombre d'entreprises Internet se partagent la très grande majorité des requêtes de recherche. En raison du monopole qu'exercent ces entreprises sur le marché de la recherche en ligne, il est extrêmement difficile pour les utilisateurs de s'affranchir des algorithmes

---

[21] Comité des droits de l'homme, Observation générale nº 34 (2011) sur la liberté d'opinion et la liberté d'expression, par. 9. Disponible à l'adresse suivante : https://tbinternet.ohchr.org/_layouts/treatybodyexternal/Download.aspx?symbolno=CCPR%2fC%2fGC%2f34&Lang=fr. Manfred Nowak, *U.N. Covenant on Civil and Political Rights: CCPR Commentary* (1993).

[22] M. Nowak, *UN Covenant on Civil and Political Rights*.

[23] *Yong Joo-Kang c. République de Corée*, communication nº 878/1999 du Comité des droits de l'homme, 16 juillet 2003 (CCPR/C/78/D/878/1999).

de classement et d'édition des résultats, et certains peuvent aussi être portés à croire (comme le souhaitent ces entreprises) que les résultats produits constituent les informations les plus pertinentes ou les plus objectives qui soient disponibles sur un sujet particulier. Le manque de transparence sur la manière dont les systèmes d'IA structurent et appliquent les critères de recherche peut aussi renforcer l'idée que les résultats produits par telle ou telle plateforme constituent une présentation objective composée d'informations factuelles.

26. Les situations de position dominante qui se sont établies dans le domaine de l'édition assistée par IA remettent au premier plan la question de savoir dans quelle mesure le traitement des contenus agit ou non sur la capacité de chacun à se faire une opinion. Le caractère inédit des problèmes soulevés, conjugué à l'absence générale de jurisprudence sur les atteintes à la liberté d'opinion, suscite plus de questions que de réponses quant à l'impact de l'édition assistée par IA sur les droits de l'homme dans l'environnement numérique contemporain. Néanmoins, ces questions devraient stimuler la recherche relative aux effets de l'édition de contenus assistée par IA sur les droits sociaux, économiques et politiques. Les entreprises devraient, au minimum, fournir des informations suffisantes et compréhensibles sur la façon dont elles structurent et appliquent les critères d'édition et de personnalisation des contenus sur leurs plateformes, en explicitant notamment leurs politiques et procédures de détection des préjugés sociaux, culturels ou politiques pouvant intervenir dans la conception et dans la mise au point des systèmes d'IA concernés.

C. Droit à la liberté d'expression

27. Le paragraphe 2 de l'article 19 du Pacte garantit largement le droit « de rechercher, de recevoir et de répandre des informations et des idées de toute espèce », en précisant que ce droit doit être protégé et respecté sans considération de frontières ou de moyens de communication. L'exercice du droit à la liberté d'expression est étroitement lié à celui des autres droits et joue un rôle fondamental dans le bon fonctionnement des institutions démocratiques ; dans ces conditions, la protection, le respect et la promotion du droit à la liberté d'expression s'accompagnent nécessairement de l'obligation de promouvoir la diversité et l'indépendance des médias et de protéger l'accès à l'information.[24]

28. À la différence du droit à la liberté de se faire et de maintenir des opinions, le droit à la liberté d'expression et d'accès aux informations et aux idées peut être soumis à des restrictions dans un nombre limité de cas (paragraphe 3 de l'article 19 du Pacte). Ces restrictions doivent satisfaire aux principes de légalité (c'est-à-dire être expressément fixées par une loi satisfaisant aux normes de clarté et de précision et interprétées par des autorités judiciaires indépendantes), de nécessité et de proportionnalité (c'est-à-dire ne restreindre la liberté que dans la mesure nécessaire pour protéger l'intérêt légitime en jeu, sans porter atteinte à l'essence même du droit), et de légitimité (c'est-à-dire protéger exclusivement l'un des intérêts légitimes

[24] Rapporteur spécial sur la promotion et la protection du droit à la liberté d'opinion et d'expression, Représentant pour la liberté des médias de l'Organisation pour la sécurité et la coopération en Europe, Rapporteur spécial sur la liberté d'expression de l'Organisation des États américains et Rapporteur spécial sur la liberté d'expression et l'accès à l'information de la Commission africaine des droits de l'homme et des peuples, « Joint Declaration on freedom of expression and "fake news", disinformation and propaganda » [Déclaration conjointe sur la liberté d'expression et les fausses nouvelles (« fake news »), la désinformation et la propagande], 3 mars 2017. Disponible à l'adresse suivante (en anglais, russe et slovène seulement) : www.osce.org/fom/302796. Voir également l'observation générale nº 34 (2011) du Comité des droits de l'homme sur la liberté d'opinion et la liberté d'expression, ainsi que le paragraphe 61 du rapport A/HRC/29/32 et le paragraphe 86 du rapport A/HRC/32/38.

énumérés dans le Pacte, à savoir les droits ou la réputation d'autrui, la sécurité nationale, l'ordre public et la santé ou la moralité publique (A/HRC/38/35, par. 7). Dans ce cadre, le droit à la liberté d'expression peut aussi être restreint en application du paragraphe 2 de l'article 20 du Pacte, qui exige des États qu'ils interdisent « tout appel à la haine nationale, raciale ou religieuse qui constitue une incitation à la discrimination, à l'hostilité ou à la violence », mais ces restrictions doivent toujours répondre aux conditions cumulatives de légalité, de nécessité et de légitimité[25].

29. La complexité de la prise de décisions inhérente à la modération des contenus peut être poussée à l'extrême par l'introduction de processus automatisés. À la différence des humains, les algorithmes sont actuellement incapables d'évaluer le contexte culturel, de détecter l'ironie d'un discours ou de procéder à l'analyse critique requise pour reconnaître avec précision, par exemple, un contenu « extrémiste » ou un discours haineux, en conséquence de quoi ils risquent davantage de procéder par défaut au blocage ou à la restriction de certains contenus, et ainsi de porter atteinte au droit qu'a chaque utilisateur d'être entendu et d'accéder aux moyens d'information sans restriction ni censure[26].

30. Dans un système régi par des algorithmes d'IA, la diffusion des informations et des idées est dictée par des processus opaques dont les priorités pourraient aller à l'encontre du maintien d'un environnement favorisant la diversité des médias et l'expression de voix indépendantes. À ce propos, le Comité des droits de l'homme a conclu que les États « devraient prendre les mesures voulues […] pour empêcher une domination ou concentration indue des organes d'information par des groupes de médias contrôlés par des intérêts privés dans des situations de monopole qui peuvent être préjudiciables à la diversité des sources et des opinions »[27].

31. Les utilisateurs n'ont pas non plus accès aux règles du jeu lorsqu'ils utilisent des plateformes et sites Web pilotés par IA. Le manque d'information sur le périmètre et la portée des opérations exécutées en ligne par des systèmes algorithmiques et des applications d'IA empêche les individus de comprendre à quel moment et selon quels critères les différentes informations sont diffusées, filtrées ou ciblées. Les petites améliorations concédées en vue d'atténuer ce problème, comme l'identification sélective des résultats de recherche à caractère promotionnel ou la mise en évidence des publicités financées par des acteurs politiques sur les plateformes de médias sociaux, peuvent aider un tant soit peu à comprendre les règles du cyberespace, mais ces mesures ne prennent pas en compte ni ne résolvent les interrogations quant à la mesure dans laquelle les processus algorithmiques façonnent cet environnement[28].

32. Même lorsque les utilisateurs sont informés de l'existence, de la portée et du fonctionnement des systèmes d'IA, ces derniers peuvent encore se dérober aux efforts déployés pour leur donner une transparence convenable. À ce jour, on ne dispose toujours pas d'outils suffisamment perfectionnés et modulables pour vérifier et rendre transparents les processus qui sous-tendent les décisions automatisées des plateformes en ligne[29]. Cela signifie que les utilisateurs subissent souvent des

---

[25] Comité des droits de l'homme, Observation générale nº 34 (2011) sur la liberté d'opinion et la liberté d'expression, par. 50.

[26] Conseil de l'Europe, *Algorithmes et droits humains*, p. 21.

[27] Comité des droits de l'homme, Observation générale nº 34 (2011) sur la liberté d'opinion et la liberté d'expression, par. 40.

[28] Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (New York, New York University Press, 2018).

[29] Mike Ananny et Kate Crawford, « Seeing without knowing: limitations of the transparency ideal and its application to algorithmic accountability », *New Media and Society*, vol. 20, nº 3 (13 décembre 2016). Disponible à l'adresse suivante : http://journals.sagepub.com/doi/abs/10.1177/1461444816676645?journalCode=nmsa.

atteintes à leur liberté d'expression sans disposer d'aucun recours pour étudier ou élucider les principes, mécanismes ou critères qui en sont la cause.

D. Droit à la vie privée

33. Le droit à la vie privée sert souvent de voie d'accès à l'exercice de la liberté d'opinion et d'expression[30]. L'article 17 du Pacte protège chaque personne contre les « immixtions arbitraires ou illégales dans sa vie privée, sa famille, son domicile ou sa correspondance » et contre les « atteintes illégales à son honneur et à sa réputation », et prévoit que « toute personne a droit à la protection de la loi contre de telles immixtions ou de telles atteintes ». Le Haut-Commissariat des Nations Unies aux droits de l'homme et le Conseil des droits de l'homme ont souligné que toute restriction du droit à la vie privée devait satisfaire aux principes de légalité, de nécessité et de proportionnalité (paragraphe 23 du rapport A/HRC/27/37 et paragraphe 2 de la résolution 34/7 du Conseil des droits de l'homme).

34. Les systèmes de prise de décision pilotés par IA dépendent du recueil et de l'exploitation de différentes informations allant des données techniques non personnelles jusqu'aux renseignements nominatifs, la grande majorité des données utilisées pour alimenter les systèmes d'IA se situant quelque part à mi-chemin (données inférées ou extraites à partir de données personnelles, ou données personnelles anonymisées, quoique souvent de manière imparfaite). Les entreprises utilisent des données provenant du profilage en ligne et de l'empreinte numérique des utilisateurs, se procurent des jeux de données auprès de tiers (notamment des courtiers en information) et produisent de nouvelles données par agrégation de vastes ensembles de données afin d'alimenter les systèmes d'IA. Les systèmes autonomes et appareils grand public pilotés par IA sont fréquemment équipés de capteurs qui génèrent et recueillent de grandes quantités de données sur les personnes se trouvant à proximité[31], et les plateformes de médias sociaux utilisent des méthodes d'IA pour inférer et générer à propos de ces personnes des informations sensibles qu'elles n'ont ni fournies ni confirmées, comme leur orientation sexuelle, leurs liens de parenté, leurs convictions religieuses, leur état de santé ou leur affiliation politique.

35. L'IA remet en question les notions traditionnelles de consentement, de limitation des finalités et des utilisations, de transparence et de responsabilité, autrement dit les piliers sur lesquels reposent les normes internationales de protection des données[32]. Comme les systèmes d'IA fonctionnent en exploitant des ensembles de données existants et en en créant de nouveaux, ils produisent un contexte où les personnes sont privées de tout moyen pratique de savoir et de comprendre comment leurs données sont utilisées ou d'avoir prise sur ces utilisations. Une fois que des données sont utilisées à de nouvelles fins dans un système de l'IA, elles perdent leur contexte d'origine, ce qui augmente le risque d'inexactitude ou de péremption des informations associées aux personnes concernées et prive celles-ci de toute possibilité de les faire corriger ou supprimer. Les systèmes d'IA utilisant ces données servent à prendre des décisions importantes dont certaines ont de lourdes conséquences sur la vie des gens, et pourtant, les individus ont peu d'options pour exercer un contrôle sur les données dérivées de leurs renseignements personnels, alors même que les techniques d'anonymisation présentent encore des failles[33].

---

[30] Voir le paragraphe 16 du rapport A/HRC/29/32, la résolution 68/167 de l'Assemblée générale et la résolution 20/8 du Conseil des droits de l'homme.

[31] Article 19 et Privacy International, « Privacy and freedom of expression ».

[32] Comité des droits de l'homme, Observation générale n° 16 (1988) sur le droit au respect de la vie privée, par. 10.

[33] Article 19 et Privacy International, « Privacy and freedom of expression », p. 19.

E. Obligation de non-discrimination

36. La non-discrimination est un principe intrinsèque du droit des droits de l'homme, qui existe non seulement pour préciser l'obligation qu'ont des États de garantir l'exercice de tous les autres droits de l'homme sans discrimination, mais aussi, comme énoncé à l'article 26 du Pacte, pour garantir de manière indépendante l'égalité devant la loi et l'égale protection de la loi. Les États ont l'obligation incontestable d'« interdire toute discrimination et [de] garantir à toutes les personnes une protection égale et efficace contre toute discrimination, notamment de race, de couleur, de sexe, de langue, de religion, d'opinion politique et de toute autre opinion, d'origine nationale ou sociale, de fortune, de naissance ou de toute autre situation ». Ainsi, les articles 17 et 19 intègrent les droits individuels de protection contre la discrimination au droit à la liberté d'opinion, au droit à la liberté d'expression et d'accès aux idées et à l'information, ainsi qu'au droit à la vie privée et à la protection des données personnelles.

37. Le potentiel qu'a l'IA d'incorporer et de perpétuer des préjugés et des facteurs de discrimination s'étend à la discrimination dans l'exercice de la liberté d'opinion et d'expression. Il peut arriver que les algorithmes de modération ne prennent pas en compte les contextes et sensibilités liées à la culture, à la langue ou à l'appartenance sexuelle, voire l'intérêt du public pour un contenu[34]. Les fils d'actualité pilotés par IA peuvent perpétuer et renforcer des attitudes discriminatoires, tandis que les systèmes de profilage et de publicité fondés sur l'IA ont manifestement contribué à la discrimination sur des critères de race, de religion et de sexe[35]. Les fonctions de saisie semi-automatique basées sur l'IA ont également produit des résultats discriminatoires sur le plan racial[36].

38. Un certain nombre de facteurs amalgament des préjugés dans les systèmes d'IA et augmentent ainsi leur potentiel discriminatoire. Il s'agit notamment de la façon dont les systèmes d'IA sont conçus, des décisions quant à l'origine et à la portée des ensembles de données utilisés pour entraîner ces systèmes, des préjugés culturels et sociétaux que les concepteurs peuvent intégrer dans ces ensembles de données, des modèles d'IA eux-mêmes et de la manière dont les résultats produits par ces modèles sont utilisés en pratique. Par exemple, les applications de reconnaissance faciale ont le défaut de s'appuyer sur des ensembles de données portant majoritairement sur des hommes blancs, si bien qu'ils présentent un taux d'erreurs allant jusqu'à 20 % chez les femmes et chez les personnes à la peau plus foncée[37]. Lorsque ces systèmes sont utilisés, par exemple, pour catégoriser des images accessibles à partir d'un moteur de recherche, leur potentiel discriminatoire peut se traduire par de véritables atteintes

---

[34] Cela a conduit, par exemple, à la suppression de photographies présentant une grande importance historique et culturelle. Voir Julia Carrie Wong, « Mark Zuckerberg accused of abusing power after Facebook deletes "napalm girl" post », *The Guardian*, 9 septembre 2016. Disponible à l'adresse suivante : www.theguardian.com/technology/2016/sep/08/facebook-mark-zuckerberg-napalm-girl-photo-vietnam-war. Voir également le paragraphe 29 du rapport A/HRC/38/35.

[35] Julia Angwin, Madeleine Varner et Ariana Tobin, « Facebook enabled advertisers to reach "Jew haters" », ProPublica, 14 septembre 2017. Disponible à l'adresse suivante : www.propublica.org/article/facebook-enabled-advertisers-to-reach-jew-haters. Ariana Tobin, « Why we had to buy racist, sexist, xenophobic, ableist and otherwise awful Facebook ads », ProPublica, 27 novembre 2017. Disponible à l'adresse suivante : www.propublica.org/article/why-we-had-to-buy-racist-sexist-xenophobic-ableist-and-otherwise-awful-facebook-ads.

[36] Paris Martineau, « YouTube's search suggests racist autocompletes », The Outline, 13 mai 2018. Disponible à l'adresse suivante : https://theoutline.com/post/4536/youtube-s-search-autofill-suggests-racist-results?zd=1&zi=3ygzt6hw.

[37] Joy Buolamwini, « The dangers of supremely white data and the coded gaze », présentation au congrès Wikimania 2018 tenu au Cap, en Afrique du Sud. Disponible à l'adresse suivante : https://www.youtube.com/watch?v=ZSJXKoD6mA8&feature=youtu.be.

aux droits qu'ont les personnes de demander, recevoir et communiquer des informations et de se réunir ou de s'associer librement.

F. Droit à un recours utile

39. Le droit des droits de l'homme garantit que l'autorité compétente (judiciaire, administrative ou législative) statuera sur les droits de toute personne qui forme un recours (paragraphe 3 de l'article 2 du Pacte). Les recours doivent être connus et à la portée de toute personne dont les droits ont été violés, doivent donner lieu à des enquêtes rapides, approfondies et impartiales sur les allégations de violations [38] et doivent pouvoir faire cesser les violations en cours (A/HRC/27/37, par. 39 à 41).

40. Les systèmes d'IA portent souvent atteinte au droit à un recours. En premier lieu, l'avis à l'utilisateur est pour ainsi dire introuvable. Dans presque toutes les applications d'IA animant le cyberespace, les utilisateurs ne connaissent pas la portée, l'étendue ou même l'existence des algorithmes de prise de décision susceptibles d'avoir une incidence sur l'exercice de leurs droits à la liberté d'opinion et d'expression. Le second obstacle, et le plus difficile à surmonter, est celui de la vérifiabilité du système d'IA lui-même. La logique conduisant aux décisions prises par les algorithmes n'est pas nécessairement évidente même pour un expert connaissant bien les mécanismes sous-jacents du système. S'il est logique de supposer qu'une plus grande transparence des systèmes d'IA et de leurs algorithmes permettrait de mieux surveiller leur fonctionnement, cela n'équivaut pas forcément à rendre intelligibles les processus de prise de décision. Les algorithmes peuvent occulter le fait qu'une décision lourde de conséquences a été prise, ou être assez complexes et assez dépendants du contexte pour se dérober aux explications. La situation est rendue encore plus compliquée par le fait que les entreprises exploitant des plateformes dans le cyberespace actualisent fréquemment leurs algorithmes [39] ; pour couronner le tout, les applications d'apprentissage automatique peuvent modifier leurs propres règles et algorithmes au fil du temps.

41. Ces difficultés sont encore aggravées par le passage à l'automatisation des systèmes de recours eux-mêmes ; dans ce cas de figure, les plaintes des utilisateurs, concernant soit des décisions de modération des contenus, soit des atteintes aux droits de l'homme causées par les systèmes d'IA, sont à leur tour examinées et évaluées par de tels systèmes[40]. Ces processus de réponse automatique suscitent des inquiétudes quant à la question de savoir si leurs mécanismes de résolution des plaintes constituent un recours utile, vu qu'ils sont dénués de capacité d'analyse contextuelle, de liberté d'appréciation et de faculté de jugement indépendant[41].

---

[38] Comité des droits de l'homme, Observation générale n° 31 (2004) sur la nature de l'obligation juridique générale imposée aux États parties au Pacte, par. 15.

[39] Barry Schwartz, « Google: we make thousands of updates to search algorithms each year », Search Engine Roundtable, 5 juin 2015. Disponible à l'adresse suivante : www.seroundtable.com/google-updates-thousands-20403.html.

[40] Conseil de l'Europe, *Algorithmes et droits humains*, p. 24.

[41] Pei Zhang, Sophie Stalla-Bourdillon et Lester Gilbert, « A content-linking-context model for "notice-and-take-down" procedures », *WebSci' 16*, mai 2016. Disponible à l'adresse suivante : http://takedownproject.org/wp-content/uploads/2016/04/ContentLinkingModelZhangStallaGilbert.pdf.

G. Législations, règlements et politiques adoptées à l'égard de l'intelligence artificielle

42. De nombreux États sont en train d'élaborer des stratégies nationales en vue de concevoir et mettre au point des politiques et initiatives destinées à maximiser les avantages potentiels de l'IA pour leurs citoyens. Bien qu'aucun État n'ait encore proposé de loi ou règlement régissant l'IA dans tous ses aspects, une telle démarche appelle la prudence, car elle pourrait être mal adaptée à un domaine aussi innovant et compenser le manque de détails par des dispositions trop restrictives ou trop permissives[42]. Une réglementation sectorielle pourrait être préférable, et il n'est pas exclu que les lois et règlements existants, par exemple dans le domaine de la protection des données, s'appliquent de manière assez large et avec suffisamment de souplesse pour qu'on se passe de législation supplémentaire.

43. Dans le même temps, les États devraient veiller à ce que le développement de l'IA s'effectue en accord avec les normes des droits de l'homme. Tous les efforts engagés par les États en vue d'établir une politique ou un règlement dans le domaine de l'IA devraient prendre en compte les questions de droits de l'homme[43]. Le droit à la liberté d'opinion et d'expression, en particulier, est souvent exclu des débats sur l'IA qui se tiennent dans la sphère publique et politique, car lorsqu'ils abordent les questions de droits de l'homme, ils ont tendance à se concentrer sur le problème des préjugés et de la discrimination dans la prestation des services.

44. Comme la mise au point de systèmes d'IA efficaces nécessite l'acquisition de grands ensembles de données et des investissements à long terme dans les moyens techniques, il est probable que les capacités de conception et de production resteront largement entre les mains d'entités du secteur privé et que le secteur public dépendra davantage de celles-ci pour accéder à ces systèmes. Pour cette raison, les intérêts des entreprises et des pouvoirs publics risquent de devenir de plus en plus inextricablement imbriqués et d'être perçus comme tels par le public. Cela est particulièrement vrai dans le cyberespace, où les organes publics sont souvent utilisateurs – de plateformes de médias sociaux, de moteurs de recherche et d'autres services informatiques – plutôt que fournisseurs. Le rapprochement des intérêts du public et du privé ne constitue pas en soi une source d'atteintes aux droits de l'homme, mais il soulève des questions de transparence et de responsabilité. Au fur et à mesure que l'IA poursuit son développement entre les mains du secteur privé, les États risquent fort de se mettre à déléguer aux entreprises des missions de surveillance et de censure devenant sans cesse plus complexes et plus ardues.

45. Tout projet de loi ou de politique publique touchant le domaine de l'IA devrait porter sur les applications du secteur privé tout autant que sur celles du secteur public, plutôt que de chercher à réglementer seulement ces dernières. Pour reprendre la conclusion du Conseil de l'Europe, « les questions de gouvernance et/ou de réglementation des algorithmes relèvent de la sphère publique et ne devraient pas être laissées aux mains des seuls acteurs privés »[44]. En réponse aux préoccupations que suscite l'IA, les États pourraient engager des efforts de réglementation visant à

[42] Tim Dutton, « An overview of national AI strategies », Medium, 28 juin 2018. Disponible à l'adresse suivante : https://medium.com/politics-ai/an-overview-of-national-ai-strategies-2a70ec6edfd.

[43] Il est préoccupant, par exemple, qu'un comité parlementaire du Royaume-Uni de Grande-Bretagne et d'Irlande du Nord ait publié un rapport de 200 pages ne mentionnant pas une seule fois les droits de l'homme. Voir Royaume-Uni de Grande-Bretagne et d'Irlande du Nord, Commission spéciale de la Chambre des Lords sur l'intelligence artificielle, « AI in the United Kingdom ».

[44] Conseil de l'Europe, *Algorithmes et droits humains*, p. 44.

renforcer les obligations de transparence et de communication des entreprises et à établir par la loi des règles de protection des données à la fois efficaces et fiables.

46. On assiste actuellement, de la part des secteurs public et privé, à une prolifération d'initiatives visant à incorporer des principes d'éthique aux activités d'acquisition, de conception, de mise en place et de mise en œuvre des systèmes d'IA. Le Rapporteur spécial encourage vivement les acteurs concernés à intégrer les impératifs de protection des droits de l'homme à ces efforts. Les efforts déployés par le secteur privé pour promouvoir des principes d'éthique et par les pouvoirs publics pour encourager leur adoption procèdent souvent d'une résistance à la mise en place d'une réglementation fondée sur les droits de l'homme[45]. Bien que les principes d'éthique constituent un cadre essentiel pour surmonter certaines difficultés dans le domaine de l'IA, ils ne sauraient se substituer aux droits de l'homme, que chaque État est tenu de faire respecter. Les entreprises et les pouvoirs publics devraient veiller à ce que les principes de protection des droits de l'homme et les mécanismes de responsabilité afférents soient fermement intégrés à tous les aspects de leurs activités d'IA, même s'ils sont déjà en train d'élaborer des codes de déontologie et des principes d'éthique[46].

IV. Une approche de l'intelligence artificielle fondée sur les droits de l'homme

47. Dans de récents rapports, le titulaire du mandat a énoncé des mesures juridiques et pratiques que les entreprises pourraient prendre pour mettre les principes des droits de l'homme au cœur de leurs politiques de régulation des contenus, et il a exposé en détail des normes et mécanismes de fond propres à garantir que les entreprises puissent s'acquitter, dans tous les aspects de leurs activités, de leurs responsabilités en application des Principes directeurs relatifs aux entreprises et aux droits de l'homme. C'est ce même cadre qui structure la démarche proposée dans le présent rapport pour le cas des systèmes d'IA. Les normes et mécanismes de fond proposés ci-après s'appliquent aux entreprises, en leur qualité d'acteurs assurant la conception, la mise en place et la mise en œuvre des systèmes d'IA, ainsi qu'aux États, qui sont tenus au premier chef de ne pas porter atteinte aux droits de l'homme dans leur adoption et leur utilisation des systèmes d'IA. Ces normes et mécanismes sont conçus pour garantir que le droit des droits de l'homme occupe toujours une place centrale dans l'évolution du secteur de l'IA. Deux principes fondamentaux sont intégrés à tous les aspects des normes et mécanismes proposés : d'une part la nécessité de protéger et de respecter la faculté d'agir et l'autonomie de toute personne (une condition préalable essentielle à l'exercice du droit à la liberté d'opinion et d'expression), et d'autre part l'importance d'une véritable communication de la part des acteurs du secteur public et du secteur privé, qui se définit par des efforts ouverts et novateurs pour expliquer le fonctionnement des systèmes d'IA au public et faciliter leur surveillance.

A. Normes de fond pour les systèmes d'intelligence artificielle

48. Les entreprises devraient articuler la conception de leurs normes, règles et systèmes autour des principes universels des droits de l'homme (A/HRC/38/35, par. 41 à 43). Les conditions générales d'utilisation et les guides destinés au public

[45] Ben Wagner, « Ethics as an escape from regulation: from ethics-washing to ethics-shopping? », dans *Being Profiling. Cogitas Ergo Sum*, Mireille Hildebrandt, éd. (Amsterdam University Press, à paraître).

[46] Article 19 et Privacy International, « Privacy and freedom of expression », p. 13.

devraient être complétés par une politique interne prévoyant des mesures de transversalisation des droits de l'homme dans toutes les activités de l'entreprise, en particulier pour ce qui concerne la mise au point et la mise en œuvre des systèmes algorithmiques et des applications d'IA. Les entreprises devraient mener une réflexion sur la manière de mettre au point, à l'intention des ingénieurs en IA, des normes professionnelles présentant leurs responsabilités en matière de droits de l'homme sous la forme de principes à appliquer dans la conception technique des systèmes et dans les choix de fonctionnement. La mise en place de codes de déontologie et de structures institutionnelles d'accompagnement pourrait constituer un complément important aux mesures de protection des droits de l'homme, mais ils ne sauraient s'y substituer. Les codes et principes directeurs publiés par les organismes des secteurs public et privé devraient souligner le fait que c'est le droit des droits de l'homme qui établit les règles fondamentales de la protection des personnes dans le contexte de l'IA, tandis que les cadres déontologiques peuvent servir à élaborer plus avant le contenu et l'application des droits de l'homme dans des circonstances particulières.

49. Concernant les décisions prises dans le cyberespace, les entreprises et les pouvoirs publics devraient indiquer clairement aux utilisateurs lesquelles sont prises par des systèmes automatisés et lesquelles s'accompagnent d'une vérification humaine, ainsi que les grandes lignes de la logique suivie par ces systèmes. Les utilisateurs devraient aussi être informés dès le départ lorsque les données personnelles qu'ils fournissent à un acteur du secteur privé (de manière explicite ou par leur utilisation d'un site ou d'un service) doivent ensuite faire partie d'un ensemble de données utilisé par un système d'IA, afin qu'ils puissent tenir compte de ce facteur pour décider de consentir ou non au recueil des données et pour déterminer quels types de données ils souhaitent communiquer[47]. Un peu comme dans le cas des avis au public requis pour l'utilisation de caméras de télévision en circuit fermé, les systèmes d'IA devraient signaler aux utilisateurs de manière active, claire et compréhensible (par des moyens novateurs comme l'affichage de fenêtres contextuelles) qu'ils font l'objet d'une prise de décision pilotée par IA ou fournissent des données devant servir à une telle décision, et leur donner des informations suffisantes et compréhensibles sur la logique suivie par le système et sur l'importance des conséquences pour la personne concernée.

50. Assurer la transparence ne se limite pas à signaler aux utilisateurs des plateformes ou services en ligne que celles-ci intègrent des applications d'IA. Les entreprises et les pouvoirs publics devraient adhérer au principe de transparence pour chaque maillon de la chaîne de valeur de l'IA. Les mesures de transparence n'ont pas besoin d'être complexes pour être efficaces ; même des explications simplifiées sur la finalité, les principes et les données d'entrée et de sortie d'un système d'IA peuvent aider à informer le public et à alimenter la réflexion[48]. Plutôt que de s'évertuer à rendre la complexité des processus techniques intelligibles pour le profane, les entreprises devraient s'efforcer d'assurer la transparence d'un système en donnant des informations non techniques sur son fonctionnement. Il s'agit alors d'informer les utilisateurs sur l'existence, la finalité, la constitution et les effets d'un système d'IA, plutôt que sur le code source et les données d'entraînement, d'entrée et de sortie[49].

51. Assurer la transparence totale des effets d'un système d'IA dans le cyberespace nécessite de communiquer, par exemple, des données sur la quantité de contenus

[47] Comité des droits de l'homme, Observation générale nº 16 (1988) sur le droit au respect de la vie privée.

[48] Aaron Rieke, Miranda Bogen et David Robinson, « Public scrutiny of automated decisions: early lessons and emerging methods » (Omidyar et Upturn, 2018), p. 5.

[49] Rieke, Bogen et Robinson, « Public scrutiny of automated decisions », p. 8.

supprimée par les systèmes d'IA, la fréquence à laquelle des suppressions de contenus sont proposées par ces derniers et approuvées par un modérateur humain, la fréquence à laquelle les suppressions de contenus sont contestées et la fréquence à laquelle ces contestations sont validées ou non. Des données agrégées illustrant les tendances dans l'affichage des contenus devraient être mises à la disposition des utilisateurs, en même temps que des études de cas montrant pourquoi la priorité est donnée à certains contenus plutôt qu'à d'autre. La communication de l'origine et des bénéficiaires des publicités politiques et commerciales est essentielle à une transparence totale. Les acteurs des secteurs public et privé qui mettent en œuvre des systèmes d'IA devraient aussi faire preuve de transparence quant aux limites de ces derniers, en communiquant par exemple leurs mesures de confiance, leurs limites d'emploi et les scénarios de défaillance connus[50].

52. L'élimination des problèmes de discrimination dans les systèmes d'IA est un défi existentiel pour les entreprises comme pour les pouvoirs publics ; faute de supprimer les éléments discriminatoires et leurs effets, ces outils sont non seulement inefficaces, mais dangereux. Pour déterminer la manière d'éliminer les préjugés et les facteurs de discrimination dans les systèmes d'IA, les entreprises et les pouvoirs publics peuvent s'appuyer sur l'abondance des ressources et des travaux de réflexion existants ; il s'agit essentiellement de dépister et de neutraliser ces problèmes tant dans les données d'entrée que dans les données de sortie. Cela nécessite au minimum d'éliminer les erreurs d'échantillonnage (lorsque les ensembles de données ne sont pas représentatifs de la société), d'épurer les ensembles de données pour en retirer les éléments discriminatoires et de mettre en place des mesures destinées à compenser l'effet des données « portant l'empreinte de pratiques discriminatoires structurelles et héritées de l'histoire », à partir desquelles les systèmes d'IA risquent de produire des résultats indirectement discriminatoires[51]. La surveillance active des résultats produits par les systèmes d'IA est également essentielle pour détecter les effets discriminatoires et pouvoir ainsi prévenir ou atténuer d'éventuelles atteintes aux droits de l'homme.

B. Processus applicables aux systèmes d'intelligence artificielle

53. **Études d'impact sur les droits de l'homme**. L'établissement d'une transparence totale d'un bout à l'autre du cycle de vie de l'IA nécessite que les entreprises et les pouvoirs publics prennent des mesures pour permettre de vérifier et de mettre en doute le fonctionnement des systèmes depuis leur conception jusqu'à leur mise en œuvre. Les études d'impact sur les droits de l'homme sont un moyen de manifester la volonté de faire face aux incidences des systèmes d'IA sur les droits de l'homme ; elles et devraient être réalisées avant l'acquisition, la mise au point ou l'utilisation de ces derniers et comporter à la fois une évaluation en interne et un examen externe. Le groupe de réflexion AI Now a proposé un cadre de référence pour les études d'impact des moyens algorithmiques utilisés par les organes publics, selon lequel les pouvoirs publics devraient entreprendre une évaluation en interne des systèmes d'IA tout en coordonnant leur examen par des chercheurs externes de

[50] Amnesty International et Access Now, « Toronto declaration: protecting the right to equality and non-discrimination in machine learning systems », art. 31, al. d), 2018. Disponible à l'adresse suivante : www.accessnow.org/the-toronto-declaration-protecting-the-rights-to-equality-and-non-discrimination-in-machine-learning-systems.

[51] Iason Gabriel, « The case for fairer algorithms », Medium, 14 mars 2018. Disponible à l'adresse suivante : https://medium.com/@Ethics_Society/the-case-for-fairer-algorithms-c008a12126f8.

manière à tester et vérifier les hypothèses et les conclusions [52]. Les entreprises devraient également procéder à des évaluations conduites selon des principes similaires.

54. L'acquisition par le secteur public d'outils d'IA fournis par des entreprises du privé doit s'accompagner au préalable d'une consultation publique visant à recueillir l'avis des différents acteurs de la société sur leur conception et leur mise en œuvre. Les entreprises et les pouvoirs publics doivent engager des consultations suivies et approfondies faisant intervenir la société civile, les groupes de défense des droits de l'homme, les collectivités locales concernées et les représentants des populations historiquement marginalisés ou sous-représentés avant de mettre au point, d'acquérir ou d'utiliser des outils ou systèmes d'IA.

55. **Audits**. La coordination de l'examen des systèmes d'IA par des intervenants externes constitue une garantie de rigueur et d'indépendance essentielle pour répondre au souci de transparence. Pour cette raison, les études d'impact sur les droits de l'homme réalisées avant l'acquisition des systèmes d'IA devraient être complétées par des audits indépendants menés en continu, ces derniers constituant un mécanisme de transparence et de responsabilité important. Certains acteurs du secteur privé ont soulevé des objections quant à la faisabilité de tels audits dans la sphère de l'IA, invoquant à ce propos l'impératif de protection des techniques brevetées. Bien que ces préoccupations puissent être fondées, le Rapporteur spécial partage l'avis du groupe de réflexion AI Now selon lequel, s'agissant du fonctionnement d'un outil d'IA devant être utilisé par un organe du secteur public, tout refus de transparence de la part du fournisseur serait incompatible avec les obligations de responsabilité de l'organe en question.

56. De toute façon, les options innovantes permettant de soumettre les outils d'IA à des audits tout en maintenant le secret des renseignements exclusifs ne manquent pas : il est concevable de recourir à des protocoles de preuve sans divulgation de connaissance pour démontrer qu'un algorithme satisfait certaines propriétés, ce qui éviterait de devoir examiner sa structure sous-jacente[53] ; ou bien l'algorithme pourrait être communiqué à des tiers experts de confiance qui le conserveraient sous séquestre et sous condition de confidentialité, ce qui permettrait de procéder aux vérifications d'intérêt général sans divulguer l'algorithme au public[54]. Des organismes de contrôle publics opérant dans les domaines des télécommunications ou de la concurrence pourraient être autorisés à accéder aux systèmes d'IA sous conditions de confidentialité, comme cela se fait déjà, par exemple, pour le contrôle des machines à sous en Australie et en Nouvelle-Zélande, où les entreprises doivent soumettre leurs systèmes algorithmiques à des audits réglementaires[55]. Les publications spécialisées proposent d'autres méthodes innovantes pour la vérification des systèmes d'IA[56].

57. Chacun de ces mécanismes peut présenter des difficultés de mise en œuvre, en particulier dans le cyberespace, mais les entreprises devraient s'employer à rendre possibles les audits des systèmes d'IA. Les pouvoirs publics devraient contribuer à l'efficacité des audits en adoptant des politiques ou des mesures législatives obligeant

---

[52] Dillon Reisman *et al*., « Algorithmic impact assessments: a practical framework for public agency accountability », AI Now, 2018. Disponible à l'adresse suivante : https://ainowinstitute.org/aiareport2018.pdf.

[53] Conseil de l'Europe, *Algorithmes et droits humains*, p. 36.

[54] Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information* (Cambridge, Massachusetts, Harvard University Press, 2015).

[55] Conseil de l'Europe, *Algorithmes et droits humains*, p. 34.

[56] Christian Sandvig *et al*., « Auditing algorithms: research methods for detecting discrimination on Internet platforms », communication au forum Data and Discrimination: Converting Critical Concerns into Productive Inquiry, organisé en prélude au 64e congrès annuel de l'International Communication Association, à Seattle (État de Washington) le 22 mai 2014.

les entreprises à rendre vérifiable le code source de leurs applications d'IA, ce qui garantirait l'existence de journaux d'audit et rendrait le fonctionnement des systèmes plus transparents en cas de problèmes touchant des utilisateurs.

58. **Autonomie individuelle**. L'IA ne doit pas supplanter, manipuler ou entraver de manière invisible la capacité des personnes à se faire des opinions et à trouver ou exprimer des idées dans le cyberespace. Le respect de l'autonomie individuelle consiste, à tout le moins, à faire en sorte que les utilisateurs soient informés, puissent choisir et exercent un contrôle suffisant. Les applications d'IA omniprésentes et dissimulées qui occultent les mécanismes d'affichage, de personnalisation et de modération des contenus et ceux de profilage et de ciblage des utilisateurs hypothèquent la faculté de ces derniers à exercer leur droit à la liberté d'opinion et d'expression et au respect de la vie privée. Les entreprises devraient être attentives aux effets préjudiciables aux droits de l'homme qui découlent de l'emploi d'applications d'IA privilégiant des intérêts commerciaux ou politiques aux dépens de la transparence et des choix individuels.

59. **Avis et consentement**. Les entreprises doivent faire en sorte que les utilisateurs soient pleinement informés de la manière dont les décisions prises par des algorithmes façonnent leur utilisation d'une plateforme, d'un site ou d'un service. Cela peut s'obtenir en recourant à des campagnes d'information, à des fenêtres contextuelles, à des pages ou séquences interstitielles ou à d'autres moyens de signaler à quel moment un système d'IA détermine l'expérience de l'utilisateur d'un moteur de recherche, d'un site d'information ou d'une plateforme de médias sociaux. Les obligations d'information imposées par les États peuvent être un bon moyen de protéger le principe de l'avis et du consentement. Les utilisateurs disposent aussi du droit de savoir à quel moment une application d'IA recueille des données les concernant, si celles-ci font ensuite partie d'un ensemble de données devant servir à alimenter un système d'IA et selon quelles conditions elles sont utilisées, stockées et supprimées.

60. **Réparation**. Les incidences négatives des systèmes d'IA sur les droits de l'homme doivent être réparables et les entreprises responsables doivent les réparer. Comme préalable à la mise en place de procédures de recours utile, il est nécessaire de faire en sorte que les utilisateurs sachent qu'ils ont fait l'objet d'une décision prise par un algorithme (y compris lorsqu'elle a été proposée par un système d'IA et approuvée par un intervenant humain) et soient informés de la logique ayant conduit à cette décision. Par ailleurs, les entreprises devraient veiller à ce que les demandes de réparation soient examinées par un opérateur humain, en vue de procéder à une vérification adéquate des systèmes et de garantir le respect du principe de responsabilité. Elles devraient également publier des données sur la fréquence à laquelle les mécanismes de réparation sont déclenchés en réponse à des décisions prises par des outils d'IA.

V. Conclusions et recommandations

61. **Dans le présent rapport, le Rapporteur spécial a étudié les incidences existantes et potentielles de l'IA sur le droit à la liberté d'opinion et d'expression, en partant du principe que ces techniques constituent désormais une part essentielle du cyberespace et qu'elle présente des avantages et des risques pour l'exercice des droits des utilisateurs. Il a proposé un cadre conceptuel pour la réflexion sur les obligations des États et la responsabilité des entreprises quant à l'impératif de faire respecter ces droits face à l'expansion des capacités d'IA et a suggéré des mesures concrètes qui pourraient être mises en œuvre par les gouvernements et par les entreprises pour garantir le respect des droits de**

l'homme alors que la puissance, l'ampleur et la portée de ces outils continuent de croître.

Recommandations à l'intention des États

62. **Lorsqu'ils acquièrent ou mettent en place des systèmes ou applications d'IA, les États devraient veiller à ce que les organes du secteur public agissent conformément aux principes des droits de l'homme. Cela inclut, entre autres choses, l'organisation de consultations publiques et la réalisation d'études d'impact sur les droits de l'homme ou d'études d'impact des moyens algorithmiques utilisés par les organes publics avant l'acquisition ou le déploiement des systèmes d'IA. Il faudrait apporter une attention particulière aux effets multiples et variés que ces systèmes peuvent avoir sur les minorités raciales et religieuses, sur l'opposition politique et sur les groupes militants. Les systèmes d'IA mis en place par les pouvoirs publics devraient faire l'objet d'audits réguliers réalisés par des experts externes indépendants.**

63. **Les États devraient veiller à ce que les droits de l'homme occupent toujours une place centrale dans la conception, la mise en place et la mise en œuvre des systèmes d'IA du secteur privé. Cela consiste à actualiser la réglementation existante (et en particulier des lois sur la protection des données) afin de faciliter son application au domaine de l'IA, à mettre en place des dispositifs de régulation ou de corégulation obligeant les entreprises à soumettre leurs moyens d'IA à des études d'impact et à des audits, et à établir des mécanismes de contrôle externe efficaces[57]. Certaines applications d'IA peuvent nécessiter une réglementation sectorielle assurant une protection des droits de l'homme efficace. Dans la mesure où de telles restrictions produisent ou facilitent des atteintes à la liberté d'expression, les États devraient veiller à ce qu'elles soient nécessaires et proportionnées à la réalisation d'un objectif légitime en accord avec le paragraphe 3 de l'article 19 du Pacte. La réglementation relative à l'IA devrait également être élaborée dans le cadre d'une vaste consultation publique faisant intervenir la société civile, les groupes de défense des droits de l'homme et les représentants des utilisateurs appartenant à des groupes marginalisés ou sous-représentés.**

64. **Les États devraient adopter des politiques et mesures législatives propres à créer un cyberespace favorisant le pluralisme et la diversité. Il s'agit notamment de prendre des mesures destinées à promouvoir la concurrence dans le domaine de l'IA. Ces mesures peuvent inclure une réglementation des monopoles technologiques visant à prévenir une concentration des compétences techniques et du pouvoir de marché entre les mains de quelques entreprises dominantes, une réglementation visant à accroître l'interopérabilité des services et moyens techniques, ou encore l'adoption de politiques renforçant la neutralité des réseaux ainsi que celle des terminaux[58].**

Recommandations à l'intention des entreprises

65. **Tous les efforts visant à élaborer des codes et des principes directeurs en réponse aux implications éthiques des technologies d'IA devraient se fonder sur les principes des droits de l'homme. La société civile devrait être consultée pour**

---

[57] Wagner, « Ethics as an escape from regulation ».

[58] Autorité de régulation des communications électroniques et des postes, « Smartphones, tablettes, assistants vocaux : les terminaux, maillon faible de l'internet ouvert », février 2018. Disponible à l'adresse suivante : www.arcep.fr/uploads/tx_gspublication/rapport-terminaux-fev2018.pdf.

chaque projet de création ou de mise en place d'un système d'IA, qu'il relève du secteur privé ou du secteur public. Les entreprises devraient réaffirmer, dans leurs politiques et dans les orientations techniques qu'elles fixent aux ingénieurs, concepteurs, programmeurs, techniciens en traitement de données (de la saisie à l'épuration) et autres participants aux cycles de vie de l'IA, que leurs responsabilités en matière de droits de l'homme guident toutes leurs activités et que leur respect des règles d'éthique peut faciliter l'application des principes des droits de l'homme à chaque étape de conception, de mise en place et de mise en œuvre des moyens d'IA. En particulier, les conditions générales d'utilisation des plateformes devraient être fondées sur les principes universels des droits de l'homme.

66. **Les entreprises devraient indiquer de manière explicite à quel endroit et de quelle façon les techniques d'IA et d'automatisation sont utilisées sur leurs plateformes, services ou applications. Afin de donner aux utilisateurs les avis nécessaires à la compréhension et à la prise en compte de l'incidence des systèmes d'IA sur l'exercice de leurs droits de l'homme, il est essentiel de recourir à des moyens novateurs pour leur signaler à quel moment ils font l'objet d'une décision prise par un tel système, à quel moment un tel système intervient dans l'affichage ou la modération des contenus et à quel moment leurs données personnelles pourraient être intégrées dans un ensemble de données devant servir à alimenter des systèmes de ce type. Les entreprises devraient également publier des données sur les suppressions de contenus, notamment la fréquence à laquelle ces suppressions sont contestées et la fréquence à laquelle les contestations sont validées ou non, ainsi que des données illustrant les tendances dans l'affichage des contenus, accompagnées d'études de cas et d'informations sur le profilage politique et commercial.**

67. **Les entreprises devraient prévenir la discrimination et en déceler les facteurs dans les données d'entrées et dans les données de sortie des systèmes d'IA. Cela consiste, d'une part, à faire en sorte que les membres des équipes chargées de concevoir et de mettre en place les systèmes d'IA reflètent la diversité de la population et adoptent une attitude non discriminatoire et, d'autre part, à donner la priorité à la prévention des préjugés et de la discrimination dans le choix des ensembles de données et dans la conception des systèmes, notamment en remédiant aux erreurs d'échantillonnage, en épurant les données de manière à les débarrasser des éléments discriminatoires et en mettant en place des mesures destinées à compenser les effets de tels éléments. Il est également essentiel de surveiller activement les résultats produits par les systèmes d'IA de manière à détecter tout effet discriminatoire.**

68. **Il faudrait réaliser des études d'impact sur les droits de l'homme et des consultations publiques pendant la conception et la mise en place des nouveaux systèmes d'IA, ainsi que lors de la mise en place des systèmes existants dans de nouveaux marchés à l'international. Les consultations publiques devraient avoir lieu avant la mise au point définitive et la mise en service d'un produit ou service, pour qu'elles puissent faire une différence, et elles devraient faire intervenir la société civile, les défenseurs des droits de l'homme et les représentants des utilisateurs appartenant à des groupes marginalisés ou sous-représentés. Les résultats des études d'impact sur les droits de l'homme et des consultations publiques devraient être rendus publics.**

69. **Les entreprises devraient faire en sorte que l'intégralité du code source de leurs systèmes d'IA soit entièrement vérifiable et chercher des moyens novateurs de rendre possible un audit externe indépendant de ces systèmes, et ce,**

séparément des prescriptions réglementaires. Les résultats de ces audits et vérifications devraient être rendus publics.

70. **Les utilisateurs devraient avoir accès à des voies de recours pour obtenir réparation en cas d'atteintes aux droits de l'homme causées par des systèmes d'IA. Concernant ce qui précède, les entreprises devraient mettre en place des procédures garantissant que les plaintes, les demandes de réparation et les recours en appel soient examinés par des opérateurs humains et traités sans délai. Il faudrait également assurer la publication régulière des données relatives à la fréquence de ces plaintes et demandes de réparation, ainsi qu'aux types de recours disponibles et à leur efficacité respective.**