COUNCIL OF EUROPE

CONSEIL DE L'EUROPE

Strasbourg, 15 juin 2020 CAHAI(2020)07-fin

COMITE AD HOC SUR L'INTELLIGENCE ARTIFICIELLE (CAHAI)

Lignes directrices sur l'éthique en matière d'IA : situation en Europe et dans le monde

Rapport provisoire par Marcello Ienca* et Effy Vayena*

***Chaire de bioéthique, Laboratoire d'éthique et de politique de la santé, Département des sciences de la santé et de la technologie, ETH Zurich**

www.coe.int/cahai

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Laboratoire d'éthique et de politique de la santé
Département des sciences de la santé et de la technologie (D-HEST)

Résumé : Ces dernières années, des entreprises du secteur privé, des instituts de recherche et des organismes du secteur public ont énoncé des principes et diffusé des lignes directrices et d'autres instruments non contraignants au sujet de l'utilisation éthique de l'intelligence artificielle (IA). Cependant, malgré un consensus manifeste sur le fait que l'IA doit être « éthique », il y a débat sur ce qu'il faut entendre par « IA éthique » et sur les exigences éthiques, les normes techniques et les meilleures pratiques nécessaires à cet effet. Il s'agit ici d'analyser le corpus de documents de droit souple et autres cadres éthico-juridiques définis par des institutions gouvernementales et des entités non gouvernementales du monde entier, et ce à deux fins. Tout d'abord, nous souhaitons étudier cet ensemble d'instruments de gouvernance non contraignants qui est en constante évolution. Ensuite, nous voulons analyser les incidences que l'IA pourrait avoir sur les principes éthiques, les droits de l'homme, l'État de droit et la démocratie. Le rapport s'appuie sur un protocole d'examen de la portée, qui a été adapté et préalablement validé et dont l'objet était d'avoir un aperçu complet et actualisé des dispositions non contraignantes en vigueur. En tout, nous avons examiné 116 documents, entre autres publiés par des institutions publiques, des entités non gouvernementales, des établissements universitaires et des entreprises du secteur privé. Notre analyse fait ressortir cinq grands groupes de principes éthiques, dont elle examine la place dans le discours actuel sur la gouvernance. En revanche, elle montre qu'en l'état actuel des choses, le droit souple est lacunaire. Nous avons en outre établi un lien entre les principes éthiques et les droits de l'homme, plus spécifiquement les droits et libertés inscrits dans la Convention européenne des droits de l'homme (CEDH), pour évaluer dans quelle mesure les cadres de gouvernance non contraignants en vigueur offrent une protection intégrale des droits de l'homme. Enfin, nous avons dressé de façon empirique la liste des incidences sur l'élaboration des politiques, et ce à l'intention des scientifiques, des établissements de recherche, des organismes de financement, des organismes gouvernementaux et intergouvernementaux, et autres acteurs concernés par la promotion d'innovations éthiquement responsables en matière d'IA.

Principaux constats :

- De plus en plus d'institutions gouvernementales et d'entités non gouvernementales (notamment des entreprises du secteur privé et des établissements universitaires) élaborent des lignes directrices éthiques et d'autres instruments non contraignants au sujet de l'IA.
- Ces documents de droit souple sont principalement rédigés en Europe, en Amérique du Nord et en Asie. L'hémisphère sud est à l'heure actuelle sous-représenté parmi les organismes proposant des lignes directrices éthiques sur l'IA.
- À l'heure actuelle, les lignes directrices sur l'IA s'accordent le plus souvent sur certains principes généraux mais absolument pas sur les mesures concrètes à prendre. En outre, aucun des différents principes éthiques n'est commun aux 116 documents sur l'IA éthique que nous avons examinés.
- Nous avons constaté qu'un consensus se dégageait de plus en plus autour des principes éthiques suivants : transparence, justice, non-malfaisance, responsabilité et respect de la vie privée. Les notions éthiques de durabilité, de dignité et de solidarité semblent nettement moins traitées.
- Si la plupart des lignes directrices s'accordent à dire que l'IA doit être *transparente* pour éviter d'éventuels problèmes, on ne sait en revanche pas clairement si la transparence doit passer par la publication du code source, par les bases de données sous-jacentes ou par divers autres moyens.
- Un peu plus de la moitié des textes de droit souple qui ont été examinés recommandent explicitement de promouvoir les droits de l'homme ou mettent en garde contre leur violation, lors de la conception, du développement et du déploiement de systèmes d'IA.
- Les expressions courantes tirées des codes montrent d'importantes différences dans les thèmes qu'abordent les documents issus des pays membres du Conseil de l'Europe et ceux que traitent les documents élaborés ailleurs. Les documents de droit souple établis dans les pays membres du Conseil de l'Europe semblent mettre l'accent sur les principes éthiques de solidarité, de confiance et de fiabilité, contrairement aux documents établis dans des pays du reste du monde. Ils semblent en revanche faire plus rarement référence aux principes de bienfaisance et de dignité.
- Les principes de respect de la vie privée, de justice et de loyauté sont ceux qui varient le moins entre les pays membres du Conseil de l'Europe, les pays observateurs du Conseil de l'Europe et le reste du monde, et donc ceux qui ont le plus haut degré de constance transgéographique et transculturelle.

Principales incidences sur l'élaboration des politiques :

- Les instruments non contraignants que diffusent les institutions gouvernementales et les entités non gouvernementales (en ce compris les entreprises du secteur privé et les établissements universitaires) sont de précieux outils permettant d'exercer une influence sur les décisions publiques relatives à l'IA et d'orienter le développement des systèmes d'IA en faveur du bien-être social et du respect des valeurs éthiques et des normes légales. Il ne faut toutefois pas considérer qu'ils remplacent les instruments de gouvernance contraignants. Pour cause de conflit d'intérêts, il est particulièrement à craindre que les mesures d'autoréglementation prises par les acteurs du secteur privé de l'IA ne servent à contourner ou à écarter les mesures de gouvernance contraignantes édictées par les autorités gouvernementales ou intergouvernementales.
- Par souci d'inclusivité, de pluralisme culturel et de participation équitable à la prise de décisions collective au sujet de l'IA, il faudrait encourager l'élaboration de documents de droit souple par des organismes situés dans les régions du monde qui sont actuellement sous-représentées, tout particulièrement l'Afrique et l'Amérique du Sud.
- Les cinq principes éthiques génériques – transparence, justice, non-malfaisance, responsabilité et respect de la vie privée – autour desquels convergent les instruments non contraignants en vigueur sont à surveiller en priorité et pourraient faire l'objet de mesures de gouvernance obligatoires prises aux niveaux gouvernemental et intergouvernemental.
- Afin que ces principes éthiques puissent être dûment traduits en règles de gouvernance, ils

doivent être précisés sur le plan conceptuel. Les décideurs se doivent de remédier aux ambiguïtés sémantiques de ces principes et à leurs caractéristiques antagonistes.

- Les profondes divergences entre les textes de droit souple en vigueur au sujet de l'interprétation et de la mise en œuvre concrète de ces principes montrent que le public sera probablement divisé au sujet des solutions de gouvernance contraignantes et qu'un débat démocratique transparent sera donc nécessaire.
- Les notions éthiques qui sont très peu traitées, par exemple la durabilité, la dignité et la solidarité, doivent être analysées de très près pour que les lacunes conceptuelles et normatives que présente le droit souple ne soient pas transposées dans les mesures obligatoires.
- Comme près de la moitié des documents de droit souple que nous avons étudiés ne recommandent pas expressément de promouvoir les droits de l'homme – pas plus qu'ils ne mettent pas en garde contre leur violation – lors de la conception, du développement et du déploiement des systèmes d'IA, il est urgent de se focaliser davantage sur les incidences de l'IA sur les droits de l'homme.
- Les pays membres du Conseil de l'Europe sont bien placés pour orienter la gouvernance internationale de l'IA vers la promotion des droits de l'homme.

INTRODUCTION

L'intelligence artificielle (IA) est la conception et le développement de systèmes informatiques capables d'exécuter des tâches nécessitant normalement une intelligence humaine. En règle générale, les systèmes informatiques sont réputés intelligents (c'est pour cela qu'ils sont appelés « agents intelligents » ou « machines intelligentes ») dès lors qu'ils sont capables de percevoir leur environnement et d'agir de façon autonome pour atteindre un but. Si depuis l'antiquité la littérature scientifique et philosophique regorge de réflexions d'ordre général sur la logique mécanique, le domaine de l'IA stricto sensu est né dans les années 1940, dans le prolongement des progrès réalisés simultanément en logique mathématique (par ex. la thèse Church-Turing), en théorie de l'information, en neurobiologie et en cybernétique. Le domaine de l'IA englobe tout un éventail de démarches informatiques complexes permettant d'accomplir ou d'imiter des fonctions cognitives telles que l'apprentissage, la mémoire, le raisonnement, la vision et le traitement du langage naturel. La plus courante de ces démarches – l'apprentissage automatique – permet, grâce à des algorithmes développés à cet effet, de réaliser des tâches sans instructions explicites d'opérateurs humains. Contrairement aux programmes informatiques conventionnels, les algorithmes d'apprentissage automatique créent des modèles mathématiques sur la base des données d'apprentissage et s'appuient exclusivement sur l'inférence et l'identification de modèles pour faire des prédictions et prendre des décisions en toute autonomie[1]. Aujourd'hui, l'IA est l'un des principaux catalyseurs de transformation technologique. En ce début des années 2020, les systèmes d'IA sont intégrés à un nombre incalculable de systèmes et de dispositifs dont les humains se servent régulièrement : les téléphones mobiles, les médias sociaux, les voitures, les avions, les logiciels d'analyse, les systèmes de communication par courrier électronique, les appareils électroménagers, etc. L'IA fait partie intégrante d'un vaste éventail d'activités humaines, dont, notamment, les télécommunications, les transports, la fabrication, la santé, la banque, les assurances, le maintien de l'ordre et l'armée.

De par sa nouveauté technologique et sa capacité d'agir et de tendre vers un but de façon autonome, l'IA a le potentiel de transformer les sociétés humaines à un rythme plus élevé et de façon plus marquée que n'importe quelle autre technologie. Le potentiel transformateur de l'IA a été qualifié de « révolutionnaire » par des experts[2], et divers auteurs ont comparé le développement de l'IA à une « révolution en marche » qui « transformera presque tous les métiers »[3]. C'est pourquoi il est capital et urgent d'évaluer les incidences de l'IA sur les valeurs et principes fondamentaux de la vie humaine, l'avenir des sociétés humaines et les systèmes régissant celles-ci, à commencer par la démocratie et l'État de droit[4-6].

Ces dernières années, plusieurs organismes gouvernementaux et intergouvernementaux aussi bien que des acteurs non étatiques ont établi des principes et diffusé des lignes directrices, des recommandations, des cadres de gouvernance et d'autres instruments non contraignants en matière d'IA. Les instruments de droit souple sont des textes normatifs qui ne sont pas juridiquement contraignants ou n'ont aucune force exécutoire mais dont la nature persuasive peut avoir une influence concrète sur la prise de décision, comparable à celle des actes contraignants (droit dur). Ils ont pour but d'orienter le développement de l'IA en faveur du bien-être social et du respect des valeurs éthiques ainsi que des normes juridiques. Cependant, malgré un consensus manifeste sur le fait que l'IA doit être « éthique », il y a débat sur ce qu'il faut entendre par « IA éthique » et sur les exigences éthiques, les normes techniques et les meilleures pratiques nécessaires à cet effet. Par ailleurs, vu la prolifération rapide des textes de droit souple relatifs à l'IA et la grande diversité des auteurs, il est difficile de suivre de façon approfondie et rigoureuse l'évolution constante de ce train de mesures non contraignantes et d'en comprendre tous les tenants et aboutissants.

Le présent rapport dresse une cartographie complète du corpus actuel de principes et de lignes directrices au sujet de l'IA éthique, et en fait une méta-analyse. Celle-ci a pour but d'informer les scientifiques, les établissements de recherche, les organismes de financement, les organismes gouvernementaux et intergouvernementaux, et autres acteurs concernés par les progrès de l'innovation éthiquement responsable en matière d'IA. Elle s'intéresse en outre à la façon dont ces principes éthiques, coutumes morales et pratiques sociales recommandées peuvent se traduire en règles de gouvernance obligatoires, notamment en instruments juridiques contraignants à l'échelon

international.

Le présent rapport accorde une large place à l'examen du lien entre gouvernance de l'IA et droits de l'homme, et il évalue les répercussions que la technologie de l'IA pourrait avoir sur les droits de l'homme et les libertés[7-9]. Les droits de l'homme sont des droits inhérents à la personne, sans distinction de race, de sexe, de nationalité, d'origine ethnique, de langue, de religion ou de toute autre condition[10]. Ils renvoient aussi bien à des principes moraux qu'à des normes juridiques du droit national et international dont une personne peut se prévaloir par essence, en tant qu'être humain. Ils sont par conséquent inaliénables, inviolables et universels. Ils sont inaliénables car personne ne peut en priver qui que ce soit, inviolables, car personne ne devrait en aucun cas les enfreindre, et universels car ils s'appliquent partout et à tout moment. Les droits de l'homme et les libertés sont protégés par des conventions internationales. L'un des instruments fondamentaux de l'espace européen est la Convention européenne des droits de l'homme (CEDH), qui a été rédigée en 1950 par le Conseil de l'Europe et qui est entrée en vigueur le 3 septembre 1953. La CEDH consacre une série de droits et libertés fondamentaux qu'il faut protéger, et ses parties s'engagent juridiquement à se conformer à des normes de conduite respectueuses de ces droits et libertés[11].

MÉTHODOLOGIE

En février 2020, nous avons mené un examen de la portée du corpus de textes de droit souple existant au sujet de l'IA. Les méthodes employées à cet effet permettent de faire une synthèse et de cartographier la documentation existant dans tel ou tel domaine en suivant une démarche exploratoire, et sont donc particulièrement adaptées au criblage et à l'évaluation de domaines de recherche complexes et hétérogènes. Comme il n'existait pas de base de données unifiée des instruments non contraignants, nous avons défini un protocole de recherche et de sélection en suivant les normes PRISMA (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*). Le protocole, qui a été testé et calibré avant la collecte de données, comprend trois phases séquentielles liées par itération : criblage, évaluation de la pertinence et analyse du contenu. Cette méthodologie visait à établir une procédure formelle, fondée sur les résultats, pour cartographier, étudier et évaluer de façon itérative les mesures de gouvernance non contraignantes prises dans le domaine de l'IA.

1.1. Phase 1 : criblage

Durant la phase de criblage, nous avons procédé à l'examen rétrospectif des répertoires existants dans le cadre d'une recherche intentionnelle et non structurée du web. Nous avons commencé par examiner les quatre répertoires de données et sources textuelles ci-après pour en extraire les entrées relatives aux documents de droit souple sur l'IA :

- *AI Policy Initiatives List*, de l'Agence des droits fondamentaux de l'Union européenne (11 décembre 2019).
- Fjeld et Nagy (janvier 2020), *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI*, *Harvard Berkman Klein Center Research Publication*.
- Jobin, Ienca et Vayena (septembre 2019). *The Global Landscape of AI Ethics Guidelines. Nature Machine Intelligence*.
- Wong et CASBS (juin 2019). *Fluxus Landscape: An Expansive View of AI Ethics and Governance.*

Conformément au cadre qu'Arksey et O'Malley ont défini pour les examens de la portée[12], la recherche structurée dans les bases de données a été complétée par une recherche non structurée dans la littérature grise pour ne rater aucun instrument non contraignant. Nous avons ensuite vérifié si les éléments trouvés étaient pertinents (voir 1.2) et, lorsque c'était confirmé, nous les avons inclus manuellement dans la synthèse finale et sélectionnés pour la seconde phase.

Nous avons ensuite étudié la liste des 45 plus grandes entreprises du secteur de l'IA, dressée en mai 2019 par Datamation (*« Top-45 AI companies »*), un magazine américain de science informatique spécialisé dans l'analyse des technologies. Pour chacune de ces 45 entreprises, nous avons examiné leurs sites web et cherché manuellement et par mots-clés des déclarations au sujet de l'éthique ou des orientations générales en matière d'IA. Pour terminer, nous avons procédé à une recherche non structurée sur le web pour extraire des informations que notre stratégie de recherche n'avait peut-être pas décelées. Les éléments pertinents ont été inclus manuellement dans la synthèse finale et sélectionnés pour la seconde phase.

1.2. Phase 2 : évaluation de la pertinence

Au cours de cette phase, nous avons examiné de façon approfondie tous les éléments extraits pour vérifier s'il était pertinent de les inclure dans la synthèse finale. Les décisions en la matière ont été guidées par les critères d'inclusion et d'exclusion énumérés dans le tableau 1.

Criblage	
	-Types : sites web, articles et autres documents écrits publiés en ligne ou éléments les composant, comme des pages web spécifiques, des articles de blog, des rapports et déclarations de diverses institutions ainsi que les références sur lesquelles ils s'appuient.
Sources incluses :	-Émetteurs : organismes à but lucratif du secteur privé (entreprises, multinationales, holdings, etc. et alliances du secteur privé) ; établissements universitaires et de recherche (universités, associations professionnelles, fondations scientifiques, etc.) ; organismes gouvernementaux nationaux (ministères, autorités de protection des données, autorités de réglementation de la concurrence, etc.) ; organisations non gouvernementales (organisations à but non lucratif, organisations caritatives, etc.).
	-Langues : anglais, allemand, français, espagnol, néerlandais, italien et grec (langues parlées par les chercheurs).
Sources exclues :	-Types : vidéos, images et audio/podcasts (sauf descriptions écrites), livres, articles de journaux, articles universitaires, plans de cours, législation, normes officielles, synthèses de conférences.
	-Émetteurs : organisations intergouvernementales et supranationales.
	-Langues : autres que celles qui sont susmentionnées.
Critères de sélection	
Sources incluses :	-font référence aux termes « intelligence artificielle » et/ou « IA » de façon explicite, dans leur titre ou dans le corps du document (par exemple : Google, « principes de l'IA ») ; ou

	-ne contiennent pas les termes susmentionnés mais « robot », « robotique », « mégadonnées » (« big data »), « apprentissage automatique » et font explicitement référence à l'IA ou l'intelligence artificielle comme faisant partie intégrante des robots et/ou de la robotique ; ou
	- ne contiennent pas les termes susmentionnés dans leur titre mais des termes en lien (« algorithmes », « analyse prédictive », « informatique cognitive », « apprentissage automatique », « mégadonnées » (« big data »), « apprentissage profond », « autonome » ou « automatisé »).
	ET
	-décrivent un principe, des lignes directrices, une norme (notamment avec les termes « éthique », « principes », « préceptes », « déclaration », « politiques », « lignes directrices », « valeurs », etc.), une stratégie interne (par ex. création d'un conseil consultatif) ou autre type d'initiative.
	ET
	- sont exprimées dans un langage normatif ou impératif (c'est-à-dire avec l'emploi de modaux ou de termes indiquant une obligation, par ex. « responsable », « juste », « confiance/fiable », etc.).
	-reposent sur des principes ou des valeurs (c.-à-d. indiquent une préférence et/ou une adhésion à une certaine vision éthique ou ligne de conduite).
	-font référence à des actions/visions/engagements/lignes d'action s'appliquant à l'acteur qui les énonce ou à d'autres acteurs du secteur privé.
Sources exclues :	-les sites web et les documents concernant la robotique n'indiquant pas que l'IA fait partie intégrante des robots/de la robotique.
	-les sites web et documents concernant les données ou l'éthique des données n'indiquant pas que l'IA fait partie intégrante des données.
	-les sites web et documents sur l'éthique de l'IA s'adressant directement à des acteurs n'appartenant pas au secteur privé (par ex. services de conseil pour le secteur public).
	-les sites web et documents sur l'éthique dont le sujet principal n'est pas l'intelligence artificielle (par ex. : l'éthique dans les affaires).

Tableau 1 – Critères de sélection

1.3. Analyse du contenu

Dans la seconde phase, les éléments inclus dans la synthèse finale ont été examinés à l'aide de la version étendue d'un protocole d'analyse de contenu précédemment validé, que les auteurs avaient mis au point[13,14]. Ce protocole prévoit une analyse quantitative et une analyse qualitative. Sur le plan quantitatif, les éléments ont été classés en fonction du type d'instrument, de l'émetteur et de l'emplacement géographique de ce dernier. Par ailleurs, la fréquence relative des données quantitatives pertinentes a été mesurée et illustrée par une représentation graphique. Enfin, nous avons procédé à l'examen de l'intégralité des textes à l'aide d'une recherche par mots-clés pour repérer les documents qui faisaient explicitement référence aux droits de l'homme. Les documents relevant de cette catégorie faisaient explicitement référence soit à la protection et à la promotion des droits de l'homme, soit à la prévention de leur violation au moment de la conception, du développement ou du déploiement des applications d'IA. Ils ont été séparés de ceux qui n'évoquaient pas les droits de l'homme ou le faisaient mais sans formuler de déclaration normative explicite sur leur promotion ou leur non-violation dans le contexte de l'IA.

Sur le plan qualitatif, une analyse du contenu thématique a été menée pour repérer des thèmes récurrents concernant les domaines suivants : i) les principes et valeurs éthiques, et ii) les droits de l'homme. Cette analyse a été réalisée manuellement par les chercheurs, à l'aide d'un logiciel d'analyse de données qualitatives (NVivo/MAXQDA pour Mac). Les thèmes émergents ont été analysés en profondeur, encodés et regroupés dans des catégories éthiques prédéfinies sur la base de la matrice d'éthique mise au point par Jobin, Ienca et Vayena (2019).

Vu la taille de la base de données de la synthèse finale, cette analyse thématique manuelle a été réalisée à l'aide d'un moteur de traitement du langage naturel (TLN). Nous avons extrait des documents et du contenu web automatiquement là où c'était possible, à l'aide du paquet Python wget, et ajouté le reste manuellement. Nous avons ensuite créé des expressions régulières[15] à partir des codes résultant du protocole d'analyse qualitatif créé par Jobin, Ienca et Vayena[13]. Les expressions régulières des codes appartenant à un même thème ont été reliées entre elles en une seule expression régulière par « or » ('|'). Par souci d'exhaustivité et d'inclusion, les codes anglais d'origine ont été traduits dans les langues suivantes : allemand, français, espagnol, italien et néerlandais. Pour déterminer si un thème était traité, nous avons cherché des occurrences de ses expressions régulières dans les documents, et nous avons considéré que s'il y en avait au moins une, cela signifiait que le thème était bien évoqué. Enfin, nous avons regroupé les résultats par catégorie de membre et nous les avons normalisés en fonction du nombre total de lignes directrices dans chaque groupe. Les principes et les valeurs éthiques évoqués par les 47 États membres du Conseil de l'Europe ont été comparés aux principes et valeurs évoqués par les États observateurs ainsi que par les pays du reste du monde.

1.4. Analyse de l'éthique normative et des politiques

Dans la quatrième et dernière phase, nous avons procédé à l'analyse empirique de l'éthique normative et des politiques afin de traduire en normes prescriptives les constatations descriptives faites lors des précédentes phases de l'étude. À cet effet, nous avons suivi trois étapes théoriques successives. Premièrement, nous avons examiné les résultats de notre analyse de contenu pour déterminer quels principes et valeurs éthiques revenaient le plus souvent dans le corpus des documents traités. Étant donné que de précédentes recherches avaient montré d'importantes variations dans l'interprétation des groupes thématiques récurrents[13], nous avons complété l'évaluation de la fréquence relative d'occurrence des divers thèmes par une évaluation détaillée de l'interprétation qui en était faite. Cette dernière a largement contribué à déterminer quelles interprétations des principes considérés étaient les plus effectives et devaient donc être adoptées et reprises par l'ensemble des acteurs. Deuxièmement, nous avons analysé les données recueillies pour déterminer quels principes et valeurs étaient moins souvent traités voire absents des lignes directrices actuelles sur l'éthique en matière d'IA. Cette étape a aidé à repérer d'éventuelles lacunes dans les initiatives internationales de droit souple et, par conséquent, à énoncer des recommandations normatives sur la façon de combler ces lacunes éthiques. Troisième et dernière étape, nous avons formulé des recommandations normatives au sujet des valeurs et principes

éthiques fondamentaux à traiter en priorité dans le cadre de la gouvernance internationale de l'IA. Cette dernière étape a permis d'étayer les futurs cadres normatifs de l'éthique et de définir une feuille de route pour l'élaboration des politiques internationales sur l'IA, l'éthique et les droits de l'homme. À cet effet, nous avons réalisé une synthèse graphique des résultats de l'étude facile à exploiter, et créé une boîte à outils qui permettra d'assurer un suivi et une évaluation (indicateurs) à l'interface entre IA, éthique et droits de l'homme.

2. Constatations

Notre recherche a permis de trouver 116 documents contenant des instruments de droit souple sur l'IA, diffusés par des organismes non intergouvernementaux jusqu'en février 2020. Il ressort des données que depuis 2016, le nombre de publications a enregistré une hausse de 93,9 %, ce qui est considérable. Le nombre de documents de droit souple publiés à l'échelon international a culminé en 2018 puis enregistré une baisse non négligeable l'année suivante (voir figure 1).

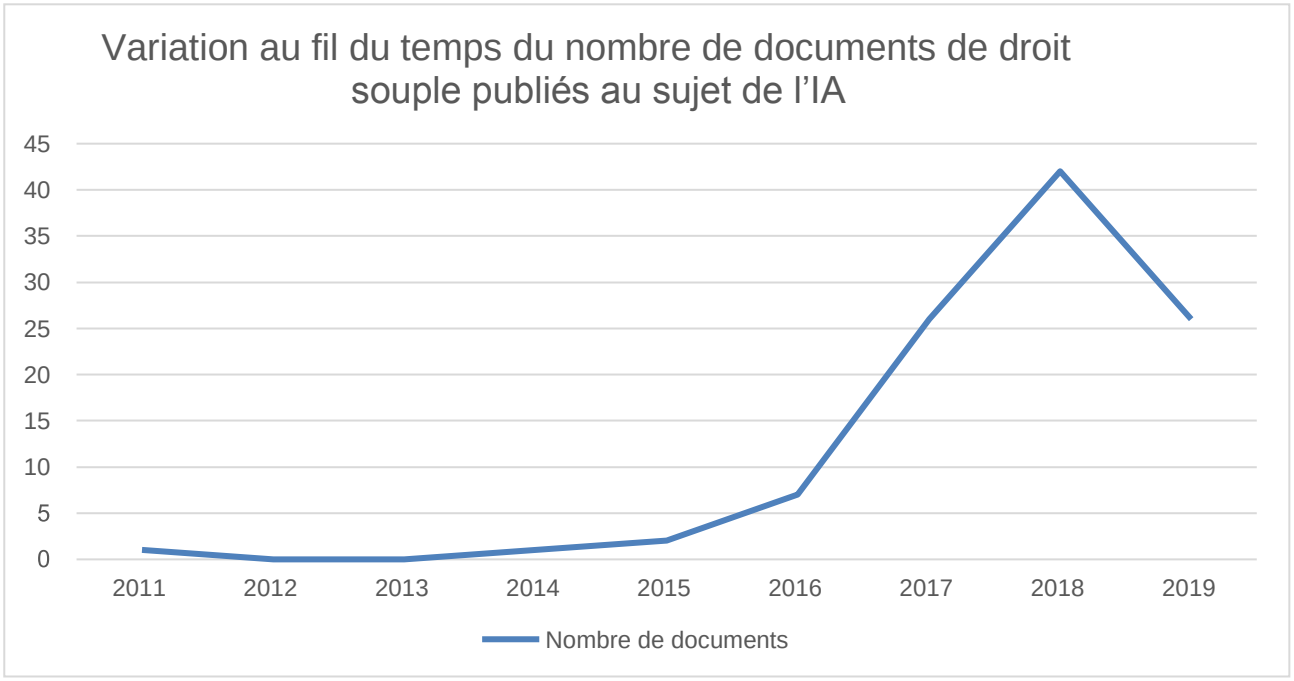

Figure 1 – Variation au fil du temps du nombre de documents de droit souple publiés au sujet de l'IA

Il ressort de la ventilation des données par type d'organisme émetteur que la plupart des documents ont été émis par des autorités publiques (n=39), puis des entreprises et des alliances du secteur privé (n=36), puis des établissements universitaires et de recherche, notamment des fondations scientifiques, des associations professionnelles et des alliances de recherche (n=28) ainsi que par des organisations non gouvernementales (ONG) et notamment des organismes à but non lucratif et des organisations caritatives (n=13). La figure 2 montre la répartition détaillée des organismes émetteurs.

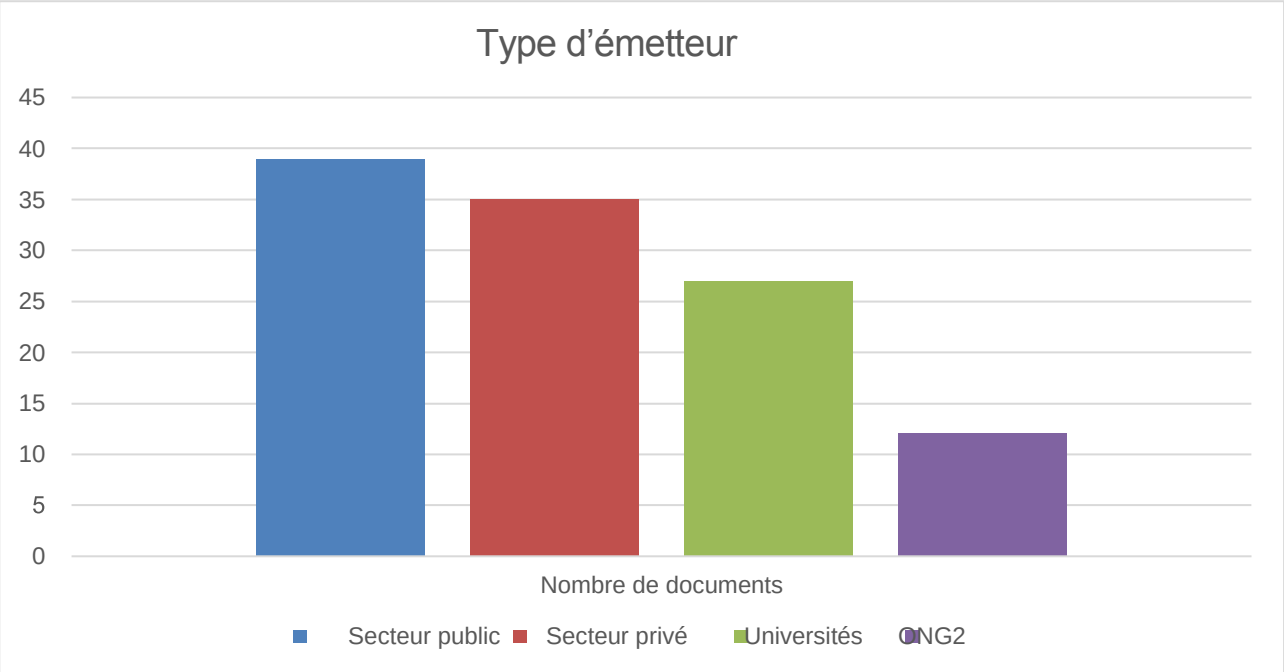

Figure 2 – Types d'organismes émetteurs

La ventilation des données en fonction de la répartition géographique des organismes émetteurs montre que 46 % (n=53,5) des documents de droit souple sont diffusés par des organisations basées dans des pays membres du Conseil de l'Europe ; 32 % (n=37,5) par des organisations basées dans des pays ayant le statut d'observateur auprès du Conseil de l'Europe ; 21 % (n=25) par des organisations basées dans des pays qui ne sont ni membres du Conseil de l'Europe ni observateurs. Dans l'ensemble, il ressort des données recueillies que les organismes émetteurs situés dans des pays économiquement développés prédominent, les États-Unis (*n*=29,5 ; 25,2 %) et le Royaume-Uni (*n*=17,5 ; 16 %) totalisant à eux deux plus d'un tiers de tous les principes éthiques énoncés en matière d'IA. Les autres pays sont, par ordre décroissant, l'Allemagne (n=8), le Japon (n=6), la Finlande (n=4), la Belgique, la Chine, la France et les Pays-Bas (n=3), l'Inde, l'Italie, Singapour et l'Espagne (n=2), l'Australie, l'Autriche, la République tchèque, l'Islande, la Lituanie, Malte, le Mexique, la Nouvelle-Zélande, la Norvège, la Russie, la Corée du Sud, la Suède, la Suisse, les EAU et le Vatican (n=1). Treize documents ont été diffusés par des organisations internationales ou des organisations ne pouvant pas être rattachées à un pays précis. Les pays d'Afrique et d'Amérique du Sud n'étaient pas représentés indépendamment d'organisations internationales. La figure 3 donne un aperçu de la répartition géographique des organismes émetteurs.

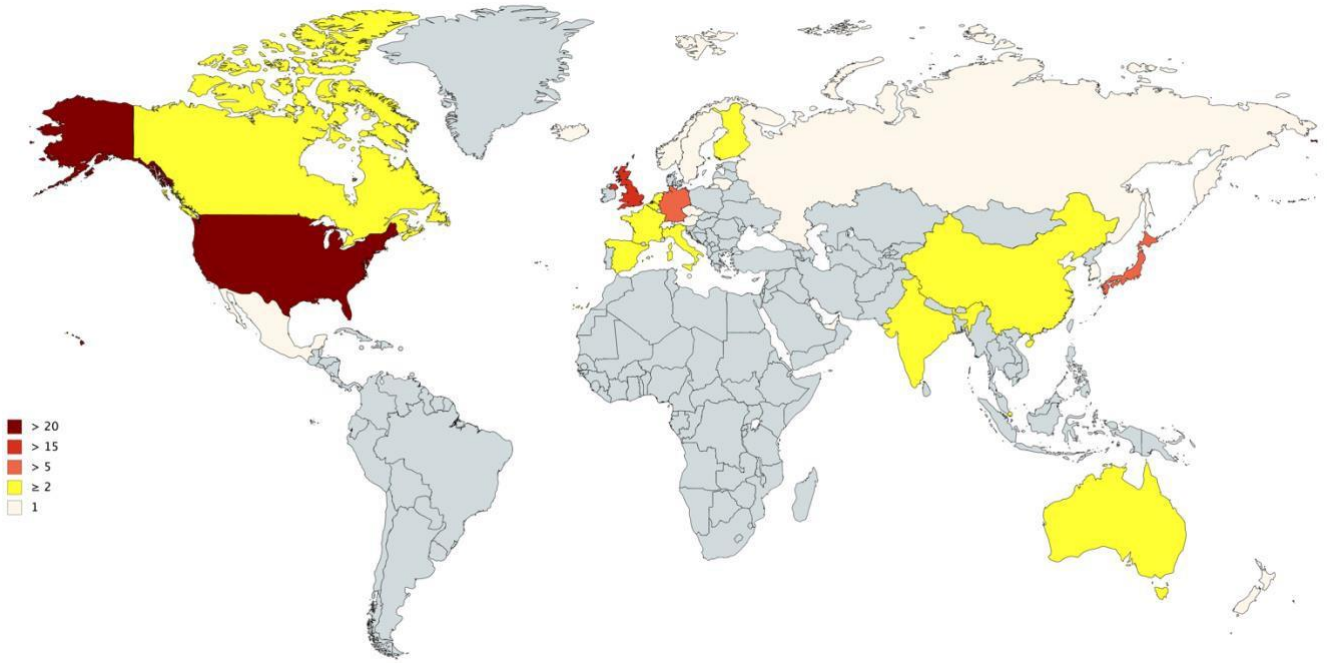

Figure 3 – Répartition géographique des documents de droit souple en fonction du pays d'appartenance de l'organisme émetteur

Plus de la moitié des documents (n=62) font explicitement référence à la promotion et au respect des droits de l'homme ou à la prévention de leur violation. Trente et un de ces documents ont été diffusés par des organismes basés dans des pays membres du Conseil de l'Europe, 14 par des organismes basés dans des pays observateurs et 17 par des organismes basés dans des pays qui ne sont ni l'un ni l'autre. Les documents émis par des organismes basés dans des pays membres du Conseil de l'Europe évoquent les droits de l'homme dans 57,9 % des cas. Ceux qui proviennent de pays non membres du Conseil de l'Europe y font référence dans 49,6 % des cas. Cela montre que les incidences de l'intelligence artificielle sur les droits de l'homme sont plus fréquemment examinées par des organismes basés dans des pays membres du Conseil de l'Europe que dans des pays du reste du monde.

Notre analyse du contenu thématique nous a permis d'extraire divers codes se rapportant à l'éthique, susceptibles d'être tous systématiquement applicables aux onze groupes de principes éthiques définis par Jobin, Ienca et Vayena (2019)[13]. Ce sont, par ordre décroissant de fréquence des sources dans lesquels ils apparaissent : la transparence, la justice et la loyauté, la non-malfaisance, la responsabilité, le respect de la vie privée, la bienfaisance, la liberté et l'autonomie, la confiance, la dignité, la durabilité et la solidarité. Le tableau 2 illustre avec précision la fréquence avec laquelle les principes éthiques et les codes y associés sont évoqués.

Principes éthiques	Nombre de documents	Codes associés
Transparence	101/116	Transparence, explicabilité, compréhensibilité, interprétabilité, communication, divulgation, montrer
Justice et loyauté	97/116	Justice, loyauté, cohérence, inclusion, égalité, équité, (im)partialité, (non-)discrimination, diversité, pluralité, accessibilité, réversibilité, recours, réparation, défi, accès et distribution, impartialité
Non-malfaisance	84/116	Non-malfaisance, sécurité, sûreté, préjudice, protection, précaution, prévention, intégrité (physique et mentale), non-subversion
Responsabilité	79/116	Responsabilité, obligation de rendre compte, obligation, agir avec intégrité
Respect de la vie privée	74/116	Respect de la vie privée, informations personnelles ou privées, confidentialité
Bienfaisance	58/116	Avantages, bienfaisance, bien-être, paix, bien-être social, bien commun
Liberté et autonomie	48/116	Liberté, autonomie, consentement, choix, autodétermination, autonomisation
Fiabilité	41/116	Confiance, fiabilité
Durabilité	20/116	Durabilité, environnement (nature), énergie, ressources (énergie)
Dignité	20/116	Dignité
Solidarité	10/116	Solidarité, sécurité sociale, cohésion

Tableau 2 – Fréquence des thèmes éthiques et codes y associés

Aucun des différents principes éthiques n'est commun à tous les documents du corpus mais une convergence se dégage autour des principes suivants : transparence, justice et loyauté, non-malfaisance, responsabilité et respect de la vie privée. Près de deux tiers de toutes les sources y font référence. Toutefois, l'analyse thématique met en évidence la persistance de nettes divergences sémantiques et conceptuelles, à la fois quant à l'interprétation des onze principes éthiques et aux recommandations spécifiques ou aux questions préoccupantes qui en découlent.

Les expressions régulières tirées des codes montrent des différences perceptibles entre les thèmes abordés dans les documents issus des pays membres du Conseil de l'Europe et ceux dont traitent les documents élaborés ailleurs. Par rapport aux documents élaborés dans les pays ayant le statut d'observateur auprès du Conseil de l'Europe, les textes de droit souple établis au sein des pays membres du Conseil de l'Europe semblent insister sur les principes éthiques ci-après : la transparence, la durabilité la liberté et l'autonomie, la confiance/fiabilité et la solidarité (voir figure 4). Ils semblent en revanche faire moins souvent référence aux principes de justice, de bienfaisance et de dignité. Par rapport aux textes de droit souple établis dans le reste du monde (pays non membres et non observateurs), ceux qui sont établis dans les pays membres du Conseil de l'Europe paraissent mettre l'accent sur les principes de confiance/fiabilité et de solidarité et traiter tous les autres principes moins fréquemment. Les principes de respect de la vie privée, de justice et de loyauté sont ceux pour lesquels les différences sont les moins marquées, et donc ceux qui ont le plus haut degré de constance transgéographique et transculturelle.

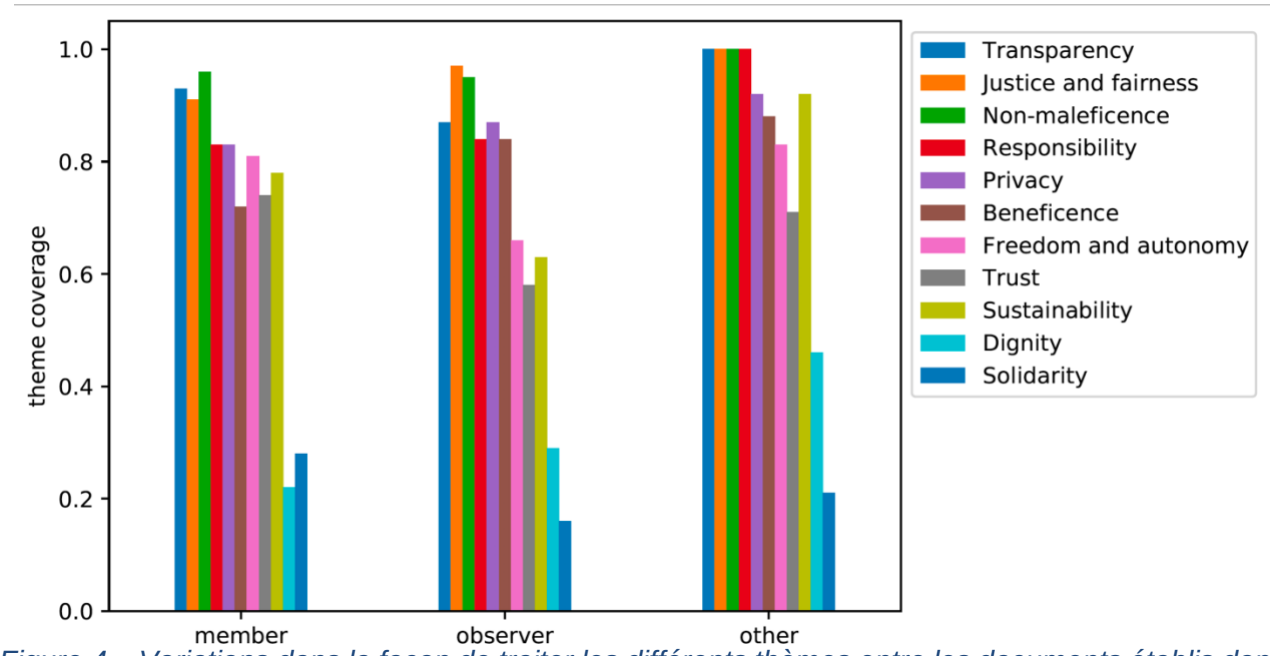

Figure 4 – Variations dans la façon de traiter les différents thèmes entre les documents établis dans des pays membres du Conseil de l'Europe et les documents établis dans le reste du monde.

Légende :
Thèmes traités
Transparence
Justice et loyauté
Non-malfaisance
Responsabilité
Respect de la vie privée
Bienfaisance
Liberté et autonomie
Confiance
Durabilité
Dignité
Solidarité
Membre Observateur Autres

Voici une évaluation détaillée de chacun des thèmes énumérés plus haut.

Transparence : La transparence, qui est évoquée dans 101 des 116 documents examinés, est le principe éthique le plus fréquemment traité dans les textes non contraignants actuels. L'analyse thématique montre que la façon dont elle est interprétée et les raisons qui la justifient varient considérablement, ce qui, constate-t-on, entraîne d'indéniables divergences dans les stratégies de mise en œuvre proposées pour parvenir à la transparence en matière d'IA. Les références faites à la transparence peuvent être rangées dans deux grandes catégories thématiques : 1) la transparence des algorithmes et des méthodes de traitement des données, 2) la transparence des pratiques humaines relatives à la conception, au développement et au déploiement des systèmes d'IA. La première catégorie implique le plus souvent la promotion des démarches méthodologiques en faveur de « l'explicabilité de l'IA », à savoir le fait que les produits et les décisions des systèmes d'IA soient compréhensibles par les experts humains. Ces méthodes et techniques sont en opposition avec le mécanisme de « boîte noire » de l'apprentissage automatique, où les étapes par lesquelles un système d'IA arrive à telle ou telle décision sont inintelligibles pour les experts humains, en ce compris les concepteurs du système. Alors que les entreprises du secteur privé, surtout les acteurs de l'IA, ont tendance à réduire la transparence à l'interprétabilité et à l'explicabilité moyennant diverses solutions techniques – par exemple, la méthode LRP (*layerwise relevance propagation*) et l'interprétabilité locale – les autorités publiques, telles que les responsables nationaux de la protection des données, soulignent l'importance des méthodes de surveillance comme les audits. La deuxième forme de transparence ne porte pas sur des algorithmes interprétables mais sur la transparence des pratiques humaines relatives aux données et à l'IA, par exemple divulguer des informations pertinentes aux personnes concernées, éviter l'opacité lors du déploiement des stratégies d'IA, et interdire les conflits d'intérêts entre les acteurs de l'IA et les organes de surveillance. Ce type de transparence est plus couramment évoqué par les acteurs publics et les ONG.

Justice, loyauté, équité : Il est essentiellement fait référence à la justice en termes de loyauté et de prévention (ou d'atténuation) des biais algorithmiques qui peuvent conduire à de la discrimination. La crainte de voir l'IA accroître les inégalités et être à l'origine de discriminations semble moins répandue dans les documents de droit souple émanant du secteur privé que dans ceux qui émanent des organismes publics et du milieu universitaire. Les textes sont divisés sur la question de savoir comment assurer la justice et la loyauté en matière d'IA. Certaines sources sont axées sur le respect de la diversité et favorisent l'inclusion et l'égalité, aussi bien dans la conception des systèmes d'IA (en particulier lors de la compilation des jeux de données de l'apprentissage) que dans leur déploiement dans la société. D'autres sources, en s'appuyant sur le droit de recours et de réparation, préconisent d'offrir la possibilité de contester les décisions. L'accès équitable aux avantages que procure l'IA est aussi un thème récurrent. Les documents émanant d'acteurs publics insistent particulièrement sur les répercussions de l'IA sur le marché du travail, et sur la nécessité d'examiner les défis démocratiques et sociétaux. Nous avons repéré cinq grandes stratégies de mise en œuvre, conciliables entre elles, permettant de préserver et de promouvoir la justice et la loyauté dans l'IA :

I. appliquer des solutions techniques telles que des normes et des bonnes pratiques ;
II. sensibiliser le public aux droits et à la réglementation en vigueur ;
III. améliorer les tests, le suivi et le contrôle des systèmes d'IA ;
IV. développer ou renforcer l'État de droit et instaurer un droit de recours et de réparation ;
V. modifier le système et introduire des processus, par exemple des mesures ou des contrôles gouvernementaux, renforcer la pluridisciplinarité de la main-d'œuvre, et améliorer l'inclusion de la société civile ou de divers autres acteurs de façon interactive.

Si les solutions II à V semblent être privilégiées par les autorités publiques (notamment les responsables de la protection des données), la première est privilégiée par les acteurs du secteur privé de l'IA.

Non-malfaisance : Les références à la non-malfaisance, bien plus fréquentes que les références à la bienfaisance, sont liées à des recommandations générales de sûreté et de sécurité ou à des déclarations selon lesquelles l'IA ne devrait jamais causer de préjudices prévisibles ou involontaires. Certains documents se concentrent sur des risques précis ou des préjudices potentiels, en particulier le risque de détournement intentionnel dans le cadre d'une cyberguerre et d'un piratage. Les sources de préjudice les plus courantes que citent les divers documents sont la discrimination sociale, la violation de la vie privée et les préjudices corporels ou psychologiques. Les documents de droit souple consacrés à l'atténuation des préjudices recommandent souvent des solutions techniques et des mesures de gouvernance obligatoires relatives à la recherche en matière d'IA, à la conception ainsi qu'au développement et au déploiement technologiques. Les solutions techniques passent entre autres par des évaluations de la qualité des données intégrées, ou des systèmes de sécurité et de respect de la vie privée inclus dans le cadre de conception, tandis que d'autres documents privilégient la création de normes sectorielles. Les stratégies de gouvernance proposées prévoient notamment une coopération active entre les diverses disciplines et parties prenantes, le respect de la législation en vigueur ou à venir, et la nécessité de mettre en place des processus et des pratiques de surveillance, notamment grâce à des tests, à un suivi, à des contrôles et à des évaluations effectués par des services internes, les clients, les utilisateurs, des tierces parties indépendantes ou des autorités publiques. Certaines sources désignent expressément l'usage militaire de l'IA – le problème des technologies dites à double usage – comme étant le principal domaine dans lequel des mesures de gouvernance doivent être prises.

Responsabilité et obligation de rendre compte : Si de nombreux textes évoquent le développement d'une « IA responsable », il n'en reste pas moins que les notions de responsabilité et d'obligation de rendre compte sont rarement définies. Divers acteurs sont désignés comme ayant une responsabilité et des comptes à rendre pour les actions et décisions de l'IA. Il s'agit des développeurs et concepteurs de l'IA et du secteur de l'IA dans son intégralité. Les documents ne s'accordent pas sur la question de savoir si la responsabilité de l'IA doit être calquée sur celle des humains ou si les humains doivent toujours être les seuls responsables en dernier ressort des actions des objets technologiques.

Respect de la vie privée : Le respect de la vie privée est généralement considéré comme une valeur à défendre et un droit à protéger. Si les questions de respect de la vie privée sont fréquemment abordées dans les lignes directrices existantes sur l'IA, aucun consensus ne se dégage sur les difficultés spécifiques que pourraient soulever les progrès en matière d'IA par rapport aux autres technologies utilisant de gros volumes de données. L'analyse thématique montre que la plupart des documents évoquent le respect de la vie privée en termes généraux, sans relier explicitement les capacités de l'IA et les nouveaux défis en la matière. Bien qu'il soit mal décrit, le problème de respect de la vie privée que pose l'IA est souvent abordé en lien avec les questions de protection et de sécurité des données. Les stratégies proposées pour protéger la vie privée face à l'IA peuvent être rangées en trois catégories : A) les solutions techniques telles que la confidentialité différentielle, le calcul multipartite sécurisé et le chiffrement homomorphique ; B) les solutions consistant à informer le public, comme les campagnes de sensibilisation des utilisateurs et des personnes concernées par les données ; et C) les solutions réglementaires, consistant par exemple à mieux définir les règles juridiques à respecter (notamment en matière de protection des données) ou à rédiger de nouveaux textes législatifs ou réglementaires portant spécifiquement sur l'IA.

Bienfaisance : Si la pratique du bien (la *bienfaisance*, en termes éthiques) est souvent évoquée, elle est rarement définie, à l'exception de quelques documents encourageant l'adoption de mesures en faveur du bien-être et de l'épanouissement humain, de la paix et du bonheur, de la création d'opportunités socio-économiques et de la prospérité économique. Il n'est pas non plus clair de savoir quels acteurs doivent bénéficier de l'IA : en règle générale les émetteurs du secteur privé mettent en avant leurs clients alors que les sources universitaires et publiques affirment le plus souvent que l'IA doit profiter à « tout le monde », à « l'humanité » et à « la société au sens large ». Les stratégies de bienfaisance prévoient entre autres d'aligner l'IA sur les valeurs humaines, de limiter la concentration du pouvoir, et d'employer les capacités de l'IA pour promouvoir les droits de l'homme.

Liberté et autonomie : Les documents de droit souple établissent un lien entre l'IA et la protection ou la promotion de plusieurs libertés, à savoir, notamment, la liberté d'expression, l'autodétermination informationnelle, le droit au respect de la vie privée et l'autonomie personnelle. Cette dernière notion correspond en règle générale à une liberté positive, et plus précisément à la liberté de s'épanouir et au libre arbitre. Un petit nombre de documents assimilent toutefois l'autonomie à une liberté négative, par exemple la liberté face à l'expérimentation technologique, à la manipulation ou à la surveillance. Les solutions proposées pour protéger la liberté et l'autonomie en matière d'IA consistent notamment à favoriser la transparence et l'explicabilité, à renforcer la maîtrise de l'IA, à veiller à ce que les personnes concernées donnent leur consentement éclairé ou, à l'inverse, à s'abstenir de recueillir et de divulguer des données si les personnes concernées n'y ont pas consenti en toute connaissance de cause.

Confiance et fiabilité : Un peu plus d'un tiers des documents de droit souple prônent la fiabilité des recherches et de la technologie en matière d'IA, ou la promotion d'une culture de la confiance chez les scientifiques et les ingénieurs. Certains documents mettent toutefois expressément en garde contre un excès de confiance dans l'IA et affirment que la confiance ne peut s'instaurer qu'entre pairs et ne doit pas être déléguée à l'IA. Il est notamment suggéré de s'appuyer sur les éléments ci-après pour instaurer ou maintenir la confiance : l'éducation, la fiabilité, la responsabilité, des processus permettant d'assurer le suivi et l'évaluation de l'intégrité des systèmes d'IA au fil du temps, et des outils et des techniques pour veiller au respect des règles et normes.

Durabilité : La durabilité est rarement évoquée, et lorsqu'elle l'est c'est en règle générale dans le cadre de la protection de l'environnement, voire de l'amélioration des écosystèmes et de la biodiversité de la planète. Certains documents exigent que les systèmes d'IA traitent les données de façon durable et optimisent leurs performances énergétiques pour réduire au maximum leur empreinte écologique[4,7]. Un plus petit nombre de documents sont axés sur la durabilité sociale, c'est-à-dire veiller à ce que les responsables rendent des comptes en cas de pertes d'emplois, et accroître les possibilités d'innovation.

Dignité : Si la dignité n'est pas définie dans les lignes directrices existantes, les documents de droit souple précisent que c'est une prérogative des humains et non des robots. Les références faites à la dignité sont étroitement liées à la protection et à la promotion des droits de l'homme. L'IA ne devrait pas réduire ou détruire la dignité humaine mais plutôt la respecter, la préserver, voire la renforcer. La dignité est présumée protégée si, en premier lieu, les développeurs de systèmes d'IA la respectent et si elle est favorisée par une nouvelle législation, des initiatives de gouvernance ou des lignes directrices techniques et méthodologiques édictées par les pouvoirs publics.

Solidarité : La solidarité est le thème éthique qui revient le moins souvent, et lorsqu'il est évoqué, c'est principalement en lien avec les incidences de l'IA sur le marché du travail. Les sources documentaires recommandent un renforcement des protections sociales face aux conséquences à long terme de l'IA sur le travail humain. Elles soulignent la nécessité de redistribuer les bénéfices de l'IA afin de ne pas mettre en danger la cohésion sociale[6,5] et de respecter les personnes et groupes potentiellement vulnérables. Enfin, elles mettent en garde contre les pratiques de collecte et de traitement des données axées sur les personnes et susceptibles de fragiliser la solidarité en faveur d'un « individualisme radical ».

Insuffisances

Cette étude affiche plusieurs insuffisances. Tout d'abord, du point de vue bibliographique, les lignes directrices et les documents de droit souple font partie de la littérature grise et ne sont donc pas indexés dans les bases de données conventionnelles de recherche. Leur extraction est inévitablement moins reproductible et impartiale que s'ils se trouvaient dans des bases de données systématiques de documents soumis à un examen collégial. En suivant les meilleures pratiques relatives à l'examen de la littérature grise, nous avons réussi à atténuer ce problème en créant un protocole de recherche et de sélection que nous avons testé avant de recueillir les données. Comme les résultats des recherches effectuées sur des moteurs de recherche sont personnalisés, il a fallu atténuer le risque d'influence en élargissant aussi bien les mots-clés que la sélection des résultats.

Comme le choix de l'anglais était susceptible de fausser les résultats en ne renvoyant qu'à des documents en anglais, nous avons limité le problème en incluant également des éléments rédigés dans les langues suivantes : allemand, français, italien, espagnol et néerlandais. Les mots-clés et les codes dans ces langues ont été traduits en anglais et inclus dans l'analyse. Notre analyse du contenu présente les insuffisances caractéristiques des méthodes d'analyse qualitative. Conformément aux meilleures pratiques en matière d'analyse de contenu, nous avons atténué le biais subjectif grâce à une stratégie de codage inductif appliquée séparément par deux évaluateurs. Enfin, vu le rythme soutenu auquel sont publiés les documents directifs relatifs à l'IA, de nouveaux documents d'orientation risquaient d'être diffusés après la fin de nos travaux. Aussi avons-nous surveillé en permanence la littérature jusqu'au 1er mars 2020, tout en procédant à l'analyse des données.

Débat et analyse de l'éthique normative

Nous avons constaté que le nombre et la diversité des documents de droit souple sur l'IA augmentaient rapidement, ce qui montre que la communauté internationale participe de plus en plus à l'élaboration de mesures non contraignantes portant sur ce domaine technologique. Les organismes qui établissent des lignes directrices, des principes et d'autres instruments non contraignants en matière d'IA appartiennent à des secteurs très diversifiés. Le fait que le secteur public (c.-à-d. les organismes gouvernementaux) et le secteur privé (entreprises et alliances) publient quasiment autant de documents l'un que l'autre montre que les difficultés que soulève l'IA sur le plan éthique préoccupent aussi bien les organismes publics que les entreprises privées. Il existe toutefois de nettes différences dans les solutions proposées face aux problèmes éthiques que soulève l'IA, les acteurs publics privilégiant des solutions techniques telles que l'explicabilité et l'interprétabilité de l'IA plutôt qu'une réglementation obligatoire et une réflexion éthique approfondie. Par ailleurs, le fait que certaines zones géographiques, par exemple l'Afrique et l'Amérique du Sud, soient relativement sous-représentées montre que le débat sur l'IA éthique n'a pas la même intensité partout. Les pays économiquement développés façonnent ce débat davantage que les autres, ce qui fait craindre que le savoir local, le pluralisme culturel et l'équité mondiale ne soient laissés pour compte. Ces observations confirment les inégalités de représentation et de répartition géographiques des acteurs de l'éthique de l'IA que de précédentes études avaient constatées[13]. Par rapport à celles-ci, notre rapport permet toutefois de constater que de nouveaux acteurs, issus de pays qui étaient auparavant absents du débat, prennent désormais part à l'élaboration de mesures de gouvernance non contraignantes à l'échelon international. Ils viennent de grandes puissances du secteur de l'IA, à savoir de pays leaders mondiaux dans ce secteur, comme la Chine, mais aussi de pays à revenus moyens situés dans des régions autrefois non représentées, comme la Russie et le Mexique.

La prolifération de textes non contraignants peut être interprétée comme une réponse sur le plan de la gouvernance aux recherches de pointe concernant l'IA, dont la production et le marché ont considérablement augmenté ces dernières années[16]. Notre analyse montre que les différents acteurs semblent commencer à converger vers la promotion des principes éthiques de transparence, de justice, de non-malfaisance, de responsabilité et de respect de la vie privée. Elle fait néanmoins apparaître de profondes divergences sur quatre grands points : i) la façon d'interpréter les principes éthiques, ii) la raison pour laquelle ils sont jugés importants, iii) les questions, domaines et acteurs auxquels ils se rapportent, et iv) la façon de les mettre en œuvre. Par ailleurs, on ne sait toujours pas quels principes éthiques doivent être prioritaires, comment résoudre les incompatibilités entre les principes, qui doit assurer la surveillance éthique de l'IA et de quelle façon les chercheurs et les institutions peuvent respecter les lignes directrices. Ces observations donnent à penser qu'il existe, au niveau de la formulation des principes et de leur mise en œuvre concrète, des lacunes que l'expertise technique ou des approches descendantes ne peuvent guère combler.

Bien qu'aucun des différents principes éthiques ne soit expressément approuvé par toutes les lignes directrices existantes, la transparence, la justice et la loyauté, la non-malfaisance, la responsabilité et le respect de la vie privée apparaissent tous dans plus de la moitié des documents. Cela pourrait indiquer qu'à l'échelon mondial, les politiques sur l'IA éthique sont en train de converger vers ces principes. En particulier, la fréquence des appels à la transparence, à la justice et à la loyauté montre

qu'il devient primordial, sur le plan de l'éthique, d'exiger des processus transparents à toutes les étapes de l'IA (du développement et de la conception des algorithmes jusqu'aux modalités d'utilisation de l'IA) et de mettre en garde la communauté mondiale face au risque que l'IA renforce les inégalités si les notions de justice et de loyauté ne sont pas dûment prises en considération. Ces thèmes apparaissent étroitement liés à celui de la responsabilité, car la promotion de la transparence et de la justice semble aller de pair avec le renforcement de la responsabilité et de l'obligation de rendre compte qui incombent aux fabricants d'IA et à ceux qui la déploient.

D'aucuns ont affirmé que la transparence n'est pas un principe éthique en tant que tel mais plutôt un prérequis éthique facilitant ou compromettant d'autres pratiques ou principes éthiques[17]. Cette définition de la transparence en tant que prérequis éthique à d'autres principes est décelable dans la Déclaration de conformité du fournisseur d'IBM, qui donne des informations sur les quatre grands piliers de la fiabilité de l'IA. L'hypothèse selon laquelle la transparence serait un prérequis éthique pourrait expliquer en partie pourquoi elle est plus fréquemment évoquée que d'autres principes éthiques. Il est à noter que les lignes directrices existantes accordent une grande importance à la promotion de la responsabilité et de l'obligation de rendre compte alors que peu de textes insistent sur le devoir d'agir avec intégrité qui incombe à tous les acteurs du développement et du déploiement de l'IA. Ce décalage est probablement dû au fait que, comme nous l'avons constaté, les lignes directrices existantes ne mettent pas pleinement en correspondance les principes et les conditions concrètes à remplir, plusieurs principes n'étant d'ailleurs ni définis ni reliés aux critères indispensables à leur réalisation.

Alors que les codes liés à la non-malfaisance sont plus nombreux que les codes liés à la bienfaisance, il semble qu'à l'heure actuelle, dans la communauté de l'IA, l'obligation morale de prévenir les dommages l'emporte sur la promotion des bonnes pratiques. Cela pourrait être en parti interprété comme un exemple de ce que l'on appelle un biais négatif, à savoir un biais cognitif général qui amène à donner plus de poids aux éléments négatifs[18,19], hypothèse récemment mise en avant par le psychologue cognitif Steven Pinker dans une analyse approfondie réalisée par l'Unité de la prospective scientifique (STOA) du Parlement européen[20]. Ce biais négatif est accentué par le fait que les lignes directrices existantes se concentrent avant tout sur la façon de protéger la vie privée, la dignité, l'autonomie et la liberté individuelle *face aux* progrès de l'IA, et qu'elles ne se posent guère la question de savoir si ces principes pourraient être activement encouragés par des innovations responsables en matière d'IA.

La question de la confiance dans l'IA, traitée dans moins d'un tiers de toutes les sources, aborde un dilemme éthique essentiel de la gouvernance de l'IA : est-il souhaitable sur le plan de l'éthique de favoriser la confiance du public dans l'IA ? Si plusieurs sources, en particulier celles qui émanent du secteur privé, soulignent qu'il importe d'encourager la confiance dans l'IA par l'éducation et la sensibilisation, un petit nombre d'entre elles affirment que la confiance dans l'IA risque en fait d'entraîner un allègement des contrôles et d'aller à l'encontre de certaines des obligations sociétales des producteurs d'IA[21]. Cette hypothèse remet en question le point de vue dominant en matière d'éthique de l'IA, à savoir que renforcer la confiance du public dans l'IA est un critère essentiel de la gouvernance éthique[22]. S'agissant de la confiance, nous avons relevé d'autres difficultés conceptuelles. Tout d'abord, les documents actuels ne semblent clarifier ni le sens de la notion de confiance ni la dynamique de la confiance. La plupart des sources omettent de préciser qui est de quel côté dans la relation de confiance qu'elles décrivent, et font donc abstraction du fait qu'il s'agit d'une relation très complexe impliquant au moins deux acteurs, qui ne doivent pas douter du fait que l'autre fera ou ne fera pas quelque chose. Tout un ensemble d'éléments, par exemple la culture, les croyances, le contexte ainsi que les caractéristiques des acteurs, viennent influencer la relation de confiance. Or la littérature actuelle semble ne pas tenir compte de ces facteurs contextuels. En outre, la « fiabilité », une caractéristique, et la « confiance », un concept relationnel, paraissent fréquemment confondues ou utilisées de façon interchangeable par les acteurs de l'IA que nous avons étudiés. Il en découle non seulement une confusion entre deux notions mais aussi de faux espoirs pour les utilisateurs de l'IA et les décideurs. La confiance et la fiabilité sont deux notions différentes, la fiabilité n'étant pas à elle seule gage de confiance. Aux fins de la gouvernance, il faudrait éclaircir ces différences notionnelles capitales et exiger plus de précisions sur les impératifs d'une relation de confiance.

Le fait que les thèmes de la durabilité et de la solidarité sont relativement peu abordés pourrait signifier qu'ils ne sont actuellement pas pris en compte dans le discours éthique dominant sur l'IA. Le fait que les principes liés à la durabilité ne sont que très peu traités est particulièrement problématique, car le déploiement de l'IA exige des ressources informatiques considérables, qui entraînent à leur tour une consommation énergétique élevée[23]. L'impact environnemental de l'IA ne se réduit toutefois pas aux effets négatifs de l'empreinte écologique élevée de ses infrastructures numériques : l'IA peut aussi être exploitée dans l'intérêt des écosystèmes et de la biosphère dans son intégralité. Ce dernier point, mis en avant dans un rapport du Forum économique mondial, et non dans ses lignes directrices sur l'IA, devra être approuvé à plus grande échelle pour pouvoir être incorporé dans le discours sur l'IA éthique[24]. Le principe éthique de solidarité est peu abordé et lorsqu'il l'est, c'est le plus souvent en lien avec le développement de stratégies inclusives visant à éviter des pertes d'emploi et un partage inéquitable des fardeaux. Peu de sources documentaires évoquent le fait que l'on pourrait encourager la solidarité en s'appuyant sur la possibilité, récente, d'utiliser l'expertise en matière d'IA pour résoudre des problèmes humanitaires, une mission que mènent actuellement, entre autres, diverses organisations intergouvernementales, comme le Bureau des Nations Unies pour les services d'appui aux projets (UNOPS) ou l'Organisation mondiale de la santé (OMS), ainsi que des entreprises du secteur privé comme Microsoft. Alors que le coût humanitaire du changement climatique causé par l'homme augmente rapidement[25], les principes de durabilité et de solidarité paraissent étroitement liés, mais ils sont très peu évoqués par rapport à d'autres principes.

Si les données numériques montrent qu'un consensus se forme autour de la promotion de certains principes éthiques, une analyse thématique approfondie dépeint un tableau bien plus complexe : il existe en effet des différences cruciales dans la *façon* dont ces principes sont interprétés et au sujet des critères jugés indispensables à leur réalisation. Nous avons constaté que des mesures différentes et souvent incompatibles étaient proposées pour mettre en œuvre l'IA éthique. C'est ainsi que la nécessité de disposer de jeux de données plus vastes et plus diversifiés dans un souci d'impartialité de l'IA semble difficilement conciliable avec l'obligation de donner davantage de contrôle aux personnes sur leurs données et l'utilisation qui en est faite, aux fins du respect de leur vie privée et de l'autonomie. Des contrastes similaires se font jour entre le besoin d'éviter les préjudices à tout prix et celui de trouver un équilibre entre risques et avantages. Par ailleurs, il faut noter que l'évaluation du rapport risques-avantages donnera des résultats différents selon les personnes pour le bien-être desquelles il sera optimisé et les acteurs qui s'en occuperont. Si ces questions ne sont pas résolues, des décalages et des tensions pourraient compromettre l'élaboration d'un programme mondial pour l'IA éthique.

Bien que tous les acteurs s'accordent à dire que l'IA doit être éthique, d'importantes divergences se font jour dans les lignes directrices en faveur de l'IA éthique. En outre, nul ne sait encore précisément comment les principes et lignes directrices éthiques devraient être mis en œuvre. Ces difficultés ont des incidences sur les politiques scientifiques, la gouvernance des technologies et l'éthique de la recherche. Au niveau des pouvoirs publics, il faut absolument que les organismes publics renforcent leur coopération afin d'harmoniser et de hiérarchiser les priorités de leurs programmes d'IA, tâche pour laquelle les organisations intergouvernementales peuvent jouer un rôle d'intermédiaires et de facilitatrices. Si une harmonisation est souhaitable, il ne faut pas qu'elle se fasse au détriment du pluralisme culturel et éthique en matière d'IA. Le premier défi consistera donc, pour pouvoir établir un programme mondial sur l'IA, à trouver un équilibre entre la nécessaire harmonisation entre pays et le respect de la diversité culturelle et du pluralisme éthique. Il faudra, pour ce faire, mettre en place des mécanismes de délibération pour trancher les désaccords entre les différents acteurs de chaque région du monde au sujet des valeurs et des incidences des progrès de l'IA. En matière de gouvernance technologique, l'harmonisation passe en règle générale par la normalisation. Des efforts ont été faits dans ce sens, notamment par l'Institut des ingénieurs électriciens et électroniciens (*Institute of Electrical and Electronics Engineers*, IEEE), dans le cadre de l'initiative « Ethically Aligned Design » (conception respectueuse de l'éthique)[26]. Enfin, il sera de plus en plus souvent fait appel à des mécanismes de gouvernance souples tels que les comités d'examen indépendant (*Independent Review Boards*, IRB) pour évaluer la valeur éthique des applications de l'IA dans la recherche scientifique, tout particulièrement dans le milieu universitaire. Toutefois, les

applications de l'IA par les pouvoirs publics ou des entreprises privées échapperont probablement à la surveillance des IRB, à moins que les compétences de ces derniers ne soient considérablement élargies.

Dans l'ensemble, nous avons constaté que la communauté internationale n'est pas d'accord sur ce qui constitue l'IA éthique ni sur les critères à réunir impérativement pour sa mise en œuvre. Néanmoins, des signes de convergence apparaissent autour des notions de transparence, de non-malfaisance, de responsabilité et de respect de la vie privée. Enrichir le débat actuel sur l'IA éthique en améliorant l'analyse des principes éthiques essentiels – mais peu évoqués – que sont la dignité humaine, la solidarité et la durabilité, permettrait vraisemblablement de mieux définir le cadre éthique de l'IA. Il serait par ailleurs souhaitable que les travaux, qui sont pour l'instant axés sur la formulation des principes, soient réorientés vers la réalisation de ces derniers. Un programme mondial pour l'IA éthique devrait concilier la nécessaire harmonisation entre pays et entre domaines avec le respect de la diversité culturelle et du pluralisme moral. Globalement, notre analyse offre un précieux point de départ pour comprendre la diversité intrinsèque des principes et lignes directrices actuels sur l'IA éthique et elle donne un aperçu des difficultés qui attendent la communauté internationale.

Incidences sur l'élaboration des politiques

Les nombreuses initiatives internationales visant à élaborer des documents de droit souple sur l'IA fournissent indirectement des informations utiles sur la façon dont l'humanité réagira aux multiples difficultés de gouvernance que pose l'IA. La communauté internationale semble commencer à s'accorder sur l'importance de la transparence, de la non-malfaisance, de la responsabilité et du respect de la vie privée dans le cadre du développement et du déploiement de l'IA éthique. Toutefois, enrichir l'actuel débat sur l'IA éthique en améliorant l'analyse des principes essentiels, pourtant peu évoqués, que sont la dignité humaine, la solidarité et la durabilité, permettrait vraisemblablement de mieux définir le cadre éthique de l'IA. De plus, les travaux, qui sont pour l'instant axés sur la formulation des principes, doivent être réorientés vers la réalisation de ces derniers. Un programme mondial pour l'IA éthique devrait concilier la nécessaire harmonisation entre pays et entre domaines avec le respect de la diversité culturelle et du pluralisme moral.

Ces éléments ont des incidences sur les politiques publiques, la gouvernance des technologies et l'éthique de la recherche. Au niveau des pouvoirs publics, il faut renforcer la coopération entre les parties prenantes pour harmoniser les différents programmes relatifs à l'éthique de l'IA et faire converger les procédures non seulement au sujet des principes éthiques mais aussi de leur mise en œuvre. S'il est souhaitable qu'un consensus se dégage à l'échelon mondial, il ne faut pas que ce soit au détriment du pluralisme culturel et éthique ; aussi faudrait-il peut-être mettre en place des mécanismes de délibération pour trancher les désaccords entre les différents acteurs de chaque région du monde. Dans ce cadre, des organisations intergouvernementales comme le Conseil de l'Europe peuvent jouer un rôle d'intermédiaires et de facilitatrices, en complément de quoi, des approches ascendantes associant toutes les parties prenantes concernées sur un pied d'égalité peuvent être adoptées.

Les pouvoirs publics devraient intervenir pour préciser de quelle façon les lignes directrices sur l'éthique de l'IA s'articulent avec les textes de droit national et international en vigueur. Malgré la prétendue originalité sociotechnique de l'IA, les documents de droit souple sur l'IA ne s'appliquent pas dans un vide éthico-juridique. Au contraire, les lignes directrices sur l'éthique et les autres instruments non contraignants devront au bout du compte s'appliquer dans un cadre dans lequel il existe déjà une multitude de règles, notamment de droit dur (règles de gouvernance obligatoires). Ne pas tenir compte du cadre dans lequel s'inscrivent ces règles pourrait compromettre la traduction des principes ici évoqués en mesures concrètes et efficaces de gouvernance internationale. La transparence, principe éthique le plus largement traité, en est un exemple. Le principe de transparence est certes fréquemment évoqué, mais le plus souvent sans lien explicite avec les règles sous-jacentes qui ont force contraignante. De nos jours, les établissements qui se servent de la technologie de l'IA sont déjà soumis à de nombreuses règles de transparence au titre des mécanismes juridiques en vigueur, par exemple, aux États-Unis, la loi intitulée *Fair Credit Reporting Act*, et, en Europe, les obligations concrètes des responsables du traitement et des sous-traitants

qui sont énoncées aux articles 12 à 14 du Règlement général sur la protection des données (RGPD) de l'UE. De la même manière, il est indispensable que les décideurs précisent la distinction entre « confiance » et « fiabilité ».

Non seulement il faudra relever le défi de faire concorder le droit dur et le droit souple, mais aussi celui de mettre les principes éthiques en pratique tout cherchant à harmoniser des codes divergents sur l'éthique de l'IA. Au niveau de la gouvernance technologique, des tentatives d'harmonisation prometteuses ont eu lieu dans le cadre d'initiatives de normalisation telles que celles de l'IEEE, qui est la plus grande association professionnelle de métiers techniques au monde, dont les travaux sont consacrés aux progrès de l'innovation technologique. L'IEEE mène des travaux en matière d'éthique de l'IA aussi bien au sujet de systèmes généraux autonomes et intelligents, dans le cadre de l'initiative sur la conception respectueuse de l'éthique[26], que de systèmes spécifiques, avec, par exemple, la feuille de route qu'elle a établie en neurotechnologie au sujet des normes relatives à l'interface cerveau-machine.

L'IA a par ailleurs des incidences sur la supervision de la recherche. Il sera de plus en plus fait appel à des mécanismes d'éthique de la recherche, par exemple les comités d'examen indépendant (*Independent Review Boards*, IRB), pour évaluer la valeur éthique des applications de l'IA dans la recherche scientifique, tout particulièrement dans le milieu universitaire. Toutefois, les applications de l'IA par les pouvoirs publics ou des entreprises privées échapperont probablement à la surveillance des IRB, à moins que les compétences de ces derniers ne soient considérablement élargies.

Dans l'ensemble, la diversité thématique et la richesse informationnelle des documents que nous avons analysés conduisent à penser que les instruments non contraignants publiés par des institutions gouvernementales et des entités non gouvernementales (en ce compris les entreprises privées et les établissements universitaires) sont de précieux outils pour exercer une influence concrète sur l'élaboration des politiques concernant l'IA par les pouvoirs publics. Si elles sont adéquatement conçues et ébauchées, les initiatives de droit souple peuvent orienter le développement des systèmes d'IA en faveur du bien-être social et dans le respect des valeurs éthiques et des normes juridiques. Il ne faut toutefois pas considérer qu'elles remplacent les instruments de gouvernance contraignants. Il est particulièrement à craindre que les acteurs du secteur privé de l'IA n'adoptent des mesures d'autoréglementation destinées à contourner ou à écarter les mesures de gouvernance contraignantes édictées par des autorités gouvernementales ou intergouvernementales. Ce risque a été mis en avant par le philosophe allemand Thomas Metzinger, membre du groupe d'experts de haut niveau de l'UE sur l'IA, qui a fait observer que le débat sur l'éthique de l'IA est en grande partie façonné par le secteur privé[20].

Le fait que les organismes diffusant des lignes directrices éthiques sur l'IA représentent certaines zones géographiques bien plus que d'autres doit être surveillé et examiné de près par des organisations internationales, notamment intergouvernementales. Par souci d'inclusivité, de pluralisme culturel et de participation équitable à la prise de décisions collective au sujet de l'IA, il faudrait encourager l'élaboration de documents de droit souple par des organismes situés dans les régions du monde qui sont actuellement sous-représentées, tout particulièrement l'Afrique et l'Amérique du Sud. Les organisations intergouvernementales comme le Conseil de l'Europe peuvent jouer un rôle capital dans la création de plateformes internationales d'échanges et de débats sur l'éthique et la gouvernance de l'IA.

Les instruments non contraignants en vigueur convergent vers cinq principes éthiques génériques – transparence, justice, non-malfaisance, responsabilité et respect de la vie privée – qui sont autant de domaines à surveiller en priorité, au sujet desquels des mesures de gouvernance obligatoires pourraient être adoptées aux niveaux gouvernemental et intergouvernemental. Se consacrer en priorité à la réalisation de ces principes pourrait faciliter l'élaboration d'un train de normes essentielles, fondées sur des préceptes éthiques faisant l'unanimité. Par ailleurs, comme un très grand nombre d'acteurs privés et publics adhèrent à ces principes, ils seront probablement plus largement respectés. Cela étant, afin que ces principes éthiques puissent être dûment traduits en règles de gouvernance, ils doivent être précisés sur le plan conceptuel. Il incombe aux décideurs de

remédier aux ambiguïtés sémantiques de ces principes et à leurs caractéristiques antagonistes. Les profondes divergences entre les textes de droit souple en vigueur au sujet de l'interprétation et de l'application concrète de ces principes montrent que le public sera probablement divisé au sujet des solutions de gouvernance contraignantes et qu'un débat démocratique transparent sera donc nécessaire.

Parallèlement, les notions éthiques qui sont relativement peu traitées, par exemple la durabilité, la dignité et la solidarité, doivent être analysées de très près pour que les lacunes conceptuelles et normatives que présente le droit souple ne soient pas transposées dans le droit dur. Des mesures de gouvernance obligatoires doivent compléter les approches non contraignantes et combler leurs lacunes au lieu de reproduire celles-ci. Pour dûment faire face aux difficultés que pose l'IA en termes de durabilité et de solidarité, peut-être faudra-t-il renforcer la coopération entre les autorités chargées de la protection de l'environnement, les ministères du Travail et de l'Emploi ainsi que les ministères de la Technologie et de l'Innovation.

Étant donné que près de la moitié des textes de droit souple qui ont été examinés ne recommandent pas explicitement de favoriser le respect des droits de l'homme – pas plus qu'ils ne mettent pas en garde contre leur violation – lors de la conception, du développement et du déploiement des systèmes d'IA, il est urgent de se focaliser davantage sur les incidences de l'IA sur les droits de l'homme. Les pays membres du Conseil de l'Europe sont bien placés pour orienter la gouvernance internationale de l'IA vers la promotion des droits de l'homme. Les incidences de l'IA sur les droits de l'homme doivent être examinées attentivement sous divers angles. Tout d'abord sous l'angle des droits et des obligations au sens philosophique, car ils existent indépendamment de toute loi en tant que normes morales justifiées. Ensuite, sous l'angle du droit international des droits de l'homme. À cet égard, la Convention européenne des droits de l'homme (CEDH) peut jouer un rôle essentiel dans la recherche doctrinale et la réflexion sur l'IA à l'échelon international. Son respect est une condition essentielle au développement et à l'adoption socialement responsables d'une nouvelle technologie. Il est donc de la plus haute importance d'évaluer les effets de la transformation sociotechnique qu'entraîne l'IA sur les droits et libertés fondamentaux consacrés par la CEDH. Cette évaluation doit avoir un double objectif : i) examiner si l'IA aura des répercussions ou fera peser de nouveaux risques sur les droits de l'homme et les libertés, et de quelle façon ; ii) examiner si le développement responsable de l'IA et la consultation du public à cet égard pourraient contribuer à la promotion de ces droits et libertés, et de quelle façon. Il faut souligner que comme les technologies ne sont pas développées dans le néant mais dans un contexte sociohistorique de pratiques, de coutumes et de normes humaines, les stratégies d'évaluation des effets de l'IA devront, pour être efficaces, s'inscrire dans le contexte des pratiques et normes actuelles et non en faire abstraction[27]. Enfin, il est important d'examiner l'interface entre l'IA et les droits de l'homme non seulement d'un point de vue général, mais aussi et surtout sous l'angle de la prépondérance des droits de l'homme dans des domaines d'application spécifiques de l'IA, entre autres la robotique[8,28], les mégadonnées[29,30], les armes autonomes[31,32] et les interfaces cerveau-machine[33].

Remerciements

Les auteurs tiennent à remercier Mmes Anna Jobin et Karolina Ignatiadis et M. Manuel Schneider, dont les travaux ont contribué à la réalisation du présent rapport.

Bibliographie

1. Michie D, Spiegelhalter DJ, Taylor C. Machine learning. *Neural and Statistical Classification.* 1994;13.
2. Appenzeller T. The AI revolution in science. *Science.* 2017;357:16-17.
3. Harari YN. Reboot for the AI revolution. *Nature News.* 2017;550(7676):324.
4. Helbing D, Frey BS, Gigerenzer G, et al. Will democracy survive big data and artificial intelligence? In: *Towards Digital Enlightenment.* Springer; 2019:73-98.
5. Livingston S, Risse M. The Future Impact of Artificial Intelligence on Humans and Human Rights. *Ethics & International Affairs.* 2019;33(2):141-158.
6. Nemitz P. Constitutional democracy and technology in the age of artificial intelligence. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences.* 2018;376(2133):20180089.
7. Ashrafian H. Intelligent robots must uphold human rights. *Nature.* 2015;519(7544):391-391.
8. Van Est R, Gerritsen J, Kool L. Human rights in the robot age: Challenges arising from the use of robotics, artificial intelligence, and virtual and augmented reality–Expert report written for the Committee on Culture. *Science, Education and Media of the Parliamentary Assembly of the Council of Europe (PACE)(Rathenau Institute), extrait de https://www.rathenau.nl/sites/default/files/2018-02/Human%20Rights%20in%20the%20Robot%20Age-Rathenau%20Instituut-2017.pdf (5 janvier 2019), 2017*.
9. Raso FA, Hilligoss H, Krishnamurthy V, Bavitz C, Kim L. Artificial Intelligence & Human Rights: Opportunities & Risks. *Berkman Klein Center Research Publication.* 2018(2018-6).
10. Assembly UG. Déclaration universelle des droits de l'Homme. *Assemblée générale des Nations Unies*1948;302(2).
11. Mowbray A. The European Convention on Human Rights. In: *International Human Rights Law.* Routledge; 2016:287-304.
12. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. *International journal of social research methodology.* 2005;8(1):19-32.
13. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. *Nature Machine Intelligence.* 2019;1(9):389-399.
14. Ienca M, Ferretti A, Hurst S, Puhan M, Lovis C, Vayena E. Considerations for ethics review of big data health research: A scoping review. *PloS one.* 2018;13(10):e0204937.
15. Li Y, Krishnamurthy R, Raghavan S, Vaithyanathan S, Jagadish H. Regular expression learning for information extraction. Document présenté lors de la conférence de 2008 sur les méthodes empiriques de traitement du langage naturel.
16. Shoham Y, Perrault R, Brynjolfsson E, et al. The AI Index 2018 annual report. *AI Index Steering Committee, Human-Centered AI Initiative, Stanford University, Stanford, CA.* 2018.
17. Turilli M, Floridi L. The ethics of information transparency. *Ethics and Information Technology.* 2009;11(2):105-112.
18. Rozin P, Royzman EB. Negativity bias, negativity dominance, and contagion. *Personality and social psychology review.* 2001;5(4):296-320.
19. Ito TA, Larsen JT, Smith NK, Cacioppo JT. Negative information weighs more heavily on the brain: the negativity bias in evaluative categorizations. *Journal of personality and social psychology.* 1998;75(4):887.
20. Peter J. Bentley, Miles Brundage, Olle Häggström, Thomas Metzinger. Should we fear artificial intelligence? *European Parliamentary Research Service.* 2018:Scientific Foresight Unit (STOA).
21. Bryson J. No one should trust artificial intelligence. *Science & Technology: Innovation, Governance, Technology.* 2018;11:14.
22. Winfield AF, Jirotka M. Ethical governance is essential to building trust in robotics and artificial intelligence systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering*

Sciences. 2018;376(2133):20180085.
23. Strubell E, Ganesh A, McCallum A. Energy and policy considerations for deep learning in NLP. *arXiv preprint arXiv:190602243.* 2019.
24. Forum WE. Harnessing Artificial Intelligence for the Earth.
25. Scheffran J, Brzoska M, Kominek J, Link PM, Schilling J. Climate change and violent conflict. *Science.* 2012;336(6083):869-871.
26. IEEE. Ethically aligned design. *IEEE Standards v1.* 2016(Global Initiative).
27. Rasmussen T. *Social theory and communication technology.* Routledge; 2019.
28. Liu H-Y, Zawieska K. From responsible robotics towards a human rights regime oriented to the challenges of robotics and artificial intelligence. *Ethics and Information Technology.* 2017:1-13.
29. Mantelero A. AI and Big Data: A blueprint for a human rights, social and ethical impact assessment. *Computer Law & Security Review.* 2018;34(4):754-772.
30. Vayena E, Tasioulas J. The dynamics of big data and human rights: The case of scientific research. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences.* 2016;374(2083):20160129.
31. Heyns C. Human rights and the use of autonomous weapons systems (AWS) during domestic law enforcement. *Hum Rts Q.* 2016;38:350.
32. Asaro P. On banning autonomous weapon systems: human rights, automation, and the dehumanization of lethal decision-making. *International Review of the Red Cross.* 2012;94(886):687-709.
33. Ienca M, Andorno R. Towards new human rights in the age of neuroscience and neurotechnology. *Life Sci Soc Policy.* 2017;13(1):5.