

Strasbourg, 18 mai 2021

CAHAI-PDG(2021)05
Provisoire

COMITÉ AD HOC SUR L'INTELLIGENCE ARTIFICIELLE (CAHAI)

GROUPE D'ÉLABORATION DES POLITIQUES (CAHAI-PDG)

Analyse d'impact des systèmes d'IA sur les droits de l'homme, la démocratie et l'État de droit¹

www.coe.int/cahai

¹ Ce projet de document sera revu par le CAHAI-PDG et ne doit en aucun cas être considéré comme un document définitif.

Analyse d'impact des systèmes d'IA sur les droits de l'homme, la démocratie et l'État de droit	3
Introduction et champ d'application	3
Section I. Modèle d'analyse d'impact sur les droits de l'homme, la démocratie et l'État de droit : considérations méthodologiques	4
A. Analyse d'impact sur les droits de l'homme : sources, documents et expériences	5
1. Analyses d'impact sur les droits de l'homme existantes	7
<i>i) Documents et cadres généraux faisant référence aux analyses d'impact sur les droits de l'homme en général</i>	7
<i>ii) Documents et cadres spécifiques ayant trait à l'IA</i>	8
B. Recherche des caractéristiques pertinentes pour la conception d'un modèle d'AIDHDE	11
1. Les grandes caractéristiques des systèmes d'IA en tant que variables d'évaluation	11
<i>i) Le contexte de l'application en tant que variable de l'analyse d'impact</i>	12
<i>ii) Les technologies sous-jacentes des systèmes d'IA en tant que variable de l'analyse d'impact</i>	13
<i>iii) Les acteurs en jeu et l'étape de développement du système d'IA</i>	18
<i>iv) Les parties prenantes</i>	19
2. La démocratie et l'État de droit en tant que dimensions d'un modèle d'analyse intégré des systèmes d'IA	19
<i>i) Difficultés méthodologiques de l'analyse d'impact des applications d'IA sur la démocratie et l'État de droit</i>	19
<i>ii) Critères fondés sur des droits applicables à la démocratie et à l'État de droit</i>	21
Section II. Vers un modèle de réalisation d'une analyse d'impact des systèmes d'IA sur les droits de l'homme, la démocratie et l'État de droit, sur la base de l'expérience acquise avec les systèmes d'analyse d'impact de l'IA sur les droits de l'homme	25
Section III. Synergies entre AIDHDE et mécanismes de conformité	34
A. Harmonisation de l'AIDHDE et des mécanismes de conformité	34
B. Harmonisation des mécanismes correctifs et de l'AIDHDE	35
<i>i) L'AIDHDE en tant qu'outil de renforcement de l'autonomie par l'information</i>	36
<i>ii) L'AIDHDE en tant qu'élément d'un processus plus large de diligence raisonnable en matière de droits de l'homme</i>	36
<i>iii) L'AIDHDE en tant que socle commun pour la conception des dispositifs de réparation</i>	37
TABLEAU 1	38

Analyse d'impact des systèmes d'IA sur les droits de l'homme, la démocratie et l'État de droit

Introduction et champ d'application

À la suite de l'adoption en décembre 2020 de l'Étude de faisabilité du CAHAI, qui comprend un chapitre spécifique (ch. 9) sur les mécanismes pratiques et de suivi nécessaires pour assurer la conformité avec un cadre juridique sur l'intelligence artificielle (IA) fondé sur les normes du Conseil de l'Europe en matière de droits de l'homme, d'État de droit et de démocratie, le CAHAI, via son Groupe d'élaboration des politiques (CAHAI-PDG), a décidé d'examiner de plus près l'un de ces mécanismes, à savoir les analyses d'impact sur les droits de l'homme, et :

1. de définir une méthodologie pour réaliser des analyses d'impact des applications d'intelligence artificielle (IA) du point de vue des droits de l'homme, de la démocratie et de l'État de droit, sur la base des normes pertinentes du Conseil de l'Europe et des travaux déjà menés dans ce domaine au niveau des pays et au niveau international ²;
2. d'élaborer un modèle d'analyse d'impact ;
3. d'examiner la complémentarité d'une telle analyse avec d'autres mécanismes de conformité décrits au chapitre 9 de l'Étude de faisabilité.

Le présent document est organisé selon ce plan.

La section I expose les considérations méthodologiques applicables à un modèle d'analyse d'impact des applications d'IA sur les droits de l'homme (DH), la démocratie (D) et l'État de droit (E) (ou AIDHDE). À cet égard, le document présente des outils et des cadres d'analyse d'impact existants qui ont trait à l'IA en particulier ou qui s'appliquent à l'analyse d'impact sur les droits de l'homme en général. Il examine également les caractéristiques les plus pertinentes de ces outils et cadres, en particulier les principales propriétés et variables d'analyse des systèmes d'IA et les façons d'intégrer l'État de droit et la démocratie dans un modèle d'analyse intégré des systèmes d'IA.

À la section II, le document propose une méthodologie d'AIDHDE dans la perspective d'une description plus détaillée d'un modèle d'analyse concret.

La section III examine la complémentarité entre une future AIDHDE fondée sur le modèle des analyses d'impact existantes mentionnées à la section II et les mécanismes de conformité décrits au chapitre 9 de l'Étude de faisabilité, le but étant d'obtenir une plus grande valeur ajoutée et une meilleure complémentarité.

² Voir par exemple la Recommandation CM/Rec (2020)1 sur les impacts des systèmes algorithmiques sur les droits de l'homme, les travaux en cours de l'OCDE sur la classification des systèmes d'IA et la boîte à outils pour l'analyse d'impact sur les droits de l'homme élaborée par l'Institut danois des droits de l'homme.

Section I. Modèle d'analyse d'impact sur les droits de l'homme, la démocratie et l'État de droit : considérations méthodologiques

Les considérations méthodologiques applicables à tout modèle d'analyse d'impact des systèmes d'IA fondée sur les droits de l'homme, la démocratie et l'État de droit (ou « AIDHDE pour l'IA », comme indiqué plus haut) devraient s'appuyer sur les pratiques établies et les expériences acquises en matière d'analyse d'impact sur les droits de l'homme (AIDH) **(A)**. Ces considérations méthodologiques devraient également prendre acte de l'origine de l'AIDH, laquelle s'inscrit dans le cadre de la diligence raisonnable en matière de droits de l'homme telle qu'elle figure dans les Principes directeurs des Nations Unies relatifs aux entreprises et aux droits de l'homme et dans les Principes directeurs de l'OCDE pour les entreprises multinationales³. Il est important d'insister sur cette origine, car la diligence raisonnable en matière de droits de l'homme est un processus continu et non un instantané à un moment donné. De plus, cette diligence ne se limite pas aux activités d'une seule entreprise, mais couvre la totalité de la chaîne de valeur. Ce principe devrait s'appliquer à toute AIDHDE pour l'IA.

Cela étant, du fait des particularités des systèmes d'IA, une simple transposition au domaine de l'IA de l'acquis relatif à la diligence raisonnable en matière de droits de l'homme pose un certain nombre de problèmes. Se pose également la question d'intégrer la démocratie et l'État de droit dans toute analyse complète d'un système d'IA, car ces dimensions sont généralement absentes des analyses d'impact existantes. Cette section se propose donc aussi de passer en revue et d'analyser les éléments du cadre de l'AIDH qui peuvent être transposés au contexte de l'IA et d'étudier comment étendre ce cadre pour y intégrer la démocratie et l'État de droit (dans la perspective d'une AIDHDE) **(B)**.

En ce qui concerne la portée de l'analyse d'impact, il importe de noter que les cadres généraux d'AIDH mettent le plus souvent l'accent sur les effets négatifs des activités des entreprises sur les droits de l'homme. C'est également le cas de la plupart des modèles d'analyse d'impact des systèmes d'IA actuellement utilisés dans le privé ou dans le public. En conséquence, toute AIDHDE devrait être élaborée dans cette logique et porter en particulier sur les effets négatifs de ces systèmes sur les droits de l'homme, la démocratie et l'État de droit.

Bien entendu, ce choix ne signifie pas que l'utilisation de l'IA ne produit que des effets négatifs. L'IA présente en effet de nombreux avantages et peut avoir des effets bénéfiques considérables pour l'humanité. Elle peut même faciliter l'exercice, la protection et le renforcement des droits de l'homme, et cette contribution positive ne doit pas être négligée. Cela dit, le but d'une AIDH est de déceler les éventuels risques qu'un système d'IA donné porte atteinte aux droits de l'homme et non de mettre ces risques en balance avec les éventuels effets bénéfiques d'une telle application. L'évaluation du rapport bénéfices-risques ne figure pas dans la méthodologie d'analyse d'impact. Cet exercice serait plutôt réalisé à une étape ultérieure, lorsqu'il est

³ Documents disponibles en ligne : https://www.ohchr.org/documents/publications/guidingprinciplesbusinessshr_fr.pdf et <https://www.oecd.org/fr/daf/inv/mne/2011102-fr.pdf>.

question d'apprécier l'opportunité de déployer une application d'IA donnée. Ainsi, il peut arriver que les pouvoirs publics, jugeant que les effets négatifs sont compensés par les effets positifs, décident d'utiliser une application d'IA pour une finalité donnée. S'il en résulte une limitation d'un ou plusieurs droits de l'homme (ce que l'AIDHDE peut aider à évaluer), il est essentiel que cela soit justifié par une démarche proportionnée et nécessaire dans une société démocratique, par exemple le souci de protéger la sécurité nationale ou un autre intérêt public légitime.

Il faut également se demander si l'AIDHDE doit s'appliquer à des acteurs privés, à des acteurs publics ou aux deux. Même si, à l'évidence, les acteurs publics et les acteurs privés n'ont ni les mêmes rôles ni les mêmes obligations au regard du respect des droits de l'homme, la réalisation d'une AIDHDE lors de l'élaboration, du déploiement ou de l'acquisition de systèmes d'IA est tout aussi importante pour les premiers que pour les seconds, comme le soulignent plusieurs documents du Conseil de l'Europe⁴. De fait, ces deux types d'acteurs peuvent déployer l'IA d'une manière qui porte atteinte aux droits de l'homme, à la démocratie et à l'État de droit. De plus, il convient de noter que dans le contexte de l'IA, la distinction entre sphère publique et sphère privée est souvent brouillée. En effet, d'une part, les acteurs privés qui développent des solutions d'IA jouent un rôle croissant dans la sphère publique⁵, et, d'autre part, le secteur public et le secteur privé sont de plus en plus souvent amenés à collaborer, car les applications d'IA utilisées dans le secteur public sont très souvent développées par des acteurs privés. L'AIDHDE a pour but d'identifier les risques liés à une application d'IA donnée et ainsi de contribuer au respect des droits de l'homme, ce qui est un devoir important dans le public comme dans le privé.

Nous examinerons tout d'abord diverses sources concernant les AIDH traditionnelles ainsi que le contenu de ces analyses d'impact. Nous fournirons ensuite une liste des AIDH existantes – de type général ou propres à l'IA – qu'il convient de prendre en considération en vue de l'élaboration d'un modèle global d'AIDHDE, et nous préciserons quelles sont leurs principales caractéristiques applicables à notre contexte.

A. Analyse d'impact sur les droits de l'homme : sources, documents et expériences

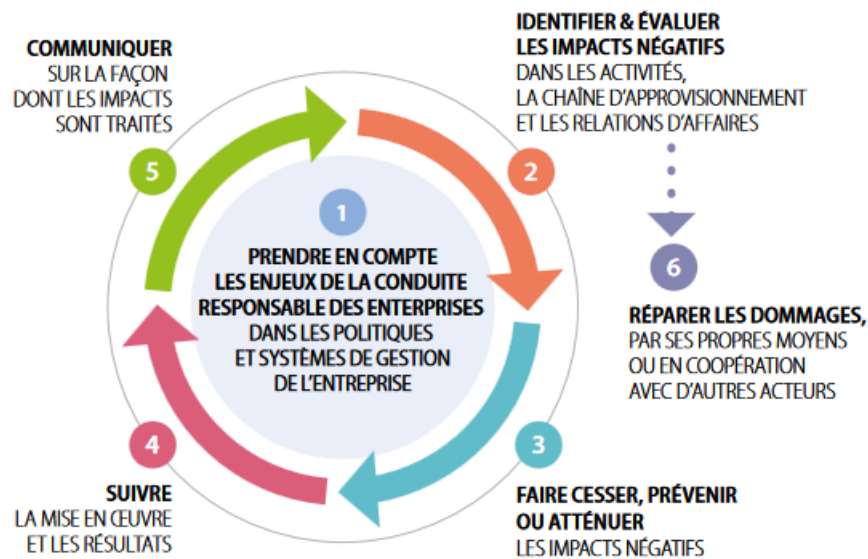
Les analyses d'impact sur les droits de l'homme (AIDH) générales, non spécifiques à l'IA, s'inspirent des cadres internationaux susmentionnés qui imposent une diligence raisonnable en matière de droits de l'homme⁶. La diligence raisonnable en matière de droits de l'homme est un

⁴ Les acteurs publics et les acteurs privés avec lesquels ils travaillent sont aussi censés effectuer une analyse d'impact sur les droits fondamentaux. Voir la Recommandation de la Commissaire aux droits de l'homme, *Décoder l'intelligence artificielle*, pp. 8 et 9, disponible à l'adresse <https://rm.coe.int/decoder-l-intelligence-artificielle-10-mesures-pour-protger-les-droit/168094b6e2>, et la Recommandation CM/Rec(2020)1 du Comité des Ministres aux États membres sur les impacts des systèmes algorithmiques sur les droits de l'homme, point 5.2, disponible à l'adresse https://search.coe.int/cm/pages/result_details.aspx?ObjectId=09000016809e1124.

⁵ Par exemple, certains outils d'IA appliquent à des communautés mondiales entières (réseaux sociaux) leurs propres règles et mécanismes de contrôle et de résolution de litiges. Les régulateurs de ces communautés ressemblent souvent à des acteurs publics.

⁶ Au sens du Principe 17 des Principes directeurs des Nations Unies relatifs aux entreprises et aux droits de l'homme.

processus continu et itératif qui comprend les étapes suivantes⁷ :



Il convient de noter que l'AIDH n'est pas identique à la diligence raisonnable en matière de droits de l'homme : elle en est une composante. On la retrouve en particulier au niveau des étapes 2 (identification des risques) et 6 (réparer les dommages par ses propres moyens ou en coopération, si l'analyse d'impact a mis en évidence une atteinte aux droits de l'homme qu'un acteur a causée ou à laquelle il contribue).

Il importe également de noter que l'AIDH ne comporte pas d'élément normatif : son but est d'identifier et d'évaluer l'impact négatif, sans entrer plus avant dans les conséquences d'une telle évaluation (déterminer par exemple s'il conviendrait d'interrompre le développement ou l'utilisation d'une application d'IA ou d'en limiter le type d'utilisation, ou s'il serait souhaitable de modifier le modèle d'IA ou d'améliorer la qualité des données). Par conséquent, si une AIDH n'indique pas, en soi, si une autorité de surveillance publique ou un organisme de certification doit agir ou prendre des mesures, elle peut constituer pour eux une source d'informations utile, de même que pour les développeurs ou les utilisateurs de la chaîne de valeur de l'application d'IA qui souhaitent continuer de développer ou d'utiliser l'application en question. On notera aussi avec intérêt que dans certains pays, l'analyse d'impact sur les droits de l'homme est obligatoire.

On trouve une description détaillée de l'AIDH générale dans des boîtes à outils spécialisées, par exemple celle élaborée par l'Institut danois des droits de l'homme. L'ensemble des pratiques et normes en matière d'AIDH fournit des points de référence utiles pour l'élaboration d'un modèle d'AIDHDE. Il est important de prendre cet aspect en compte.

⁷ Guide OCDE sur le devoir de diligence pour une conduite responsable des entreprises, p. 25, disponible à l'adresse <https://www.oecd.org/fr/daf/inv/mne/Guide-OCDE-sur-le-devoir-de-diligence-pour-une-conduite-responsable-des-entreprises.pdf>.

Une méthodologie spécialement conçue pour mener une analyse d'impact relatif à la protection des données (AIPD) est apparue récemment. Cette proposition, qui a pour but d'élaborer un modèle pratique permettant d'évaluer l'impact de la prise de décision automatisée (ADM) sur les droits fondamentaux, est essentiellement centrée sur le Règlement général sur la protection des données (RGPD). Dans la mesure où cette méthodologie est une nouvelle approche de l'évaluation de l'ADM, elle se prête à une éventuelle généralisation en vue d'élaborer un modèle élargi, qui, au-delà du RGPD, serait applicable à l'AIDHDE. D'autres analyses d'impact ayant trait à l'IA ont été élaborées, notamment la Liste d'évaluation pour une IA digne de confiance conçue par le Groupe d'experts indépendants de haut niveau sur l'IA de l'Union européenne, dans le cadre, plus large, de ses Lignes directrices en matière d'éthique pour une IA digne de confiance.

Nous examinerons tout d'abord les analyses d'impact sur les droits de l'homme à caractère général (i) puis celles qui se rapportent à l'IA (ii).

1. Analyses d'impact sur les droits de l'homme existantes

i) Documents et cadres généraux faisant référence aux analyses d'impact sur les droits de l'homme en général

Principes directeurs relatifs aux entreprises et aux droits de l'homme

Les Principes directeurs relatifs aux entreprises et aux droits de l'homme (PDNU⁸) ont été adoptés par les Nations Unies en 2011. Ils soulignent : a) l'obligation, pour les États, de respecter et protéger les droits de l'homme et de s'y conformer, b) le rôle important joué par les entreprises s'agissant d'assurer la conformité avec les lois et les droits de l'homme, et c) la nécessité de mettre en place une protection juridique et des voies de recours judiciaires en cas d'impact négatif sur les droits de l'homme.

Identification et évaluation des risques pour les droits de l'homme liés à l'utilisation finale

Ce document⁹, élaboré par le Haut-Commissariat aux droits de l'homme, s'adresse aux cadres des entreprises de technologie qui souhaitent comprendre les attentes fondamentales des PDNU s'agissant d'identifier et d'évaluer les risques que présentent des produits et des services pour les droits de l'homme. À cet égard, les entreprises sont censées, selon les PDNU, 1) fournir un aperçu général des impacts possibles, 2) s'attacher en priorité aux préjudices les plus graves, et 3) prendre contact et communiquer utilement avec les parties prenantes.

Principes directeurs de l'OCDE à l'intention des entreprises multinationales

⁸ https://www.ohchr.org/documents/publications/guidingprinciplesbusinessshr_fr.pdf.

⁹ <https://www.ohchr.org/Documents/Issues/Business/B-Tech/identifying-human-rights-risks.pdf>.

Ce document de l'OCDE¹⁰ (2011) a pour objet d'adresser des recommandations aux entreprises multinationales exerçant leurs activités dans les pays adhérents ou à partir de ces derniers. Il contient des principes et des normes destinés à favoriser une conduite raisonnable des entreprises dans un environnement mondialisé, en conformité avec le droit international. Il demande aux États de promouvoir des valeurs ayant trait au respect de règles juridiques par les acteurs privés. En particulier, ces principes directeurs appellent les entreprises à se conformer à des règles juridiques relatives 1) aux droits de l'homme, 2) à l'emploi et à la réglementation du secteur de l'industrie, 3) à l'environnement, 4) à l'antitrust, 5) à l'intérêt des consommateurs et 6) à la fiscalité.

ii) Documents et cadres spécifiques ayant trait à l'IA

Décoder l'intelligence artificielle : 10 mesures pour protéger les droits de l'homme

La Recommandation « Décoder l'intelligence artificielle : 10 mesures pour protéger les droits de l'homme »¹¹ a été publiée par la Commissaire aux droits de l'homme du Conseil de l'Europe en 2019¹². Elle définit un ensemble de recommandations visant à prévenir et à atténuer les effets négatifs de l'intelligence artificielle sur les droits de l'homme. Elle porte sur 10 domaines d'action : 1) Évaluation de l'impact sur les droits de l'homme, 2) Consultations publiques, 3) Obligation, pour les États, de faciliter la mise en œuvre des normes en matière de droits de l'homme dans le secteur privé, 4) Information et transparence, 5) Contrôle indépendant, 6) Non-discrimination et égalité, 7) Protection des données et respect de la vie privée, 8) Libertés d'expression, de réunion et d'association, et droit au travail, 9) Possibilités de recours, 10) Promotion de la connaissance et de la compréhension de l'intelligence artificielle.

Recommandation CM/Rec(2020)1 du Comité des Ministres aux États membres sur les impacts des systèmes algorithmiques sur les droits de l'homme

Cette Recommandation¹³ a pour objet de guider les États et les acteurs privés dans les actions qu'ils mènent en lien avec la conception et le développement des systèmes algorithmiques et de veiller à ce que les droits de l'homme et les libertés individuelles de la Convention européenne des droits de l'homme soient protégés contre les évolutions technologiques. Ce document préconise, entre autres recommandations, que 1) les États devraient revoir leur cadre juridique pour l'adapter au contexte technologique, et 2) les acteurs privés devraient se conformer à la législation et respecter les droits de l'homme conformément aux dispositions des PDNU.

Lignes directrices en matière d'éthique pour une IA digne de confiance

Les Lignes directrices en matière d'éthique pour une IA digne de confiance ont été établies par le Groupe d'experts indépendants de haut niveau sur l'intelligence artificielle de la Commission

¹⁰ <https://www.oecd.org/fr/daf/inv/mne/2011102-fr.pdf>.

¹¹ Proposition d'Access Now.

¹² <https://rm.coe.int/decoder-l-intelligence-artificielle-10-mesures-pour-protoger-les-droit/168094b6e2>.

¹³ https://search.coe.int/cm/pages/result_details.aspx?ObjectId=09000016809e1124.

européenne (2019)¹⁴. Ce document fixe les caractéristiques d'une intelligence artificielle digne de confiance. Selon le groupe d'experts, une intelligence artificielle digne de confiance doit être 1) licite, 2) éthique et 3) robuste. De plus, les lignes directrices définissent un certain nombre d'exigences tirées des droits de l'homme que les systèmes d'intelligence artificielle doivent satisfaire pour être dignes de confiance, parmi lesquelles la protection des données à caractère personnel, la garantie de transparence et l'obligation de rendre des comptes. Pour déterminer si un système d'IA remplit ces exigences, une Liste d'évaluation pour une IA digne de confiance a été intégrée aux lignes directrices. Cette liste a vocation à aider les développeurs et les acteurs du déploiement de l'IA à évaluer et à améliorer la conformité de leurs systèmes avec les exigences énoncées dans les lignes directrices.

Examen de la boîte noire

Le document *Examining the black box*, élaboré par l'Institut Ada Lovelace¹⁵, fournit une explication des expressions « audit algorithmique » et « analyse d'impact algorithmique » et dresse un état des lieux de la recherche et de la pratique. En ce qui concerne les audits algorithmiques, l'Institut identifie deux grandes approches : 1) l'audit des biais : approche ciblée consistant à déterminer si un système algorithmique comporte des biais, et 2) l'inspection réglementaire : étude d'envergure portant sur la conformité d'un système algorithmique avec une réglementation ou des normes, nécessitant un certain nombre d'outils et de méthodes et généralement réalisée par des autorités de réglementation ou des spécialistes de l'audit. Pour ce qui est des analyses d'impact algorithmique, l'Institut identifie également deux approches : 1) l'évaluation des risques algorithmiques : évaluation des impacts sociétaux possibles d'un système algorithmique avant sa mise en service, et 2) l'évaluation de l'impact algorithmique : évaluation des impacts sociétaux possibles d'un système algorithmique sur les usagers ou la population qu'il touche, après sa mise en service.

Boîte à outils pour l'équité algorithmique

L'Union américaine des libertés civiles (American Civil Liberties Union) a conçu une boîte à outils pour l'équité algorithmique connue sous le nom d'AEKit (*Algorithmic Equity Toolkit*)¹⁶. Ce kit rassemble quatre composants conçus pour identifier les technologies de surveillance et de prise de décision utilisées par les pouvoirs publics, expliquer le fonctionnement de ces technologies et aider les citoyens à poser des questions sur leurs effets, leur efficacité et les contrôles auxquels elles sont soumises. Ces composants sont 1) un arbre de décision, 2) une cartographie du système, 3) un formulaire à trous et 4) un questionnaire. Ils sont destinés à être utilisés dans cet ordre, mais peuvent aussi être appliqués comme l'utilisateur le souhaite.

Boîte à outils sur les algorithmes et l'éthique

GovEx (municipalité et comté de San Francisco), Harvard DataSmart et Data Community DC

¹⁴ <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>.

¹⁵ <https://www.adalovelaceinstitute.org/wp-content/uploads/2020/04/Ada-Lovelace-Institute-DataKind-UK-Examining-the-Black-Box-Report-2020.pdf>.

¹⁶ <https://www.aclu-wa.org/AEKit>.

ont mis au point une boîte à outils sur les algorithmes et l'éthique appelée Ethics & Algorithms Toolkit¹⁷. Ce kit s'adresse à toute personne qui conçoit ou achète des algorithmes utilisés dans l'administration. À travers une série de questions, il aide l'utilisateur d'un algorithme à 1) comprendre les risques éthiques que pose cet usage et à 2) déterminer ce qu'il peut faire pour réduire ces risques le plus possible. Cette boîte à outils doit être appliquée chaque fois qu'un algorithme est utilisé pour éclairer une décision dans le secteur public.

Outil canadien d'évaluation de l'incidence algorithmique

L'outil d'évaluation de l'incidence algorithmique est un outil d'évaluation des risques obligatoire, qui est destiné à soutenir la Directive sur la prise de décisions automatisée du Conseil du Trésor du Canada¹⁸. Il se présente sous la forme d'un questionnaire, composé de 48 questions sur les risques et de 33 questions sur l'atténuation, qui permettent de déterminer le niveau d'incidence d'un système de décision automatisé. Cette évaluation est organisée selon des considérations touchant aux politiques, à l'éthique et au droit administratif du gouvernement en ce qui a trait aux secteurs de risque du système de décision automatisé, tels qu'établis dans le cadre de consultations menées par le Secrétariat du Conseil du Trésor du Canada auprès du milieu universitaire, de la société civile et d'autres institutions publiques. L'outil est conçu pour aider les ministères et les organismes à mieux comprendre et gérer les risques associés aux systèmes de décision automatisés. L'évaluation doit être effectuée au début de la phase de conception d'un projet et les résultats publiés dans un format accessible et dans les deux langues officielles du Canada.

Évaluations de l'impact algorithmique : un cadre pratique pour la responsabilisation des organismes publics

AI Now a mis au point un modèle d'évaluations de l'impact appelé « Algorithmic Impact Assessments : A Practical Framework for Public Agency Accountability » (Évaluations de l'impact algorithmique : un cadre pratique pour la responsabilisation des organismes publics)¹⁹. Dans son article, AI Now adresse les recommandations suivantes : 1) les organismes publics devraient procéder à une auto-évaluation des systèmes de décision automatisés existants et proposés, en évaluant leurs répercussions possibles sur l'équité, la justice, les préjugés ou d'autres paramètres dans les collectivités visées ; 2) les organismes publics devraient définir des processus utiles de contrôle par des chercheurs externes afin de découvrir, de mesurer ou de suivre les impacts dans le temps ; 3) les organismes publics devraient fournir au public des informations sur leur définition des « systèmes de décision automatisés », sur les systèmes existants et proposés et sur tous les processus connexes d'auto-évaluation et de contrôle par des chercheurs, et ce avant l'acquisition d'un système ; 4) les organismes publics devraient solliciter les commentaires du public pour répondre aux inquiétudes et aux questions en suspens ; et 5) les pouvoirs publics devraient mettre en place de meilleurs mécanismes de garantie d'une procédure régulière pour permettre aux personnes ou aux communautés

¹⁷ <https://ethicstoolkit.ai/>.

¹⁸ <https://www.canada.ca/fr/gouvernement/systeme/gouvernement-numerique/innovations-gouvernementales-numeriques/utilisation-responsable-ai/evaluation-incidence-algorithmique.html>.

¹⁹ <https://ainowinstitute.org/aiareport2018.pdf>.

concernées de remettre en question les évaluations insuffisantes ou les utilisations inéquitables, biaisées ou autrement préjudiciables du système que les organismes publics n'ont pas réussi à limiter ou à corriger.

Outre les recommandations fournies par les institutions susmentionnées, il convient de rappeler que différents articles scientifiques abordent également la question de l'évaluation de l'impact des systèmes d'IA sur les droits de l'homme²⁰.

B. Recherche des caractéristiques pertinentes pour la conception d'un modèle d'AIDHDE

Les systèmes d'IA actuels présentent des caractéristiques spécifiques qu'il convient de prendre en compte pour concevoir un modèle d'AIDHDE, en particulier lorsque les cadres généraux d'AIDH servent de point de départ. Ces caractéristiques seront examinées dans la première partie de la section (1). En outre, l'utilisation de l'IA peut avoir des incidences sur les valeurs de démocratie et d'État de droit. Ces aspects du modèle d'évaluation des systèmes d'IA sont donc tout aussi importants. Leur intégration dans le modèle pose des défis conceptuels, que nous analyserons en détail dans une section distincte (2).

1. Les grandes caractéristiques des systèmes d'IA en tant que variables d'évaluation

Comme l'explique le CAHAI dans son Étude de faisabilité, si les systèmes d'IA peuvent apporter aux particuliers et aux entreprises un certain nombre d'avantages, ils peuvent aussi menacer les droits de l'homme, la démocratie et l'État de droit (DHDE). Toute décision technique résulte d'un compromis entre, d'un côté, la valeur que le système technologique peut apporter et, de l'autre, les préjudices qu'il peut concrètement produire. Une approche plus complète et pleinement pertinente serait donc de mettre en balance et de prendre en compte à la fois la valeur ajoutée du système d'IA et la probabilité et l'ampleur de ses éventuelles atteintes aux DHDE²¹.

Lors de l'évaluation de la probabilité et de l'ampleur des risques associés à un système d'IA et de la classification du système selon ces paramètres, les éléments suivants pourraient, au minimum, être pris en considération :

- i) le contexte dans lequel le système d'IA est utilisé ;
- ii) sa technologie sous-jacente, en examinant divers aspects comme la finalité, la fiabilité, la traçabilité, l'explicabilité, les données d'entrée, le degré d'automatisation, la sécurité et l'accessibilité du système ;
- iii) les acteurs en jeu et l'étape de développement du système ;

²⁰ Voir par exemple, Alessandro Mantelero, « AI and Big Data: A blueprint for a human rights, social and ethical impact assessment », *Computer Law and Security Review*, Vol. 34, n° 4, août 2018, pp. 754-772, ainsi que Jacob Metcalf, Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh et Madeleine Clare Elish, 2021, « Algorithmic Impact Assessments and Accountability: The Co-construction of Impacts », dans *ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, 3-10 mars 2021, Virtual Event, Canada.ACM.

²¹ Suggéré par ES.

iv) les parties prenantes à l'évaluation.

Dans l'échelle de cotation du rapport bénéfices-risques, toute AIDHDE devrait être guidée par le principe préventif de l'absence de préjudice : si l'analyse d'impact identifie un risque élevé pour les droits de l'homme, la démocratie ou l'État de droit (DHDE) qui ne peut pas être atténué immédiatement, le système d'IA ne devrait jamais être développé, mis en œuvre ni utilisé en l'état²² par une quelconque autorité publique ou privée, au moins tant que des mesures efficaces n'ont pas été adoptées pour prévenir les risques potentiels ou supplémentaires en matière de DHDE (interdictions/moratoires frappant les systèmes d'IA à haut risque²³). Des mesures préventives devraient être mises en place pour les applications d'IA présentant le risque le plus élevé en termes d'échelle, de gravité et d'irrémediabilité et la valeur ajoutée ou le bénéfice le plus faible²⁴.

En plus du principe préventif d'absence de préjudice, toute AIDHDE devrait aussi respecter le principe de proportionnalité. En particulier, il convient d'évaluer si le système d'IA doit faire l'objet ou non d'une AIDHDE complète et détaillée. Ce point est important pour garantir la proportionnalité des cadres juridiques²⁵.

i) Le contexte de l'application en tant que variable de l'analyse d'impact

Une application donnée d'un système d'IA peut présenter un risque élevé pour les droits de l'homme, l'État de droit et la démocratie, alors que le même système, configuré pour une autre application, entraîne un risque moindre, voire nul²⁶. Il faut donc s'interroger sur le contexte géopolitique, social ou économique dans lequel le système d'IA sera exploité. De même, un système d'IA peut apporter une valeur et des avantages dans un certain contexte, qui pourraient être inversés dans un autre.

Lorsque l'on examine le contexte d'une application, il est important de tenir compte de la finalité du système annoncée par le concepteur ou le développeur ou par le client dans sa demande²⁷. Considérons l'exemple d'un système d'IA destiné à déceler des schémas de transaction financière pour signaler d'éventuelles opérations de blanchiment d'argent. Le concepteur ou le développeur a créé le système pour qu'il effectue une tâche précise, en l'occurrence détecter des opérations de blanchiment. Mais l'exploitant peut modifier la finalité du système et donc le contexte qui sera pertinent pour l'évaluation des risques. C'est le cas notamment lorsque

²² Voir les Lignes directrices en matière d'éthique pour une IA digne de confiance, p. 25 : « Lorsqu'aucun arbitrage acceptable du point de vue éthique ne peut être déterminé, la mise au point, le déploiement et l'utilisation du système d'IA ne devraient pas se poursuivre en l'état » (référence suggérée par NS).

²³ Repris de CONGO, *Guidelines for impact assessment*.

²⁴ En réponse à la demande d'éclaircissement de la FRA au sujet des « premières mesures ».

²⁵ Suggéré par EE.

²⁶ Voir également OCDE, *OECD Framework for the Classification of AI Systems – Preliminary Findings* [Cadre de l'OCDE pour la classification des systèmes d'IA. Résultats préliminaires], pp. 9-15, non encore publié, ainsi que la Liste d'évaluation pour une IA digne de confiance élaborée par le Groupe d'experts indépendants de haut niveau sur l'intelligence artificielle de l'UE dans le cadre des Lignes directrices en matière d'éthique pour une IA digne de confiance, en particulier ce qui concerne l'examen des résultats produits par les applications d'IA : à cet égard, il conviendrait de déterminer si le système produit un conseil, une décision ou un autre type de résultat (page 33).

²⁷ Suggéré par EE.

l'exploitant du système au sein d'une Direction du Trésor modifie le flux de données d'entrée, en présentant au système non pas des transactions de capitaux suspects, mais des transactions régulières, dont l'analyse livrera la stratégie commerciale et financière adoptée par des sociétés et des particuliers pour répartir et réaffecter leurs capitaux. Ainsi, la finalité du système d'IA a été modifiée et consiste maintenant à détecter des transactions commerciales et des stratégies de flux de trésorerie.

Il est tout aussi important de déterminer qui sont les utilisateurs²⁸ du système d'IA. Reprenons l'exemple du système de prévention du blanchiment de capitaux. Ce système, conçu pour les agents du Trésor et les cellules de renseignement financier, présentera un contexte de risques différent s'il est destiné à des spécialistes des données d'entraînement travaillant dans un laboratoire sur les politiques publiques.

ii) Les technologies sous-jacentes des systèmes d'IA en tant que variable de l'analyse d'impact

Si le contexte de l'application permet de mieux évaluer le niveau de risque du système d'IA, il est essentiel que l'évaluation des risques tienne aussi compte du type de technologie sous-jacente. Le domaine de l'intelligence artificielle comprend en effet différentes approches, depuis les processus de pensée et de raisonnement jusqu'au comportement, et dans les deux cas avec différentes façons de mesurer la réussite (fidélité au fonctionnement humain ou fonctionnement idéal ou rationnel)²⁹. De plus, ces approches sont mises en œuvre au moyen de différentes techniques³⁰ dont la liste évolue en permanence³¹, et il n'est pas rare de voir que certaines sont utilisées conjointement pour produire des systèmes d'IA plus complexes (par exemple, pour produire un système de détection de la fausse monnaie, on utilise un réseau neuronal antagoniste génératif constitué de deux réseaux neuronaux, le premier s'appuyant sur un modèle discriminatoire et le second sur un modèle probabiliste). Pour que le cadre d'AIDHDE reste pertinent par rapport à ces techniques d'IA en évolution permanente, la méthodologie d'analyse d'impact devrait rester aussi « algorithmiquement neutre » que possible³², tout en ayant une connaissance approfondie des différences et des incidences de chacune de ces techniques.

Le type de technologie sur lequel repose le système d'IA est important, car il détermine huit dimensions préliminaires susceptibles de nous alerter sur des risques potentiels.

²⁸ En fait, selon l'Étude de faisabilité du CAHAI, tout sujet pertinent du mécanisme de conformité est une source de contexte. Cf. section 9.2 de l'Étude de faisabilité du CAHAI.

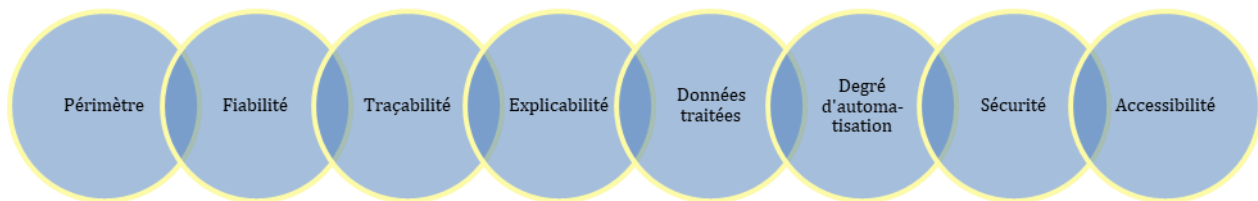
²⁹ Russell et Norvig, 2015. (Référence suggérée par ES).

³⁰ Chowdhary, 2020.

³¹ Voir l'Étude de faisabilité, chapitre 8.

³² Suggéré par le Royaume-Uni et Access Now.

Les huit dimensions d'une technologie d'IA signalant des risques potentiels



La première dimension est le périmètre. Prenons comme exemple un système d'IA utilisant un algorithme d'apprentissage supervisé entraîné avec des données réelles. Ce système présente une marge d'erreur de sur-classification ou sous-classification de sa prédiction cible³³. Les performances de la technologie employée dépendent de l'ensemble d'apprentissage, de l'algorithme utilisé pour classer les exemples d'apprentissage, de l'ensemble des données de contrôle servant à ajuster les cycles d'apprentissage – que l'ajustement repose sur des données nouvellement collectées ou sur des données générées par l'algorithme en déploiement réel³⁴ – et aussi du scénario réel du système auquel la technologie est appliquée : certains processus sont plus chaotiques et moins prévisibles que d'autres³⁵. Dans notre exemple, il est intéressant de connaître le fonctionnement interne de la technique d'apprentissage et le type d'algorithme pour interpréter les résultats du système d'IA, mais aussi pour passer au crible l'acquisition des données ainsi que l'élagage et les biais éventuels. S'agissant de cette première dimension, le périmètre, on distingue deux cas préoccupants. Le premier correspond à l'utilisation d'un système d'IA en dehors du périmètre de sa technologie sous-jacente³⁶. Par exemple, un système d'IA qui présenterait les chances statistiques d'un ensemble de tests de preuves médico-légales produit au cours d'un procès comme le fondement des chances qu'un individu donné soit coupable (ce que l'on appelle le « sophisme du procureur »³⁷). Le deuxième cas est celui d'un système qui, tout en affichant un taux d'erreur faible, ferait une erreur fondamentale³⁸. Par exemple, un système d'IA qui effectuerait un séquençage d'ADN brut avec un faible taux d'erreur³⁹, mais qui pourrait faire une erreur fondamentale avec un médicament expérimental cible produisant un effet indésirable.

La deuxième dimension est la fiabilité⁴⁰. La capacité d'un système d'IA à fournir

³³ Algorithmes de classification utilisés pour modéliser les problèmes de modélisation prédictive de classification. Les algorithmes de modélisation prédictive de classification sont évalués à partir de leurs résultats. Le taux de succès de la classification est une métrique couramment utilisée pour évaluer la performance des modèles fondés sur les prédictions des étiquettes de classe. Sinon, au lieu des étiquettes de classe, certaines tâches peuvent nécessiter la prédiction d'une probabilité d'appartenance à une classe pour chaque exemple. (Éclaircissement demandé par ES).

³⁴ Suggéré par la FRA.

³⁵ Suggéré par le Royaume-Uni.

³⁶ Flynn et al., 2020, <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7156005/>.

³⁷ Sesardic, 2008, <https://philpapers.org/archive/DEMGBS.pdf>.

³⁸ Suggéré par le Royaume-Uni.

³⁹ Tel que le programme PHRED : <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC310698/>.

⁴⁰ Conceptuellement, la fiabilité est proche de la sûreté de fonctionnement. Le premier concept correspond à la probabilité que le système d'IA fournira les services correctement (comportement attendu), conformément à ce que

systématiquement un résultat correspondant aux attentes peut être déterminée à partir de la technologie sous-jacente⁴¹. Ainsi, l'utilisation d'algorithmes adaptatifs dans un système d'IA chargé d'organiser les services publics d'une ville intelligente est susceptible d'évoluer dans le temps, car c'est précisément ce pour quoi ces algorithmes ont été choisis. Un algorithme déterminatif, basé sur des règles est fiable si l'infrastructure qui met en œuvre les systèmes d'IA fonctionne comme prévu et ne subit pas de corruption de données. Et même si les technologies n'ont pas une fiabilité universelle – car cette caractéristique varie selon les groupes de personnes –, une mesure de la fiabilité ou de la confiance est utile pour évaluer le risque. « Outre les algorithmes [de réseaux neuronaux profonds] de haute précision, il existe un besoin important de systèmes d'apprentissage automatique robustes et d'architectures matérielles capables de produire des résultats fiables et dignes de confiance en cas de défaillance matérielle tout en préservant la sécurité et la confidentialité⁴². » On pourra tolérer une légère imprécision dans un système de traduction linguistique à base d'IA (traitement du langage naturel, TLN), mais la marge de tolérance n'est pas aussi grande dans le cas d'un véhicule à conduite autonome⁴³.

La troisième dimension est la traçabilité. Le résultat produit par un système d'IA peut être traçable du point de vue de l'architecture du système. Plus précisément, la traçabilité suppose que l'on puisse déterminer non seulement comment le système a fonctionné, mais aussi comment il a été créé et dans quel but, de façon à expliquer pourquoi il possède telle dynamique ou tel comportement⁴⁴. La traçabilité est un prédicat qui accepte la granularité. Si un système d'IA repose sur une technologie qui peut être ajustée à la sortie près (autrement dit, si l'on sait pourquoi et à quelles étapes du traitement le système d'IA a produit une sortie ou une décision donnée), alors le niveau de traçabilité permet d'expliquer la sortie. Cela dit, l'utilisation de multiples réseaux neuronaux conjuguée à de grandes quantités de données d'apprentissage produit des systèmes d'IA dont le traçage étape par étape devient très coûteux. On a alors un faible niveau de traçabilité et donc d'explicabilité⁴⁵. De plus, l'IA est souvent développée à partir d'anciennes versions d'autres logiciels, ce qui peut compliquer la recherche de la cause d'un impact donné⁴⁶.

La quatrième dimension est l'explicabilité⁴⁷. Le fait qu'un système d'IA puisse, dans une certaine mesure, être traçable au niveau de ses processus/sorties est une condition propre à

prévoit le concepteur/le développeur/l'exploitant/l'utilisateur. Le second concept mesure la confiance que le concepteur/le développeur/l'exploitant/l'utilisateur a dans le système. Une façon d'obtenir la confiance est de faire en sorte que le système soit fiable. Voir O'Regan, 2017, https://link.springer.com/chapter/10.1007/978-3-319-57750-0_11.

⁴¹ Hong et al., 2021, <https://arxiv.org/abs/2102.01740>.

⁴² Hanif et al., 2018, <https://ieeexplore.ieee.org/document/8474192>.

⁴³ Certains ont proposé, par exemple, de mettre en place des mesures de la confiance pour les systèmes critiques tels que l'IA militaire ; voir Jah et al., 2019 : <https://papers.nips.cc/paper/9355-attribution-based-confidence-metric-for-deep-neural-networks>.

⁴⁴ Kroll, 2021, <https://arxiv.org/abs/2101.09385>.

⁴⁵ Felzmann et al., 2020, <https://link.springer.com/article/10.1007/s11948-020-00276-4>.

⁴⁶ En règle générale, les nouvelles versions ne devraient pas influencer sur ce problème. Les difficultés apparaissent lorsqu'un système boucle dans un autre et ressort, généralement pour des raisons logistiques (et non mathématiques). Dans ce cas, l'intervention de différentes versions peut compliquer les choses (observation du Royaume-Uni).

⁴⁷ Proposé par CINGO.

faciliter l'explication de son comportement. Toutefois, l'explicabilité comporte une dimension de communication. Ainsi, un système d'IA peut être explicable à des spécialistes de l'IA, mais pas au grand public et à la société civile dans son ensemble⁴⁸.

La cinquième dimension correspond au type de données extraites et traitées par les systèmes d'IA. Il va de soi que des données sensibles à caractère privé peuvent être plus problématiques que, par exemple, des données sur les émissions de gaz à effet de serre. Il en va de même de l'utilisation de contenus ou de données simplement formatées ou encore de données contenant des informations à caractère personnel⁴⁹. Les jeux de données de très grande taille⁵⁰ ont des propriétés statistiques (et donc des impacts potentiels) différentes des jeux de données plus restreints⁵¹. Cette dimension appelle une réflexion sur la nécessité de disposer de jeux de données diversifiés et de reconnaître qu'il est difficile de créer des ensembles de données non discriminatoires et suffisamment diversifiés, car les préjugés et la discrimination sont ancrés dans la société⁵².

La sixième dimension correspond au niveau d'automatisation⁵³ du système d'IA. L'automatisation du système doit être examinée en lien étroit avec la dimension « technologie ». Par exemple, un système d'IA entièrement autonome utilisé pour étalonner une machine peut être hautement prévisible et très fiable et dépasser le niveau de précision d'un opérateur⁵⁴. S'il n'est pas totalement automatisé, mais tributaire d'un opérateur humain, le risque d'erreur humaine est plus grand. Et même lorsque l'intervention d'un humain reste nécessaire, celui-ci peut se fier aux résultats produits par le système et prendre une décision sans faire sa propre évaluation⁵⁵. La littérature fait état de cas où des opérateurs humains ne sont pas intervenus pour éviter un préjudice ou un effet catastrophique parce qu'un système automatisé était présent⁵⁶. Des « humains dans la boucle » ont déjà été tenus responsables alors même qu'ils n'étaient pas en mesure de voir les informations importantes sur l'interface du système qu'ils exploitaient ou parce qu'ils n'avaient pas pu prévenir la défaillance du système automatisé. Dans ces cas de figure, les humains fonctionnent comme des pare-chocs moraux à des fins de responsabilisation⁵⁷.

⁴⁸ Les travaux menés par Umang Bhatt et al. montrent que les systèmes d'IA les plus explicables sont destinés au « débogage » et s'adressent donc aux développeurs d'IA par opposition aux utilisateurs de l'IA et à la société civile : <https://arxiv.org/abs/1909.06342>. (Référence fournie par CINGO.)

⁴⁹ Observation de la FRA.

⁵⁰ Bien que l'expression « métadonnées » soit utilisée de façon informelle pour désigner de gros volumes de données ou des flux de données à grande vitesse, ce terme est encore flou, même pour les praticiens et les spécialistes des données. Voir Favaretto M. et al., 2020, <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0228987>. Certains s'emploient à conceptualiser ce terme de manière plus précise, notamment M. Al-Mekhlal et A. Ali Khwaja, 2019, <https://ieeexplore.ieee.org/abstract/document/8919591>.

⁵¹ Suggéré par le Royaume-Uni.

⁵² Suggéré par CINGO.

⁵³ Suggéré par la FRA.

⁵⁴ Observation d'ES.

⁵⁵ Cf. Elish, 2019, <https://estsjournal.org/index.php/ests/article/view/260>.

⁵⁶ Cf. Elish, 2019, <https://estsjournal.org/index.php/ests/article/view/260>.

⁵⁷ Les travaux du Groupe d'experts indépendants de haut niveau sur l'intelligence artificielle sont très pertinents à cet égard, plus précisément la Liste d'évaluation pour une IA digne de confiance (ALTAI), qui comprend une section sur le contrôle humain (page 8). Cette liste souligne qu'il est important d'évaluer si le système fonctionne de manière autonome ou s'il autorise un « humain dans la boucle » (capacité d'intervention humaine dans chaque cycle de décision du système), un « humain sur la boucle » (capacité d'intervention humaine dans le cycle de conception du

La septième dimension concerne la sécurité des systèmes d'IA. Plus le risque de piratage, d'attaques malveillantes ou d'autres incidents de sécurité (effets secondaires négatifs de l'apprentissage par renforcement par exemple⁵⁸) est élevé, plus les risques d'effets non désirés sont potentiellement importants.

La huitième dimension est l'accessibilité de la technologie⁵⁹. Utilisé dans un contexte social marqué par de larges fossés numériques, ce type de technologie ne peut que renforcer l'exclusion technologique des groupes marginalisés. Ces derniers sont souvent exclus de la technique par manque d'accessibilité ou d'inclusion numérique.

Convergence avec les travaux récents de l'OCDE sur un Cadre pour la classification des systèmes d'IA

Il est intéressant de constater que bon nombre des éléments mentionnés ci-dessus convergent avec ceux envisagés par l'OCDE dans son Cadre pour la classification des systèmes d'IA, qui est en cours d'élaboration.

L'OCDE a élaboré plusieurs questions types en matière de transparence, d'explicabilité et de robustesse⁶⁰. Ainsi, si l'explicabilité et la transparence ne sont pas assurées, il n'est pas possible de traiter avec les parties prenantes pour réfléchir à l'impact de telle application d'IA sur les droits de l'homme. Il est également important d'évaluer si les capacités et les limites d'un système d'IA ont été communiquées aux utilisateurs (finaux)⁶¹. Par ailleurs, la manière dont la qualité des données est assurée compte aussi. Par exemple, la façon dont les données sont recueillies – par des systèmes automatisés ou par des humains –, l'ampleur de la collecte et la nature dynamique des données peuvent influencer sur leur qualité en vue d'une utilisation par des systèmes d'IA⁶². En outre, le format et la structure des données sont importants, de même que les droits et l'identifiabilité des personnes auxquelles les données font référence⁶³. Le type de modèle utilisé pour développer l'IA et la manière dont elle est construite entrent aussi en ligne de compte⁶⁴. Il importe de noter à cet égard que les applications d'IA sont souvent conçues au moyen de systèmes composites, ce qui peut être

système et la surveillance du fonctionnement du système) ou un « humain aux commandes » (capacité de contrôle de l'activité globale du système d'IA [y compris de ses incidences économiques, sociétales, juridiques et éthiques au sens large] et la faculté de décider quand et comment utiliser le système dans une situation donnée). Dans ce contexte, il est utile d'observer s'il existe un mécanisme de détection et d'intervention en cas de résultats ou de fonctionnement non désirés du système d'IA, si le système peut être arrêté et, en cas d'apprentissage autonome, s'il existe une supervision de ce processus.

⁵⁸ Suggéré par CINGO.

⁵⁹ Proposé par CINGO.

⁶⁰ OCDE, *OECD Framework for the Classification of AI Systems – Preliminary Findings* [Cadre de l'OCDE pour la classification des systèmes d'IA. Résultats préliminaires], p. 24, non encore publié.

⁶¹ p. 15.

⁶² OCDE, *OECD Framework for the Classification of AI Systems – Preliminary Findings* [Cadre de l'OCDE pour la classification des systèmes d'IA. Résultats préliminaires], pp. 16-18.

⁶³ OCDE, *OECD Framework for the Classification of AI Systems – Preliminary Findings* [Cadre de l'OCDE pour la classification des systèmes d'IA. Résultats préliminaires], pp. 18-21.

⁶⁴ Voir à ce propos la section I.1.B *supra*. Aussi : OCDE, *OECD Framework for the Classification of AI Systems – Preliminary Findings* [Cadre de l'OCDE pour la classification des systèmes d'IA. Résultats préliminaires], pp. 23-28.

gênant pour évaluer comment l'IA parvient à certains résultats, en particulier lorsque le modèle utilise un apprentissage non supervisé. En outre, il est important de savoir comment un modèle est construit⁶⁵. Enfin, les tâches et les objectifs du système d'IA sont aussi des éléments pertinents : le système est-il utilisé pour la détection d'événements, la prévision ou l'optimisation par objectifs⁶⁶ ? L'OCDE précise également qu'il existe plusieurs outils pour la réalisation de cette analyse technique⁶⁷.

iii) Les acteurs en jeu et l'étape de développement du système d'IA

Outre le contexte et la technologie, il convient aussi d'examiner deux autres paramètres lorsque l'on conçoit un modèle d'AIDHDE : d'une part, les acteurs en jeu et leur rôle vis-à-vis du système d'IA (a), d'autre part, l'étape de développement du système dans son cycle de vie (b).

a) Le rôle des acteurs vis-à-vis du système d'IA

L'AIDHDE devrait être un outil d'évaluation continue, utilisé tout au long⁶⁸ du cycle de vie du système d'IA. Dans ce cycle de vie interviennent divers acteurs qui jouent des rôles différents vis-à-vis du système : concepteurs, développeurs, distributeurs, exploitants et utilisateurs. De plus, certains acteurs sont aux commandes du système ou affichent une fonction de commande des performances du système (« humain aux commandes », « humain sur la boucle », « humain dans la boucle »). Mais le plus souvent, ce ne sont pas des humains qui sont aux commandes des systèmes d'IA. En effet, un système peut être commandé par d'autres systèmes d'IA ou être en relation avec un système non automatisé⁶⁹.

b) L'étape du système dans son cycle de vie

Le deuxième paramètre est l'étape du système dans son cycle de vie, ce dernier étant lié au type de technologie d'IA utilisé. Il existe de nombreux processus d'ingénierie pour bâtir un système d'IA⁷⁰. Ainsi, pour ne citer qu'un exemple, affirmer que pour utiliser un modèle d'IA (un modèle prédictif d'apprentissage automatique), il suffit d'alimenter un prédicteur avec des données est une simplification grossière. La mise en œuvre d'un système d'IA suppose en effet un enchaînement d'opérations complexe : configuration, automatisation, collecte des données, vérification des données, identification des caractéristiques, essai et débogage, gestion des ressources, analyse du modèle, gestion des processus, gestion des métadonnées, mise en service et suivi⁷¹. L'évaluation des risques d'un système d'IA à l'étape du cycle de vie « mise en œuvre de l'IA » ne ressemble en rien à l'évaluation des risques aux étapes précédentes

⁶⁵ OCDE, *OECD Framework for the Classification of AI Systems – Preliminary Findings* [Cadre de l'OCDE pour la classification des systèmes d'IA. Résultats préliminaires], pp. 26 et 27.

⁶⁶ OCDE, *OECD Framework for the Classification of AI Systems – Preliminary Findings* [Cadre de l'OCDE pour la classification des systèmes d'IA. Résultats préliminaires], pp. 29 et 30.

⁶⁷ OCDE, *Tools for Trustworthy AI* [Outils pour une IA digne de confiance], p. 8, non encore publié.

⁶⁸ Suggéré par EE.

⁶⁹ Observation du Royaume-Uni.

⁷⁰ Observation d'ES.

⁷¹ Google, [ML Ops: Continuous delivery and automation pipelines in machine learning](#).

« délimitation du champ d'application-compréhension » (cas d'un système d'apprentissage automatique)⁷².

iv) Les parties prenantes

Pour que l'AIDHDE soit une réussite, il est essentiel de mobiliser la collectivité⁷³. Des mécanismes efficaces doivent être mis en place pour identifier les parties prenantes au sein des populations concernées et assurer une participation active au processus d'analyse. L'identification correcte des parties prenantes et l'existence de processus visant à les associer de manière ouverte et suffisamment informée et en se mettant à leur portée sont un facteur déterminant pour évaluer l'impact potentiel sur les droits de l'homme, la démocratie et l'État de droit⁷⁴.

2. La démocratie et l'État de droit en tant que dimensions d'un modèle d'analyse intégré des systèmes d'IA⁷⁵

Toute analyse d'impact des systèmes d'IA sur les droits de l'homme (AIDH) suppose une référence commune aux droits de l'homme (DH). Le fait d'y ajouter l'analyse de l'impact sur la démocratie et l'État de droit présente des difficultés considérables d'un point de vue méthodologique, difficultés qu'il convient d'examiner en détail (i) avant d'envisager une solution possible sur la base de critères fondés sur des droits (ii).

i) Difficultés méthodologiques de l'analyse d'impact des applications d'IA sur la démocratie et l'État de droit

Les droits de l'homme recouvrent généralement des valeurs et des biens communs concrets, qu'il convient de prendre en considération, par exemple la liberté d'expression (article 10 de la CEDH) ou le droit à un recours effectif (article 13 de la CEDH). En revanche, la « démocratie » n'est pas réductible (bien qu'elle s'y rattache) à un ensemble de droits ou de règles.

Le simple fait de promulguer des règles ne crée pas la démocratie, mais l'émergence de la démocratie nécessite un ensemble commun de règles. Plusieurs normes existent pour assurer la sauvegarde du processus démocratique, comme les règles d'accès et d'exercice relatives au pouvoir et aux décisions publiques, mais la démocratie comporte également une dimension culturelle, institutionnelle et sociale. Le Rapport de la Secrétaire Générale du Conseil de l'Europe sur la situation de la démocratie, des droits de l'homme et de l'État de droit note à cet égard que « [l]a démocratie ne se réduit pas à l'existence de lois et institutions. Celles-ci sont nécessaires mais non suffisantes. Le bon fonctionnement des démocraties dépend de ce que

⁷² Microsoft, Google et DataRobot, à tout le moins, s'accordent à reconnaître une étape de délimitation-compréhension dans la gestion du cycle de vie. <https://cloud.google.com/blog/products/ai-machine-learning/making-the-machine-the-machine-learning-lifecycle> ; <https://azure.microsoft.com/en-ca/blog/how-to-accelerate-devops-with-machine-learning-lifecycle-management/> ; <https://www.datarobot.com/wiki/machine-learning-life-cycle/>.

⁷³ Proposé par HomoDigitalis.

⁷⁴ Mise en avant de ces aspects suggérée par NS.

⁷⁵ L'ensemble de la section a grandement bénéficié des éclaircissements, nuances et recommandations apportés par NS.

l'on qualifie souvent de culture de la démocratie⁷⁶. » Le Rapport note également l'étroite interdépendance qui existe entre démocratie, État de droit et droits de l'homme, ce que soulignaient déjà plusieurs textes clés du Conseil de l'Europe⁷⁷.

Loin de se limiter à des élections libres et équitables, la démocratie suppose une participation ouverte, diversifiée et accessible, l'intégrité, une société civile active et une répartition équitable des pouvoirs entre les organes de l'État et les citoyens, lesquels vivent dans l'égalité et la dignité⁷⁸. La bonne gouvernance à tous les niveaux (local, régional et national) est également une composante essentielle de la démocratie⁷⁹, et on ne soulignera jamais assez le lien organique entre qualité de la démocratie et qualité de la gouvernance : un affaiblissement de la démocratie entraînera une moindre responsabilisation et une dégradation de la gouvernance ; une dégradation de la gouvernance entraînera à son tour le mécontentement des citoyens et donc un affaiblissement de la démocratie.

L'État de droit⁸⁰ aussi est une composante essentielle de la démocratie, comme le souligne également la Commission de Venise dans sa Liste des critères de l'État de droit⁸¹.

La Commission rappelle que dans le Préambule du Statut du Conseil de l'Europe, l'État de droit est mentionné comme étant l'un des « principes sur lesquels se fonde toute démocratie véritable », avec la liberté individuelle et la liberté politique. Selon elle, « la notion d'État de droit repose sur un droit sûr et prévisible, dans lequel toute personne [doit] être traitée par les décideurs de manière digne, égale et rationnelle, dans le respect du droit existant, et [...] disposer de voies de recours pour contester les décisions devant des juridictions indépendantes et impartiales, selon une procédure équitable. » Ses caractères⁸² sont : 1) la légalité, avec procédure d'adoption des textes de loi fondée sur la transparence, l'obligation de rendre compte et la démocratie ; 2) la sécurité juridique ; 3) l'interdiction de l'arbitraire ; 4) l'accès à la justice

⁷⁶ [Situation de la démocratie, des droits de l'homme et de l'Etat de droit : Un nouveau démocratique pour l'Europe](#), p. 149.

⁷⁷ Voir, entre autres, le Préambule de la Convention européenne des droits de l'homme.

⁷⁸ Page 6 du Rapport de la Secrétaire Générale.

⁷⁹ Voir à ce propos la page 66 du Rapport de la Secrétaire Générale, qui cite, entre autres critères de mesure d'une bonne gouvernance, l'efficacité et l'efficacité, la compétence, et l'efficacité et la responsabilité des institutions publiques.

⁸⁰ L'État de droit ne se limite pas à un ensemble de règles ; c'est aussi, en même temps, une situation sociale complexe qui nécessite des règles. D'une part, il suppose que les représentants de l'État internalisent le droit pour que chacun de leurs actes y soit conforme. De l'autre, les citoyens sont censés se plier aux prescriptions des mandats officiels qui sont produites conformément à une législation préalable. Acceptation de la loi et obéissance à la loi sont des conditions sociales, et non formelles, de l'État de droit.

⁸¹ Cf. [https://www.venice.coe.int/webforms/documents/default.aspx?pdffile=CDL-AD\(2016\)007-f](https://www.venice.coe.int/webforms/documents/default.aspx?pdffile=CDL-AD(2016)007-f).

⁸² Le Rapport 2020 de l'UE sur l'Etat de droit offre un point de vue comparable en indiquant que l'État de droit « garantit que toutes les autorités publiques agissent toujours dans les limites fixées par la loi, conformément aux valeurs de la démocratie et aux droits fondamentaux, et sous le contrôle de juridictions indépendantes et impartiales. L'État de droit est une notion qui recouvre des principes tels que la légalité, qui suppose l'existence d'une procédure d'adoption des textes de loi transparente, responsable, démocratique et pluraliste ; la sécurité juridique ; l'interdiction de l'arbitraire du pouvoir exécutif ; une protection juridictionnelle effective assurée par des juridictions indépendantes et impartiales, un contrôle juridictionnel effectif y compris le respect des droits fondamentaux ; la séparation des pouvoirs et l'égalité devant la loi. Ces principes ont été confirmés par la Cour de justice de l'Union européenne et par la Cour européenne des droits de l'homme. Par ailleurs, le Conseil de l'Europe a élaboré des normes et formulé des avis et des recommandations qui fournissent des orientations bien établies destinées à promouvoir et à faire respecter l'Etat de droit. » Voir <https://eur-lex.europa.eu/legal-content/FR/TXT/HTML/?uri=CELEX:52020DC0580&from=EN>.

devant des juridictions indépendantes et impartiales, avec contrôle juridictionnel des actes administratifs ; 5) le respect des droits de l'homme ; 6) la non-discrimination et l'égalité devant la loi. Cela étant, certains de ces principes couvrent un vaste champ, ce qui complique leur utilisation pour déterminer, en pratique, si un système d'IA donné aura des incidences sur la démocratie et l'État de droit⁸³.

Ces difficultés expliquent peut-être pourquoi, le plus souvent, les cadres d'AIDH actuels, de caractère général ou propres à l'IA, ne contiennent pas d'analyse d'impact sur la démocratie et l'État de droit. Seule la liste d'auto-évaluation du Groupe d'experts indépendants de haut niveau sur l'IA de la Commission européenne comporte une question générale sur la démocratie⁸⁴.

Il nous faut néanmoins résoudre ces difficultés pour proposer un modèle d'analyse complet, pratique et cohérent. Les effets de l'IA sur chacun⁸⁵ des droits de l'homme peuvent être très différents des effets de cette technologie sur la démocratie dans son ensemble⁸⁶. Il peut arriver, par exemple, qu'à des violations mineures de tel droit fondamental correspondent des menaces majeures au bon fonctionnement des institutions et processus démocratiques. D'où l'importance d'examiner les critères qui pourraient être utilisés dans une AIDHDE.

ii) Critères fondés sur des droits applicables à la démocratie et à l'État de droit

La démocratie et l'État de droit sont composés d'une multitude d'éléments, dont certains sont difficiles à relier à des paramètres clairement mesurables. Nous proposons ici une méthodologie pratique pour résoudre ce problème. Cette méthodologie consiste, tout d'abord, à utiliser les droits de l'homme comme des indicateurs indirects de la démocratie et de l'État de droit, autrement dit, à examiner si certaines violations des droits de l'homme pourraient être un moyen d'évaluer la situation de la démocratie et de l'État de droit. L'idée est d'étudier si l'ampleur de certaines violations de tel ou tel droit fondamental est à même de fragiliser les institutions et les pratiques sociales qui constituent la démocratie et l'État de droit. On peut aussi envisager d'examiner s'il existe un lien systémique entre la garantie et l'efficacité de certains droits de l'homme et les institutions garantes de la démocratie et de l'État de droit.

Par exemple, sans le droit de réunion, les citoyens seraient privés de la possibilité de se réunir publiquement, de délibérer et de forger une opinion publique, ou d'exprimer, de promouvoir et de défendre collectivement leurs intérêts par des actions pacifiques (par exemple en organisant des manifestations ou des mouvements de contestation). Si la démocratie n'est pas réductible à ce droit et aux possibilités qu'il offre, elle dépend néanmoins de la protection de ce droit, lequel pourrait donc être considéré comme l'un des multiples « indicateurs indirects de la démocratie » dans un modèle d'analyse d'impact.

Les droits de l'homme pouvant être considérés comme des indicateurs indirects de la

⁸³ En réponse à la demande d'éclaircissement d'ES au sujet des « difficultés conceptuelles » que présente l'extension de l'AIDH à la démocratie et à l'État de droit.

⁸⁴ Voir la page 20, qui propose seulement d'examiner si l'IA a une incidence sur la société en général.

⁸⁵ Suggéré par la FRA.

⁸⁶ Suggéré par la FRA.

démocratie sont le droit au respect de la vie privée et familiale (article 8 de la CEDH), le droit à la liberté de pensée (article 9 de la CEDH), la liberté d'expression (article 10 de la CEDH), la liberté de réunion et d'association (article 11 de la CEDH), le droit à la non-discrimination (article 14 de la CEDH), et le droit de voter et d'être élu dans le cadre d'élections libres et équitables (article 3 du Protocole additionnel à la CEDH).

Il est intéressant de noter que si l'on transpose ce principe dans le contexte de l'IA, certaines applications d'IA peuvent censurer des groupes particuliers en les empêchant d'exprimer leurs idées. Il s'agirait d'un cas extrême où le non-respect de la liberté d'expression serait un indicateur indirect de la force du discours public, laquelle est un facteur socio-institutionnel qui concourt au maintien des systèmes démocratiques. Un autre exemple, plus subtil, est l'utilisation de systèmes de ciblage par reconnaissance faciale ou de suivi des empreintes biométriques dans les manifestations publiques, les rassemblements politiques et les mairies. La simple utilisation de ces systèmes d'IA peut entraver le rassemblement spontané des citoyens, ceux-ci craignant d'être surveillés, enregistrés et suivis. Certains systèmes d'IA destinés à surveiller les manifestations politiques font peser un fardeau psychologique sur ceux qui veulent exercer leurs droits civils. L'érosion de la liberté de réunion a pour effet d'affaiblir un facteur social qui favorise la démocratie, à savoir la discussion et le soutien, de manière active et sur la place publique, des programmes, causes et idées politiques⁸⁷.

Un autre exemple de la manière dont les applications d'IA peuvent affaiblir les facteurs socio-institutionnels qui favorisent la démocratie concerne le processus de réflexion avant les élections. Les applications d'IA peuvent avoir une incidence sur la mise en place des institutions représentatives. C'est le cas notamment de certaines applications de vote. Ce type d'application a pour but de faciliter le choix des électeurs au moyen d'un quiz qui leur permet de comparer leurs opinions avec les positions politiques des différents partis en lice⁸⁸. Mais la manière dont les questions sont conçues (sous forme affirmative ou négative), les thèmes associés aux différents partis, voire l'interface des applications peuvent orienter les électeurs vers tel ou tel candidat⁸⁹. Cela pose un problème sur le plan sociologique, car lorsque les électeurs se fient à ces applications, leur capacité de décision est altérée et la légitimité collective des institutions élues s'en trouve diminuée. Un autre exemple concerne les applications d'IA destinées à donner un coup de pouce aux électeurs pour les orienter dans une certaine direction. Ces « applications de microciblage politique » servent essentiellement à diffuser des messages extrêmement circonscrits afin de maximiser l'intention de vote en faveur d'un candidat, d'un parti ou d'une cause. La conséquence collective pour la démocratie est la perte de la pluralité et de la diversité des informations dans le débat politique⁹⁰.

De même, l'État de droit peut être rattaché à certains droits de l'homme, de sorte que les violations de ces droits ou la diminution des garanties qui les protègent sont des indicateurs

⁸⁷ En réponse à un commentaire du Royaume-Uni.

⁸⁸ Cf. Cedroni et Garzia, 2010,

https://www.researchgate.net/publication/285054035_Voting_Advice_Applications_in_Europe_The_State_of_the_Art.

⁸⁹ Cf. Garzia et Marschall, 2019, à l'adresse

<https://oxfordre.com/politics/view/10.1093/acrefore/9780190228637.001.0001/acrefore-9780190228637-e-620>.

⁹⁰ Exemples suggérés par le CDDG.

indirects de l'impact sur l'État de droit. On trouve un cadre de référence fondé sur des droits dans la Liste des critères de l'État de droit adoptée par la Commission de Venise, qui mentionne les droits suivants : l'égalité devant la loi et la non-discrimination, l'accès à la justice devant des tribunaux indépendants et impartiaux, avec contrôle juridictionnel des actes administratifs, le droit à un procès équitable et le droit au respect de la vie privée et familiale, notamment la protection des données à caractère personnel.

Sur un plan pratique, ces indicateurs indirects peuvent être utilisés comme indiqué ci-après pour évaluer une application d'IA⁹¹.

La première étape consiste à identifier la technologie précise et/ou les applications potentielles de l'application d'IA à l'étude.

La deuxième étape consiste à identifier les droits de l'homme qui sont associés soit à la démocratie, soit à l'État de droit, et à les subdiviser en composantes. Examinons quelques exemples.

La liberté d'expression peut être subdivisée en plusieurs composantes. L'une d'elles, pertinente dans notre contexte, est la liberté de rechercher, de recevoir et de communiquer des informations et des idées de toute sorte. Elle exprime le droit de recevoir et communiquer librement des informations ou des idées. L'exercice de ce droit pourrait être affaibli par une application d'IA qui déclasserait ou démonétiserait du contenu sur la base de règles non collectivement approuvées⁹². Autre exemple, plus radical, celui d'un système d'IA qui aiderait à censurer des organes d'information indépendants, à cibler (via des attaques de type DDOS par exemple), harceler ou surveiller des journalistes ou des médias, à restreindre l'accès à des informations de l'État destinées au public et à limiter l'accès à l'internet en bloquant des sites web. Les composantes de ce droit peuvent servir de critères pour mesurer l'impact d'une application d'IA.

De même, la liberté de pensée peut être décomposée comme suit : le droit de ne pas dévoiler ses pensées, le droit de ne pas être sanctionné pour ses pensées et le droit de ne pas voir ses pensées manipulées. Les applications d'IA qui effectuent un microciblage des électeurs ou génèrent des infox vidéo (*deepfakes*) portent en elles un risque de manipulation. En outre, une application utilisant la reconnaissance des émotions fait peser le risque que des personnes soient pénalisées en raison de leurs pensées ou de leurs émotions (perçues).

Le droit à la liberté de réunion et d'association peut notamment être décomposé comme suit : droit d'organiser, de planifier et d'avoir un échange (en ligne) et droit d'organiser des réunions (spontanées). Une IA qui serait utilisée pour bloquer (temporairement) des sites web ou des forums afin d'empêcher cet échange pourrait mettre ce droit en péril, de même que des chatbots ou des trolls à base d'IA utilisés pour fabriquer de toutes pièces un faux assentiment ou un faux débat, qu'il soit positif (dans le domaine de la politique énergétique nationale par

⁹¹ Suggéré par l'IBA.

⁹² Exemple suggéré par le Royaume-Uni.

exemple) ou négatif (autour d'organisations qui critiqueraient cette politique)⁹³.

La troisième étape consiste à analyser l'impact que pourrait avoir l'application d'IA sur le droit examiné. Une analyse d'impact de l'IA sur les droits de l'homme peut faciliter ce travail.

La quatrième étape consiste à définir précisément les indicateurs permettant d'évaluer cet impact, sachant toutefois qu'ils doivent être adaptés à l'impact de l'application d'IA à l'étude. À noter que les étapes 2 et 3 peuvent être menées de front en identifiant l'application d'IA, les droits de l'homme et les impacts possibles jusqu'à ce que toutes les éventualités aient été évaluées. Après quoi il est possible de subdiviser le droit concerné en composantes et de dresser la carte des indicateurs.

Cette méthode est cohérente avec la précédente analyse des liens étroits qui existent entre droits de l'homme, démocratie et État de droit : le large éventail des droits de l'homme qui seraient à examiner dans le cadre de l'AIDHDE permettrait également de couvrir de nombreux éléments substantiels importants de la démocratie et de l'État de droit. Pour faciliter l'AIDHDE, on pourrait envisager de créer une liste de questions qui aborderaient l'ensemble des « droits indirects » dont on a indiqué plus haut qu'ils constituent le fondement de la démocratie et de l'État de droit.

⁹³ Exemple suggéré par le Royaume-Uni.

Section II. Vers un modèle de réalisation d'une analyse d'impact des systèmes d'IA sur les droits de l'homme, la démocratie et l'État de droit, sur la base de l'expérience acquise avec les systèmes d'analyse d'impact de l'IA sur les droits de l'homme

Nous nous sommes interrogés plus haut sur les modèles qui existent en matière d'analyse d'impact d'ordre général ou propre à l'IA ainsi que sur les questions à résoudre dans le cas de l'AIDHDE. Pour analyser l'impact sur les droits, tous ces modèles suivent une démarche fondée sur les risques. Il est donc logique de reprendre cette approche dans le cas de l'analyse d'impact de l'IA sur les droits de l'homme, la démocratie et l'État de droit. Après quelques remarques, la présente section décrit les difficultés à prendre en considération et définit les étapes à suivre dans le contexte d'une AIDHDE.

i) Remarques préliminaires sur l'opportunité de réaliser une AIDHDE

Il n'est pas certain qu'une AIDHDE doive systématiquement être menée lorsque l'on développe ou que l'on déploie une application d'IA. Au vu du temps et des ressources nécessaires pour effectuer ce type d'analyse, nous estimons qu'une AIDHDE complète ne s'impose que si une première analyse fait apparaître des risques accrus pour les droits de l'homme.

Les enseignements tirés des analyses d'impact relatives à la protection des données (AIPD), et notamment les Lignes directrices en matière d'AIPD⁹⁴, montrent qu'il serait souhaitable, également dans notre contexte, d'élaborer un premier ensemble de critères permettant de déterminer s'il est nécessaire ou non d'entreprendre une AIDHDE. Ces critères pourraient être liés à la portée de l'application, au fait qu'elle fasse systématiquement intervenir des êtres humains ou interagisse avec eux et/ou qu'elle ait sur eux des effets, et au type et à la finalité de l'application ou de cas d'utilisation particuliers susceptibles de porter atteinte aux droits de l'homme (reconnaissance faciale, techniques d'infox vidéo, réseaux sociaux, etc.).

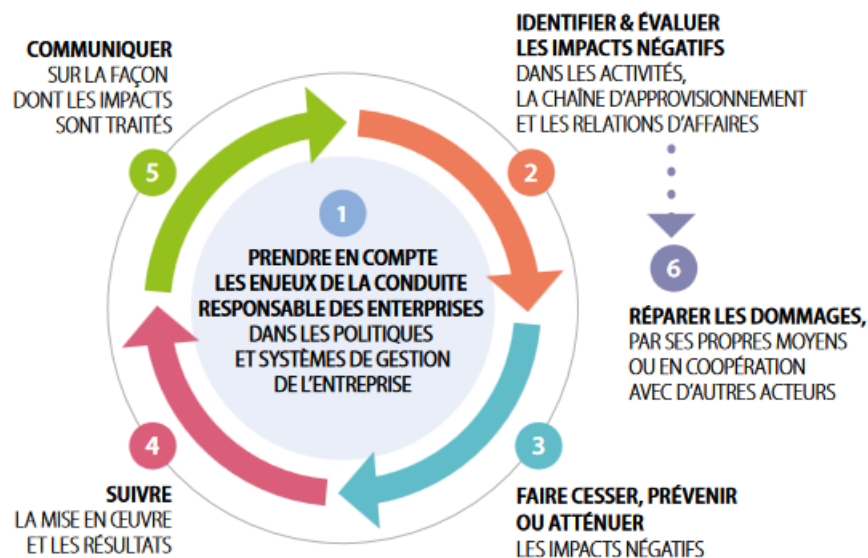
Si l'analyse initiale conclut à la nécessité de réaliser une AIDHDE complète et détaillée, un certain nombre d'éléments doivent être pris en considération. Tout d'abord, l'expérience pratique des AIPD nous montre combien il peut être difficile pour une entreprise de réaliser ce type d'analyse et que, de plus, il existe de nombreuses méthodes pour y parvenir, d'où des résultats difficiles à comparer et donc à exploiter et à interpréter. Ces difficultés peuvent empêcher l'établissement de conditions équitables entre les différents acteurs dans les États membres. Il est donc important de fournir des orientations pour les aider à déterminer à quel

⁹⁴ Les Lignes directrices concernant l'analyse d'impact relative à la protection des données (AIPD) fournissent des orientations générales pour déterminer si le traitement est « susceptible d'engendrer un risque élevé » aux fins du règlement (UE) 2016/679, p. 8. Elles sont disponibles à l'adresse https://ec.europa.eu/newsroom/just/document.cfm?doc_id=48464. En particulier, une AIPD devrait être réalisée lorsque l'une des conditions suivantes est remplie : « a) l'évaluation systématique et approfondie d'aspects personnels concernant des personnes physiques, qui est fondée sur un traitement automatisé, y compris le profilage, et sur la base de laquelle sont prises des décisions produisant des effets juridiques à l'égard d'une personne physique ou l'affectant de manière significative de façon similaire ; b) le traitement à grande échelle de catégories particulières de données visées à l'article 9, paragraphe 1, ou de données à caractère personnel relatives à des condamnations pénales et à des infractions visées à l'article 10 ; ou c) la surveillance systématique à grande échelle d'une zone accessible au public. »

moment et comment il convient de réaliser une AIDHDE complète, et aussi pour œuvrer dans le sens d'une plus grande comparabilité des résultats. De telles lignes directrices ont déjà été élaborées pour la réalisation d'AIPD⁹⁵. Il est donc judicieux de faire de même dans le cas des AIDHDE.

ii) Difficultés générales concernant la réalisation d'une AIDHDE propre à l'IA

L'état des lieux des analyses d'impact propres à l'IA nous apprend que la plupart d'entre elles, contrairement aux AIDH d'ordre général, ne mettent pas l'accent sur les droits de l'homme en tant que tels, mais adoptent un point de vue plus large comprenant souvent d'autres considérations, en particulier d'ordre éthique. De plus, en général, les analyses d'impact propres à l'IA n'intègrent pas plusieurs caractéristiques des AIDH générales, notamment les analyses d'impact tout au long des chaînes de valeur, l'engagement des parties prenantes et la mise en œuvre de mécanismes de réclamation dans le cadre de la diligence raisonnable en matière de droits de l'homme⁹⁶. Les modèles d'AIDH générale comprennent le plus souvent les six étapes de la diligence raisonnable en matière de droits de l'homme, qui, à l'exception de l'étape 1 et parfois 5, ne sont habituellement pas intégrées dans les analyses d'impact propres à l'IA⁹⁷. Cependant, il est important que l'AIDHDE mette en œuvre et documente ces six étapes.



⁹⁵ Lignes directrices concernant l'analyse d'impact relative à la protection des données (AIPD) et la manière de déterminer si le traitement est « susceptible d'engendrer un risque élevé » aux fins du règlement (UE) 2016/679, disponibles à l'adresse <https://ec.europa.eu/newsroom/article29/items/611236>.

⁹⁶ Cela dit, la liste d'auto-évaluation établie par le Groupe d'experts indépendants de haut niveau de l'UE comprend la consultation des parties prenantes et la participation (p. 18) ainsi que la possibilité pour les tiers (parmi lesquels les acteurs de la chaîne de valeur) de signaler des vulnérabilités (p. 22), mais pas l'obligation d'évaluer le risque tout au long de la chaîne de valeur et du cycle de vie de l'IA.

⁹⁷ À comparer avec la proposition de méthodologie de la Turquie.

Cela dit, l’empreinte numérique et la complexité des systèmes d’IA rendent l’AIDHDE dans son ensemble fondamentalement différente de l’AIDH à caractère général. Ainsi, une application d’IA peut avoir un impact mondial, alors que l’AIDH traditionnelle se concentre le plus souvent sur des projets ou des lieux de production particuliers au sein des chaînes d’approvisionnement. Il est donc plus difficile d’entamer un dialogue avec les parties prenantes dans le cas des applications d’IA déployées à l’échelle mondiale, notamment en raison des différences linguistiques et des écarts en matière de réglementation et de supervision publiques, mais aussi des différents niveaux de connaissances (techniques) sur le fonctionnement de l’IA.

Souvent, les outils d’analyse propres à l’IA ne précisent pas s’ils portent sur une analyse ponctuelle ou s’ils s’inscrivent dans un processus continu d’apprentissage et d’amélioration. En outre, comme on l’a vu précédemment, l’approche plus large et non centrée sur les droits de l’homme mise en œuvre dans les AIDH propres à l’IA pourrait être un défi supplémentaire à relever. De plus, il se peut qu’une AIDH propre à l’IA ne soit pas réalisée par le service qui est chargé d’effectuer des AIDH générales, ce qui peut creuser encore davantage cet écart de connaissances. Par ailleurs, les analyses d’impact plus étendues peuvent passer à côté de risques pour les droits fondamentaux qui sont spécifiques à l’IA. Il peut donc être utile de travailler en collaboration, voire d’intégrer les deux types d’analyse dans un même service. En outre, si deux types d’AIDH non harmonisés, à savoir une analyse générale et une analyse propre à l’IA, doivent être réalisés par un acteur privé, l’opération peut se révéler coûteuse et entraîner une charge administrative inutile. Enfin, une analyse propre à l’IA peut passer à côté de certains risques que relèverait une AIDH à caractère général, et inversement.

iii) Proposition de méthodologie pour l’AIDHDE

Le modèle utilisé pour l’AIDHDE devrait proposer une approche cohérente et intégrée pour l’évaluation des effets négatifs des systèmes d’IA sur les droits de l’homme, la démocratie et l’État de droit, en traitant simultanément les risques liés aux caractéristiques spécifiques et inhérentes des systèmes d’IA, y compris les systèmes autonomes, et l’impact de ces systèmes sur les droits de l’homme, la démocratie et l’État de droit.

Étape 1 : Identification des droits concernés

Dans un premier temps, le modèle d’AIDHDE doit permettre d’identifier les droits de l’homme sur lesquels l’application d’IA pourrait avoir des effets négatifs, y compris les droits qui sont des indicateurs indirects de la démocratie et de l’État de droit.

Étape 2 : Évaluation de l’impact sur ces droits

Au vu de la méthodologie présentée plus haut à la section B.2, l’évaluation de l’impact sur les droits identifiés à l’étape 1 doit s’appuyer sur les analyses d’impact générales et propres à l’IA

existantes et les intégrer⁹⁸. Cette évaluation doit comprendre des aspects techniques et non techniques.

Premièrement, l'évaluation technique doit permettre d'évaluer les problèmes que pose l'application d'IA en tant que telle sur le plan des droits de l'homme. Cette évaluation porterait sur la technologie sous-jacente utilisée et sur les caractéristiques techniques spécifiques que doit avoir l'application pour prévenir d'éventuels effets négatifs : explicabilité, transparence, cybersécurité, protection contre l'utilisation non prévue, etc. Cette partie de l'évaluation pourrait s'appuyer sur des analyses d'impact propres à l'IA existantes.

Deuxièmement, l'évaluation doit aussi comprendre une partie non technique consistant à analyser l'environnement sociotechnique général dans lequel le système évolue ainsi que les compétences et les savoir-faire requis pour le déploiement et l'utilisation d'une application d'IA donnée. De plus, il faut aussi envisager d'identifier, de traiter et de retracer les risques liés au déploiement de l'IA dans l'ensemble de la chaîne de valeur et du cycle de vie (étiquetage des données et autres services d'enrichissement, modération du contenu humain⁹⁹)¹⁰⁰.

La caractéristique distinctive des AIDHDE, qui les différencie en partie des AIDH générales ou propres à l'IA existantes, serait qu'elles comprennent une analyse spécifique de l'impact sur certains droits fondamentaux qui sont des indicateurs indirects de l'État de droit et de la démocratie, notamment la liberté d'expression, la liberté d'association et le droit à la non-discrimination, mais aussi le large éventail des droits de l'homme qui sous-tendent la démocratie et l'État de droit. Cette analyse spécifique s'appuierait sur les orientations générales fournies à la section I.B. Cette façon de procéder permettrait d'éviter la duplication des modèles existants ainsi que les coûts et charges administratives inutiles pour les acteurs privés. De plus, elle renforcerait l'application des cadres actuels en matière d'AIDH générale et propre à l'IA dans le champ de l'IA.

Étape 3 : Évaluation des mécanismes de gouvernance

En plus de l'analyse d'impact sur les droits réalisée à l'étape 2, il est aussi utile d'examiner si des mécanismes de gouvernance existants ou établis pourraient contribuer à atténuer les risques potentiels.

À cet égard, l'engagement des parties prenantes et la mise en place de mécanismes de réclamation ou la participation à de tels mécanismes sont des éléments à prendre en compte. Actuellement, les mécanismes de réclamation en lien avec l'IA sont rares, notamment en raison de l'opacité qui entoure l'utilisation de cette technologie dans certains contextes. Par exemple,

⁹⁸ Elle peut aussi s'appuyer sur des cadres de référence fondateurs comme la Recommandation CM/Rec(2020)1 du Comité des Ministres aux États membres sur les impacts des systèmes algorithmiques sur les droits de l'homme : https://search.coe.int/cm/pages/result_details.aspx?objectid=09000016809e1154.

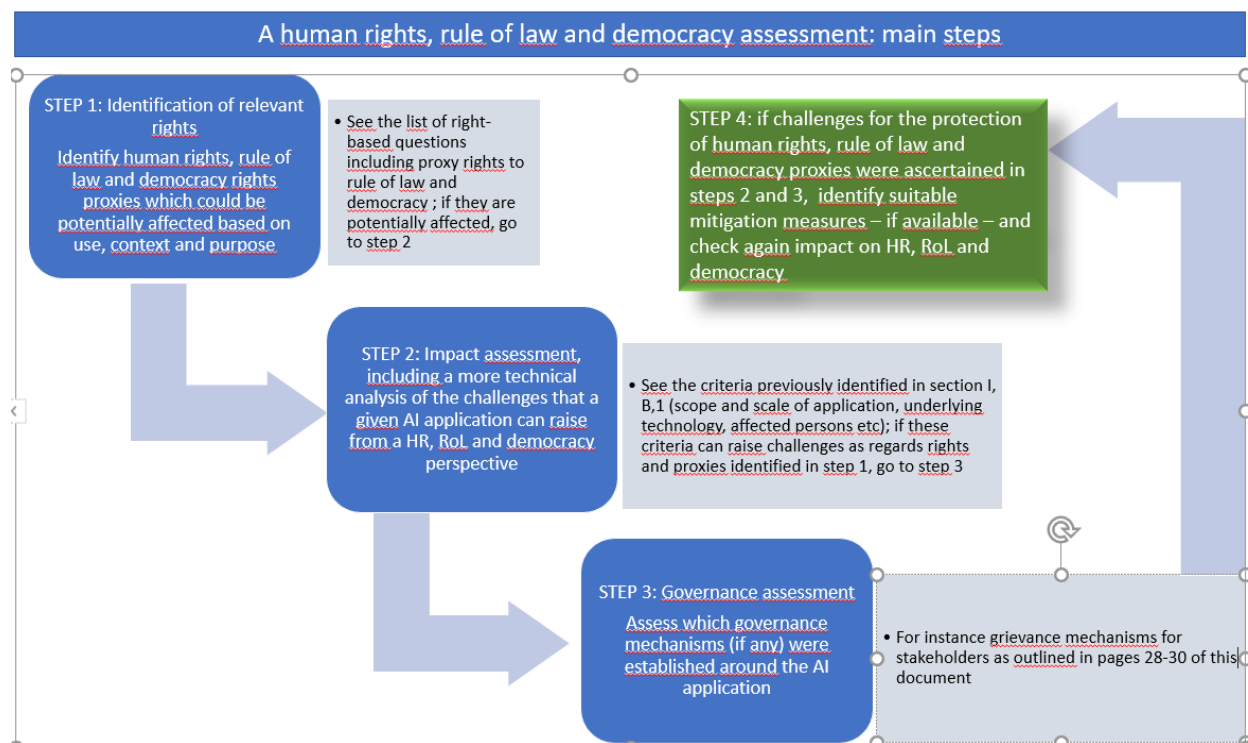
⁹⁹ Suggéré par CINGO.

¹⁰⁰ Cf. Projet B-Tech du HCDH, *Identifying and assessing human rights risks related to end-use*, disponible à l'adresse <https://www.ohchr.org/Documents/Issues/Business/B-Tech/identifying-human-rights-risks.pdf>.

tel organisme public ou privé qui utilise des caméras de surveillance pourra être identifiable, alors que le développeur de la technologie d'IA, lui, sera plus difficile à identifier. À cet égard, des enseignements peuvent être tirés des AIDH à caractère général pour lesquelles cette question a été traitée. Les issues des différentes réclamations sont aussi un élément indispensable du processus d'apprentissage continu de l'AIDHDE.

Étape 4 : Évaluation continue

Il convient enfin d'évaluer si les mécanismes de gouvernance identifiés – ou d'autres mesures d'atténuation – peuvent fournir une solution pour limiter l'impact sur les droits de l'homme qui a été mis en évidence, y compris sur les droits de l'homme qui sont des indicateurs indirects de la démocratie et de l'État de droit. Le processus d'analyse d'impact doit être évalué en permanence, au regard des changements intervenus dans le système d'IA, dans l'environnement du système et dans les mécanismes de gouvernance qui entourent le système. Car toute AIDHDE devrait être vue comme un processus itératif et non comme une activité ponctuelle.



Traduction du graphique

A human rights...	Analyse d'impact sur les droits de l'homme, l'État de droit et la démocratie : grandes étapes
STEP 1 ...	ÉTAPE 1 : Identification des droits concernés Identifier les droits de l'homme et les indicateurs indirects de l'État de droit et de la démocratie qui pourraient être touchés compte tenu de l'utilisation, du contexte et de la finalité du système

<i>See the list...</i>	<i>Voir la liste des questions basées sur les droits, y compris ceux qui sont des indicateurs indirects de la démocratie et de l'État de droit ; s'il y a un risque que ces droits soient impactés, aller à l'étape 2.</i>
<i>STEP 2 :...</i>	<i>ÉTAPE 2 : Évaluation de l'impact, notamment l'analyse technique des problèmes soulevés par une application d'IA donnée du point de vue des droits de l'homme, de la démocratie et de l'État de droit</i>
<i>See the criteria...</i>	<i>Voir les paramètres définis à la section I.B.1 (périmètre et portée de l'application, technologie sous-jacente, personnes impactées, etc.) ; si l'examen de ces paramètres révèle des problèmes potentiels au regard des droits et des indicateurs indirects identifiés à l'étape 1, aller à l'étape 3.</i>
<i>STEP 3 :</i>	<i>ÉTAPE 3 : Évaluation de la gouvernance Examiner si des mécanismes de gouvernance ont été mis en place autour de l'application d'IA.</i>
<i>For instance...</i>	<i>Par exemple, des mécanismes de réclamation destinés aux parties prenantes, comme indiqué aux pages 28 à 30.</i>
<i>STEP 4 :...</i>	<i>ÉTAPE 4 : Si des problèmes pour les droits de l'homme, y compris les indicateurs indirects de la démocratie et de l'État de droit, ont été mis en évidence aux étapes 2 et 3, identifier des mesures d'atténuation appropriées – s'il en existe – et procéder à une nouvelle analyse d'impact sur les droits de l'homme, la démocratie et l'État de droit.</i>

iv) Considérations complémentaires

Il importe de noter que pour réaliser une AIDH générale ou propre à l'IA d'une manière intégrée et cohérente dans le champ de l'IA, il faut veiller à l'inclusivité et à la mise en place de processus participatifs pleinement pertinents. De plus, contrairement à ce que l'on observe dans les environnements traditionnels, les conflits entre droits fondamentaux peuvent être plus fréquents dans le champ de l'IA (par exemple, respect de la vie privée contre liberté d'expression). L'AIDHDE devrait donc inclure des recommandations sur la manière de traiter ces conflits et de trouver le juste équilibre entre les droits fondamentaux concernés.

Il conviendrait également d'examiner si des caractéristiques ou des garanties supplémentaires doivent être ajoutées à l'AIDHDE lorsque l'IA est utilisée dans le secteur public, compte tenu des effets négatifs que peut produire l'utilisation de certaines applications d'IA sur les droits de l'homme, la démocratie et l'État de droit dans ce secteur.

En outre, il peut être nécessaire de renforcer les capacités d'analyse des problèmes de droits fondamentaux, notamment pour la partie la plus technique de l'AIDHDE. À cette fin, des conseils pratiques sur la façon d'appliquer la méthodologie décrite précédemment, notamment une liste de contrôle permettant aux exploitants d'en traiter tous les aspects, sont fournis à l'Annexe I et pourraient être affinés dans les mois à venir.

Apprentissage continu en vue de limiter l'impact. L'accès à une voie de recours devrait également structurer l'AIDHDE de manière à renforcer la responsabilité envers les parties prenantes concernées. Il importe de noter que l'AIDHDE fournit une analyse à un moment donné, alors que la diligence raisonnable en matière de droits de l'homme est un processus permanent d'apprentissage continu. Par exemple, l'AIDHDE peut révéler que des modifications à une application d'IA donnée seraient vivement recommandées. Une fois ces modifications apportées, une nouvelle analyse peut être réalisée afin de vérifier que l'impact a effectivement diminué.

Il peut aussi être judicieux de refaire une AIDHDE lorsque de nouveaux risques pour les droits fondamentaux ou une augmentation de ces risques sont mis en évidence pendant le processus de diligence raisonnable en matière de droits de l'homme. Cela peut se produire lorsque l'application d'IA est utilisée à d'autres fins que celles prévues à l'origine ou qu'elle est intégrée dans un réseau de systèmes plus vaste. L'AIDHDE n'est donc pas un exercice ponctuel, et il peut être nécessaire de la répéter. À cet égard, il est important d'examiner les résultats et les conclusions des AIDHDE successives en les considérant comme un tout cohérent et non comme des exercices isolés. De toute évidence, il convient de ne pas s'arrêter à l'exécution de la procédure d'AIDHDE, mais de mettre en œuvre ses résultats, par exemple en remaniant l'application d'IA ou le jeu de données qu'elle utilise. L'AIDHDE doit conduire à des plans d'action décrivant comment traiter et atténuer les risques identifiés en matière de droits de l'homme.

Participation des parties prenantes. Lorsque l'on entreprend une AIDHDE, le Guide sur l'analyse d'impact des activités numériques sur les droits de l'homme est l'un des premiers cadres de référence à examiner¹⁰¹. Bien qu'il soit conçu pour les technologies numériques au sens large, ce guide est également pertinent pour les applications d'IA. Il souligne que les AIDHDE doivent être menées non seulement par les développeurs d'IA, mais aussi par ceux qui vendent, achètent et déploient cette technologie¹⁰². Il adapte les étapes de la diligence raisonnable en matière de droits de l'homme mentionnées au paragraphe I.A aux technologies numériques¹⁰³. De plus, il souligne à juste titre que l'implication des parties prenantes, y compris au niveau local où les effets se produisent, est essentielle pour l'AIDHDE¹⁰⁴. Ainsi, l'AIDHDE devrait non seulement porter sur les aspects techniques de l'IA ou sur l'organisation qui développe, vend, achète ou déploie l'IA, mais aussi mettre les parties prenantes internes et externes intégrées dans la boucle. Du reste, comme nous l'avons expliqué plus haut, l'intervention des parties prenantes est une nécessité lorsque des risques accrus en matière de droits de l'homme ont été identifiés. L'ampleur de cet engagement auprès des parties prenantes doit être proportionnée à la gravité, à l'échelle et au caractère irréversible de l'impact de l'IA sur les droits de l'homme. Plus l'impact est grave et irréversible, ou plus son ampleur est

¹⁰¹ Disponible à l'adresse

https://www.humanrights.dk/sites/humanrights.dk/files/media/document/A%20HRIA%20of%20Digital%20Activities%20-%20Introduction_ENG_accessible.pdf.

¹⁰² p. 7.

¹⁰³ p. 13.

¹⁰⁴ p. 15.

grande, plus les parties prenantes doivent être associées au processus, en particulier les parties prenantes externes.

Cependant, il n'est pas toujours facile d'identifier les parties externes concernées et de les associer au processus. Dans le cas de l'AIDH classique, des recommandations ont été formulées à ce sujet¹⁰⁵. Pour ce qui est de l'IA en revanche, la portée de l'application d'IA peut rendre cette identification et cette participation plus difficiles. En effet, une AIDH pour l'IA peut avoir, par exemple, une portée mondiale, alors que l'AIDH traditionnelle porte généralement sur un périmètre bien précis. Lorsque les risques liés aux droits de l'homme se limitent à certains pays ou à certains lieux, l'identification et la participation des parties prenantes ressemblent à celles d'une AIDH classique, et des approches traditionnelles peuvent être mises en œuvre.

Les parties prenantes doivent être associées à l'ensemble du processus d'AIDHDE, de la planification et la délimitation de l'analyse à la conception et à la consultation des parties prenantes, jusqu'à l'évaluation finale de l'application¹⁰⁶. Il importe également de permettre un dialogue continu et de veiller à ce que la technologie et les termes utilisés ainsi que l'impact potentiel de l'application d'IA sur les droits de l'homme soient compréhensibles par les différentes parties. Comme nous l'avons mentionné plus haut, la transparence de l'AIDHDE est un élément important, qui suppose de fournir des informations complètes et accessibles sur l'application d'IA pour permettre aux parties prenantes de comprendre son impact potentiel sur leurs droits fondamentaux, et de prévoir du temps pour discuter des résultats de l'analyse¹⁰⁷. Il est possible, dans le cadre de ces AIDHDE, de nouer et renforcer des partenariats entre les entreprises et les parties prenantes afin d'identifier les risques en matière de droits de l'homme¹⁰⁸.

Accès à une voie de recours. L'accès à une voie de recours est un aspect important à examiner lorsque des effets négatifs sont détectés¹⁰⁹. À cet égard, il est essentiel d'évaluer si un mécanisme de réclamation de niveau opérationnel est en place, s'il est conforme aux exigences du Principe directeur 31 des Nations Unies et si l'entreprise coopère lors de la procédure judiciaire. Pour ce qui est de la mise en place et du fonctionnement de ces mécanismes, l'entreprise qui s'engage auprès des parties prenantes au sujet d'une application d'IA devrait également participer à un mécanisme de réclamation pour les parties en question ou bien mettre sur pied un tel mécanisme, au cas où des problèmes surviendraient après l'engagement ou que des manquements seraient constatés dans ledit engagement.

Outre l'AIDH réalisée dans le contexte des technologies numériques, une analyse d'impact relative à la protection des données (AIPD) peut aussi fournir des orientations utiles. Une AIPD est exigée par l'article 25 du RGPD dans le cas des applications (d'IA) susceptibles d'engendrer

¹⁰⁵ Voir par exemple le document *Guidance Document for Social Accountability 8000 (SA8000:2014)*, p. 124, disponible à l'adresse <https://sa-intl.org/wp-content/uploads/2020/02/SA8000-2014-Guidance-Document.pdf>.

¹⁰⁶ p. 28. Cela signifie, entre autres, que les consultations ont lieu lorsque les parties prenantes sont censées être disponibles, par exemple lorsque les femmes ne sont pas travail ou que les jeunes enfants ne sont pas à l'école.

¹⁰⁷ p. 31.

¹⁰⁸ Voir p. 20, qui énumère aussi d'autres avantages.

¹⁰⁹ p. 36.

un risque élevé. Un aspect important des AIPD réside dans l'utilisation de questionnaires concernant des risques particuliers en lien avec la protection des données. Le format utilisé consiste en des questions types portant sur des sujets pertinents en lien avec la protection des données, par exemple la loyauté, la transparence, la limitation de la finalité, la minimisation des données, l'exactitude, la limitation de la conservation et la sécurité¹¹⁰. Comme mentionné plus haut, il peut être judicieux de suivre une approche comparable dans le cas de l'AIDHDE en s'inspirant des enseignements tirés de la Liste d'évaluation pour une IA digne de confiance élaborée par le Groupe d'experts indépendants de haut niveau sur l'intelligence artificielle de l'UE dans le cadre plus large de ses Lignes directrices en matière d'éthique pour une IA digne de confiance¹¹¹, lesquelles insistent en particulier sur les risques que présentent les systèmes d'IA en général.

En conclusion, l'AIDHDE doit combiner les problématiques relevant des AIDH traditionnelles et les questions plus techniques, afin de traiter conjointement les aspects liés aux droits de l'homme et la dimension technique. Une troisième dimension, la gouvernance, doit également être examinée. Ces divers éléments et dimensions sont décrits de manière synthétique dans le tableau de l'Annexe I du présent document. Il est à noter que l'AIDHDE doit être plus détaillée si les réponses aux questions types indiquent que les risques liés à l'IA sont graves, étendus et/ou irrémédiables. En outre, l'AIDHDE sera plus ou moins poussée selon que l'entreprise concernée est à l'origine d'un impact sur les droits de l'homme, qu'elle y contribue ou qu'elle y est liée. Enfin, l'AIDHDE est un processus continu. En effet, les réponses aux questions types peuvent ne pas indiquer de risque accru pour les droits de l'homme à un moment donné, mais cette situation peut évoluer si le système d'IA est utilisé pour un usage autre que celui auquel il est destiné, qu'un nouveau système d'IA est conçu ou que d'autres applications sont développées. Dans tous ces cas de figure, il faut à nouveau répondre aux questions types. Les réponses peuvent éventuellement indiquer une augmentation du risque (voir Tableau 1).

Si l'AIDHDE est menée comme on vient de le décrire, elle peut révéler des effets négatifs, mais aussi des effets positifs sur les droits de l'homme, l'État de droit et la démocratie. L'AIDHDE n'est pas conçue pour équilibrer les effets positifs et négatifs, car cette tâche incombe aux autorités nationales et peut également dépendre des caractéristiques des systèmes juridiques. Comme indiqué plus haut, l'AIDHDE n'est pas une approche visant à équilibrer les impacts sur les droits de l'homme. Elle ne fait qu'identifier les risques en la matière.

¹¹⁰ Contrôleur européen de la protection des données, *Accountability on the ground Part II: Data Protection Impact Assessments & Prior Consultation*, pp. 11-15, https://edps.europa.eu/sites/edp/files/publication/18-02-06_accountability_on_the_ground_part_2_en.pdf.

¹¹¹ La liste d'auto-évaluation peut être consultée à l'adresse <https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificial-intelligence-atai-self-assessment>. Les pages 5 et 6 en particulier sont pertinentes par rapport à la question des droits de l'homme. Les Lignes directrices sont disponibles à l'adresse <https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html>.

Section III. Synergies entre AIDHDE et mécanismes de conformité

Les analyses d'impact sur les droits de l'homme, la démocratie et l'État de droit proposées ici ne sont pas totalement distinctes des autres mécanismes de conformité tels que la certification et l'attribution d'un label de qualité, les audits, les sas réglementaires et la surveillance continue et automatisée, qui sont mentionnés dans l'Étude de faisabilité du CAHAI.

D'une part, ces mécanismes de conformité peuvent s'appuyer sur l'AIDHDE dans la mesure où ils exigeraient la réalisation d'une telle analyse d'impact pour le développement ou le déploiement de l'IA. Et même le niveau de surveillance publique peut varier selon que l'AIDHDE est plus ou moins étendue. Il est important que l'AIDHDE et les autres mécanismes de conformité soient en phase. En effet, il serait difficile de justifier les coûts et la charge de travail engendrés par une AIDHDE qui serait en contradiction avec les stratégies de surveillance publique ou de réglementation **(A)**.

D'autre part, une approche fondée sur les risques suppose une réévaluation systématique et périodique des risques posés par les systèmes d'IA. Toute mesure d'atténuation doit être adaptée à ces risques, en particulier ceux qui touchent les groupes vulnérables et marginalisés (par exemple, les personnes autochtones, noires et de couleur). Outre l'approche fondée sur les risques, il convient aussi de prendre en compte le principe de précaution (s'il y a lieu), et notamment d'envisager des interdictions¹¹² **(B)**.

A. Harmonisation de l'AIDHDE et des mécanismes de conformité

En théorie, l'AIDHDE pourrait¹¹³ servir de base aux mécanismes de conformité et ainsi contribuer à établir les mêmes règles entre les États membres. De ce point de vue, l'AIDHDE pourrait, conjointement avec d'autres mécanismes de conformité, créer un écosystème d'outils qui constituerait un cadre commun pour accroître la confiance et renforcer la transparence autour de l'utilisation de l'IA¹¹⁴.

L'écosystème de conformité est le résultat de l'harmonisation des politiques des États, laquelle a pour but de garantir la conformité des réglementations nationales avec les cadres juridiques futurs et avec les rôles des différents acteurs. Cette harmonisation comprend, dans le cas particulier de l'AIDHDE, deux dimensions. La première est le socle commun : le modèle utilisé pour réaliser l'AIDHDE doit être commun aux assureurs, aux développeurs, aux exploitants et aux utilisateurs d'applications d'IA. La deuxième dimension est le renforcement mutuel : l'utilisation commune de l'AIDHDE devrait inciter les différents acteurs à respecter les normes et à contrôler leur bonne application, et, collectivement, ces acteurs devraient, par effet de synergie, contribuer à produire une nouvelle culture des applications d'IA respectueuse des

¹¹² Suggéré par CINGO.

¹¹³ Les propositions faites par EE ont largement contribué à la mise au point du texte.

¹¹⁴ Cf. paragraphe 155 de l'Étude de faisabilité.

droits de l'homme, de la démocratie et de l'État de droit.

L'exploitation, par des tiers, des enseignements tirés de l'analyse d'impact est un bon exemple de renforcement mutuel entre l'AIDHDE et les autres mécanismes de conformité. Ainsi, il est souhaitable que les États membres mettent en commun les informations relatives à l'AIDHDE réalisée dans leur juridiction, afin d'encourager l'émergence de bonnes pratiques, et aussi à des fins de suivi et de contrôle des applications d'IA qui ont une portée juridictionnelle multiple. L'AIDHDE est un processus dynamique, ce qui signifie que des analyses d'impact doivent être réalisées tout au long du cycle de vie de l'application d'IA¹¹⁵. À chaque analyse, les informations produites doivent être mises à la disposition des organisations et des acteurs privés qui élaborent des normes et des certifications fondées sur des données factuelles.

Acteurs, assureurs, développeurs, exploitants et utilisateurs, tous jouent un rôle essentiel dans l'émergence d'une nouvelle culture de conformité des systèmes d'IA avec le nouveau cadre juridique. Lorsqu'une AIDHDE révèle un risque important d'impact sur les droits de l'homme, il convient d'identifier et de faire intervenir les parties prenantes concernées afin d'évaluer dans quelle mesure leurs droits fondamentaux, mais aussi la démocratie et l'État de droit sont impactés. Cela dit, l'AIDHDE doit accorder une attention particulière aux exploitants et aux utilisateurs, et ce pour deux raisons indépendantes. La première est que les exploitants et les utilisateurs d'applications d'IA sont naturellement intéressés par toute information concernant les dommages ou risques potentiels liés à une application d'IA donnée, quelle qu'en soit la source (particulier, organisation ou organisme public)¹¹⁶. La seconde raison est qu'il conviendrait d'instaurer – comme condition préalable à la nouvelle culture d'une IA conforme aux droits de l'homme – l'obligation d'expliquer, de manière opportune, impartiale, accessible et lisible, aux personnes dont les droits ou les intérêts légitimes peuvent être impactés ainsi qu'aux autorités publiques compétentes, l'existence, le processus, la raison d'être, le raisonnement et les produits éventuels des systèmes algorithmiques au niveau individuel et collectif.

Les considérations de confidentialité ou les secrets commerciaux ne devraient pas entraver la réalisation d'analyses d'impact effectives sur les droits de l'homme et la mise en œuvre des mesures correctives¹¹⁷.

Enfin, les mécanismes de conformité peuvent donner lieu à l'interdiction ou à la stricte réglementation du développement et/ou du déploiement des applications d'IA que l'AIDHDE a signalées comme étant problématiques.

B. Harmonisation des mécanismes correctifs et de l'AIDHDE

Dans le contexte de l'analyse des risques posés par les applications d'IA du point de vue des

¹¹⁵ Cf. paragraphe 168 de l'Étude de faisabilité.

¹¹⁶ C'est ce que nous apprennent les analyses d'impact relatives à la protection des données. L'expérience acquise dans le domaine de la protection des données pourrait peut-être fournir des indications concrètes sur la manière de réaliser une AIDHDE. Avis du CEPD.

¹¹⁷ Proposé par HomoDigitalis en référence à la Recommandation CM/Rec(2020)1 au sujet de la contestabilité (p. 9).

droits de l'homme, de l'État de droit et de la démocratie, la question des mécanismes correctifs présente trois grands aspects. Le premier concerne l'utilisation de l'AIDHDE comme un outil de renforcement de l'autonomie par l'information destiné aux utilisateurs d'applications d'IA et autres acteurs clés de l'écosystème des mécanismes de conformité (i). Le deuxième aspect porte sur l'AIDHDE en tant qu'élément d'un processus, plus large, de diligence raisonnable en matière de droits de l'homme (ii). Le troisième aspect de la question concerne le cadre commun de l'AIDHDE pour la conception des dispositifs de réparation (iii).

i) L'AIDHDE en tant qu'outil de renforcement de l'autonomie par l'information

Contester la conception, le développement, le déploiement, le fonctionnement ou l'utilisation d'une application d'IA représente un coût (et c'est aussi un droit). Cette question, qui touche à la fois aux mécanismes de conformité (publics) et aux droits de l'homme, est fondamentale. À cet égard, il est important que les informations produites par l'AIDHDE soient compréhensibles par des spécialistes et des non-spécialistes et qu'elles puissent être utilisées pour étayer une réclamation éventuelle ainsi que la demande de réparation en cas de préjudice¹¹⁸.

ii) L'AIDHDE en tant qu'élément d'un processus plus large de diligence raisonnable en matière de droits de l'homme

La diligence raisonnable en matière de droits de l'homme comprend la réalisation d'une analyse d'impact. Lorsqu'un risque est détecté, il doit être traité en prenant des mesures adéquates pour prévenir le préjudice ou l'atténuer. De même, pour que les parties prenantes puissent intenter un recours en cas d'atteinte aux droits de l'homme dans le domaine de l'IA, il faut fournir les résultats d'une AIDHDE.

Même en dehors du champ de l'IA, l'accès à une voie de recours pose d'énormes problèmes, notamment en cas d'abus que des acteurs privés auraient causé ou auquel ils auraient contribué. Le respect de la conformité doit intégrer la question de l'accès à une voie de recours.

Cette approche est conforme aux Principes directeurs des Nations Unies relatifs aux entreprises et aux droits de l'homme ainsi qu'aux Principes directeurs de l'OCDE, lesquels prévoient l'accès à des voies de recours. Les acteurs privés sont censés prévoir une voie de recours en cas d'abus qu'ils auraient causé ou auquel ils auraient contribué. On dit qu'ils ont « causé » un abus s'ils sont à l'origine de l'impact constaté, et qu'ils y ont « contribué » s'ils ont aidé un autre acteur à le commettre ou qu'ils y ont participé. Les cadres de référence comprennent également un troisième type d'implication : un acteur n'ayant ni causé un impact négatif ni contribué à cet impact peut néanmoins y être « lié ». Dans ce cas, l'acteur en question n'est pas tenu de fournir un accès à une voie de recours, mais il peut jouer un rôle en permettant cet accès, par exemple en exerçant une influence sur les tiers qui ont causé l'impact ou y ont contribué. Ainsi, l'accès à une voie de recours doit être assuré par les développeurs et les utilisateurs de l'IA lorsqu'un effet indésirable se produit, en particulier lorsqu'un développeur

¹¹⁸ Cf. paragraphe 168 de l'Étude de faisabilité du CAHAI.

ou un utilisateur cause un impact ou y contribue.

iii) L'AIDHDE en tant que socle commun pour la conception des dispositifs de réparation

Pour assurer l'accès à une voie de recours, des dispositifs de réparation judiciaires et non judiciaires doivent être mis en place, par exemple au niveau opérationnel du développeur et de l'utilisateur (commercial) de l'IA¹¹⁹. En outre, les résultats de ces dispositifs devraient être utilisés pour alimenter les AIDHDE en cours. Compte tenu du caractère opaque de certaines applications d'IA, il est important d'encourager la créativité, l'innovation et la collaboration entre les différents dispositifs de réparation étatiques et non étatiques afin d'améliorer l'accès aux voies de recours¹²⁰. Les mécanismes de conformité devraient donc aussi encourager ce type d'accès à une voie de recours dans le domaine de l'IA.

¹¹⁹ Projet B-Tech du HCDH, *Access to remedy and the technology sector: a « remedy ecosystem » approach*, p. 2, disponible à l'adresse <https://www.ohchr.org/Documents/Issues/Business/B-Tech/access-to-remedy-ecosystem-approach.pdf>. Cf. Recommandation CM/Rec(2020)1 du Comité des Ministres aux États membres sur les impacts des systèmes algorithmiques sur les droits de l'homme, point 4.4, disponible à l'adresse https://search.coe.int/cm/pages/result_details.aspx?ObjectId=09000016809e1124.

¹²⁰ Projet B-Tech du HCDH, *Access to remedy and the technology sector: a « remedy ecosystem » approach*, p. 2, disponible à l'adresse <https://www.ohchr.org/Documents/Issues/Business/B-Tech/access-to-remedy-ecosystem-approach.pdf>.

TABLEAU 1

Aspect relatif aux droits de l'homme/aspect technique	Niveau élevé d'autonomie (« humain sur la boucle », etc.)	Utilisation en dehors de l'usage prévu à l'origine	Partie de multiples réseaux de neurones profonds/s'appuie sur d'autres systèmes/double usage	Explicabilité	Transparence/reproductibilité	Qualité des données	Robustesse /sécurité	Information claire concernant l'utilisation ou non d'un système d'IA	Signal d'alerte/cas d'utilisation particuliers (influx vidéo, reconnaissance faciale, systèmes d'armes, etc.)
Discrimination	Si les questions types mettent en évidence un risque élevé : participation des parties prenantes à toutes les étapes/accès à une voie de recours en phase de déploiement	Si le questionnaire met en évidence un risque élevé : participation des parties prenantes lors du déploiement en vue d'une nouvelle utilisation, et accès à une voie de recours	Si les questions types mettent en évidence un risque élevé : participation des parties prenantes à toutes les étapes et accès à une voie de recours	Aspect essentiel pour la participation des parties prenantes	Aspect essentiel pour l'accès à une voie de recours (tout particulièrement dans le cas des systèmes très autonomes ou des multiples réseaux de neurones profonds)	Si les questions types mettent en évidence un risque élevé : participation des parties prenantes à la constitution du jeu de données/à l'élaboration du dispositif d'accès à une voie de recours	Évaluation par le développeur et l'utilisateur	Cet aspect est moins important	Si cela est autorisé, participation forte des parties prenantes à toutes les étapes et accès à une voie de recours
Liberté d'expression	Si les questions types mettent en évidence un risque élevé : participation des	Si les questions types mettent en évidence	Si les questions types mettent en évidence un risque élevé : participation des parties	Aspect essentiel pour la participation des parties	Aspect essentiel pour l'accès à une voie de recours (tout	Si les questions types mettent en	Évaluation par le développeur et l'utilisateur	Accès à une voie de recours	Si cela est autorisé, participation forte des parties

	parties prenantes à toutes les étapes/accès à une voie de recours en phase de déploiement	un risque élevé : participation des parties prenantes lors du déploiement en vue d'une nouvelle utilisation, et accès à une voie de recours	prenantes à toutes les étapes et accès à une voie de recours	prenantes	particulièrement dans le cas des systèmes très autonomes ou des multiples réseaux de neurones profonds)	évidence un risque élevé : participation des parties prenantes à la constitution du jeu de données/à l'élaboration du dispositif d'accès à une voie de recours			prenantes à toutes les étapes et à l'élaboration du dispositif d'accès à une voie de recours
Liberté de réunion et d'association	Si les questions types mettent en évidence un risque élevé : participation des parties prenantes à toutes les étapes/accès à une voie de recours en phase de déploiement	Cet aspect est moins important	Cet aspect est moins important	Aspect essentiel pour la participation des parties prenantes	Aspect essentiel pour l'accès à une voie de recours (tout particulièrement dans le cas des systèmes très autonomes ou des multiples réseaux de neurones profonds)	Cet aspect est moins important	Évaluation par le développeur et l'utilisateur	Cet aspect est moins important	Si cela est autorisé, participation forte des parties prenantes à toutes les étapes et accès à une voie de recours
Respect de la vie privée	Cet aspect est moins important	Si les questions types mettent en évidence un risque élevé : participation des parties prenantes lors	Si les questions types mettent en évidence un risque élevé : participation des parties prenantes lors du déploiement en vue d'une nouvelle utilisation, et accès à une voie de recours	Aspect essentiel pour la participation des parties prenantes	Aspect essentiel pour l'accès à une voie de recours (tout particulièrement dans le cas des systèmes très autonomes ou des multiples	Cet aspect est moins important	Évaluation par le développeur et l'utilisateur	Cet aspect est moins important	Si cela est autorisé, participation forte des parties prenantes à toutes les étapes et accès à une voie de recours

		du déploiement en vue d'une nouvelle utilisation, et accès à une voie de recours			réseaux de neurones profonds)				
Liberté de pensée	Si les questions types mettent en évidence un risque élevé : participation des parties prenantes à toutes les étapes/accès à une voie de recours en phase de déploiement	Si les questions types mettent en évidence un risque élevé : participation des parties prenantes lors du déploiement en vue d'une nouvelle utilisation, et accès à une voie de recours	Si les questions types mettent en évidence un risque élevé : participation des parties prenantes à toutes les étapes et accès à une voie de recours	Aspect essentiel pour la participation des parties prenantes	Aspect essentiel pour l'accès à une voie de recours (tout particulièrement dans le cas des systèmes très autonomes ou des multiples réseaux de neurones profonds)	Cet aspect est moins important	Évaluation par le développeur et l'utilisateur	Si les questions types mettent en évidence un risque élevé : participation des parties prenantes à toutes les étapes et accès à une voie de recours	Si cela est autorisé, participation forte des parties prenantes à toutes les étapes et accès à une voie de recours
Vie privée et familiale (au-delà du respect de la vie privée)	Si les questions types mettent en évidence un risque élevé : participation des parties prenantes à toutes les étapes/accès à une voie de recours en phase de déploiement	Si les questions types mettent en évidence un risque élevé : participation des parties prenantes lors du déploiement	Si les questions types mettent en évidence un risque élevé : participation des parties prenantes à toutes les étapes et accès à une voie de recours	Aspect essentiel pour la participation des parties prenantes	Aspect essentiel pour l'accès à une voie de recours (tout particulièrement dans le cas des systèmes très autonomes ou des multiples réseaux de neurones	Si les questions types mettent en évidence un risque élevé : participation des parties prenantes à la	Évaluation par le développeur et l'utilisateur	Si les questions types mettent en évidence un risque élevé : participation des parties prenantes à toutes les	Si cela est autorisé, participation forte des parties prenantes à toutes les étapes et accès à une voie de recours

		en vue d'une nouvelle utilisation, et accès à une voie de recours			profonds)	constitution du jeu de données/à l'élaboration du dispositif d'accès à une voie de recours		étapes et accès à une voie de recours	
Droit à la vie	Si les questions types mettent en évidence un risque élevé : participation des parties prenantes à toutes les étapes/accès à une voie de recours en phase de déploiement	Si les questions types mettent en évidence un risque élevé : participation des parties prenantes lors du déploiement en vue d'une nouvelle utilisation, et accès à une voie de recours	Si les questions types mettent en évidence un risque élevé : participation des parties prenantes à toutes les étapes et accès à une voie de recours	Aspect essentiel pour la participation des parties prenantes	Aspect essentiel pour l'accès à une voie de recours (tout particulièrement dans le cas des systèmes très autonomes ou des multiples réseaux de neurones profonds)	Si les questions types mettent en évidence un risque élevé : participation des parties prenantes à la constitution du jeu de données/à l'élaboration du dispositif d'accès à une voie de recours	Évaluation par le développeur et l'utilisateur	Cet aspect est moins important	De préférence non autorisé, mais si cela est nécessaire : participation forte des parties prenantes à toutes les étapes et accès à une voie de recours
Droit à un procès équitable	Si les questions types mettent en évidence un risque élevé : participation des parties prenantes à toutes les étapes	Si les questions types mettent en évidence un risque élevé : participation des parties prenantes lors	Si les questions types mettent en évidence un risque élevé : participation des parties prenantes à toutes les étapes	Aspect essentiel pour la participation des parties prenantes	Aspect essentiel pour l'accès à une voie de recours (tout particulièrement dans le cas des systèmes très autonomes ou des multiples	Si les questions types mettent en évidence un risque élevé : participation	Évaluation par le développeur et l'utilisateur	Si les questions types mettent en évidence un risque élevé : participation des parties	Probablement non applicable

		du déploiement en vue d'une nouvelle utilisation			réseaux de neurones profonds)	des parties prenantes à la constitution du jeu de données		prenantes à toutes les étapes	
Droit à la propriété	Si les questions types mettent en évidence un risque élevé : participation des parties prenantes à toutes les étapes	Si les questions types mettent en évidence un risque élevé : participation des parties prenantes lors du déploiement en vue d'une nouvelle utilisation	Si les questions types mettent en évidence un risque élevé : participation des parties prenantes à toutes les étapes	Aspect essentiel pour la participation des parties prenantes	Aspect essentiel pour l'accès à une voie de recours (tout particulièrement dans le cas des systèmes très autonomes ou des multiples réseaux de neurones profonds)	Si les questions types mettent en évidence un risque élevé : participation des parties prenantes à la constitution du jeu de données/à l'élaboration du dispositif d'accès à une voie de recours	Évaluation par le développeur et l'utilisateur	Aspect moins pertinent	Si cela est autorisé, participation forte des parties prenantes à toutes les étapes et accès à une voie de recours