GROUPE D'EXPERTS

INDEPENDANTS DE HAUT NIVEAU SUR L'INTELLIGENCE ARTIFICIELLE

CONSTITUE PAR LA COMMISSION EUROPEENNE EN JUIN 2018

LIGNES DIRECTRICES EN MATIERE D'ETHIQUE POUR UNE IA DIGNE DE CONFIANCE

LIGNES DIRECTRICES EN MATIERE D'ETHIQUE pour UNE IA DIGNE DE CONFIANCE

Groupe d'experts de haut niveau sur l'intelligence artificielle

Le présent document a été rédigé par le groupe d'experts de haut niveau sur l'intelligence artificielle (GEHN IA). Les membres du GEHN IA qui y sont cités soutiennent le cadre général pour une IA digne de confiance présenté dans les présentes lignes directrices, sans approuver nécessairement chacune des affirmations formulées dans le document.

Afin de recueillir des commentaires pratiques, les parties prenantes soumettront à une phase pilote la liste d'évaluation pour une IA digne de confiance présentée au chapitre III du présent document. Une version révisée de cette liste d'évaluation tenant compte des commentaires recueillis au cours de la phase pilote sera présentée à la Commission européenne début 2020.

Le GEHN IA est un groupe d'experts indépendants constitué par la Commission européenne en juin 2018.

Personne de contact Nathalie Smuha – coordinatrice du groupe d'experts de haut niveau sur l'IA
Adresse électronique CNECT-HLG-AI@ec.europa.eu

Commission européenne
B-1049 Bruxelles

Document rendu public le X avril 2019.

Un premier projet de ce document a été publié le 18 décembre 2018 et a fait l'objet d'une consultation ouverte à laquelle plus de 500 contributeurs ont apporté des commentaires. Nous souhaitons remercier explicitement et chaleureusement toutes les personnes ayant fait part de leurs commentaires sur le premier projet de ce document. Ces commentaires ont été pris en compte dans le cadre de l'élaboration de cette version révisée.

De plus amples informations sur le groupe d'experts de haut niveau sur l'intelligence artificielle sont disponibles en ligne (https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence).

TABLE DES MATIERES

RÉSUMÉ **2**

A. INTRODUCTION **5**

B. CADRE POUR UNE IA DIGNE DE CONFIANCE **7**

I. Chapitre I: Fondements d'une IA digne de confiance **11**
1. Les droits fondamentaux en tant que droits moraux et légaux 12
2. Des droits fondamentaux aux principes éthiques 12

II. Chapitre II: Parvenir à une IA digne de confiance **17**
1. Exigences d'une IA digne de confiance 17
2. Méthodes techniques et non techniques pour parvenir à une IA digne de confiance 25

III. Chapitre III: évaluation d'une IA digne de confiance **30**

C. EXEMPLES DE POSSIBILITES ET DE PREOCCUPATIONS MAJEURES SOULEVEES PAR L'IA **42**

D. CONCLUSION **46**

GLOSSAIRE **48**

RÉSUMÉ

(1) Les présentes lignes directrices visent à promouvoir une IA digne de confiance. Une IA digne de confiance présente les **trois caractéristiques** suivantes, qui devraient être respectées tout au long du cycle de vie du système: a) elle doit être **licite**, en assurant le respect des législations et réglementations applicables; b) elle doit être **éthique**, en assurant l'adhésion à des principes et valeurs éthiques; et c) elle doit être **robuste**, sur le plan tant technique que social car, même avec de bonnes intentions, les systèmes d'IA peuvent causer des préjudices involontaires. Toutes ces caractéristiques sont nécessaires en elles-mêmes, mais elles ne sauraient suffire à la réalisation d'une IA digne de confiance. L'idéal serait que ces trois caractéristiques fonctionnent en harmonie et se chevauchent. Si, dans la pratique, des tensions venaient à apparaître entre ces caractéristiques, la société devrait s'efforcer d'y remédier.

(2) Les présentes lignes directrices établissent un **cadre pour parvenir à la réalisation d'une IA digne de confiance**. Ce cadre ne traite pas explicitement de la première caractéristique d'une IA digne de confiance (IA licite)[1]. Il vise plutôt à proposer des orientations pour encourager et garantir une IA éthique et robuste (les deuxième et troisième caractéristiques). S'adressant à l'ensemble des parties prenantes, les présentes lignes directrices cherchent, en plus de présenter une liste de principes éthiques, à fournir des orientations sur la manière dont ces principes peuvent être mis en œuvre dans des systèmes sociotechniques. Ces orientations se présentent sous la forme de trois niveaux d'abstraction, du plus abstrait, au chapitre I, au plus concret, au chapitre III, et se concluant par des exemples de possibilités et de préoccupations graves soulevées par les systèmes d'IA.

I. Sur la base d'une approche fondée sur les droits fondamentaux, le chapitre I recense les **principes éthiques** et les valeurs correspondantes qu'il convient de respecter lors de la mise au point, du déploiement et de l'utilisation de systèmes d'IA.

Orientations essentielles dérivées du chapitre I:

- ✓ Mettre au point, déployer et utiliser des systèmes d'IA en respectant les principes éthiques suivants: *respect de l'autonomie humaine, prévention de toute atteinte, équité et explicabilité.* Reconnaître et résoudre les tensions potentielles entre ces principes.
- ✓ Accorder une attention particulière aux situations concernant des groupes plus vulnérables tels que les enfants, les personnes handicapées et d'autres groupes historiquement défavorisés ou exposés au risque d'exclusion, et aux situations caractérisées par des asymétries de pouvoir ou d'information, par exemple entre les employeurs et les travailleurs, ou entre les entreprises et les consommateurs.[2]
- ✓ Reconnaître et être conscient que les systèmes d'IA apportent certes des avantages considérables aux individus et à la société, mais qu'ils présentent également certains risques et peuvent avoir des incidences négatives, y compris des incidences pouvant s'avérer difficiles à anticiper, à déterminer ou à mesurer (par exemple des incidences sur la démocratie, l'état de droit et la justice distributive, ou sur l'esprit humain même). Adopter des mesures appropriées pour atténuer ces risques, le cas échéant, d'une manière proportionnée à l'ampleur du risque.

[1] Tous les éléments normatifs du présent document ont pour but de refléter les orientations destinées à réaliser les deuxième et troisième caractéristiques d'une IA digne de confiance (une IA éthique et robuste). Ces éléments ne sont par conséquent pas destinés à fournir des conseils juridiques ou à proposer des orientations en matière de conformité avec la législation applicable, bien qu'il soit reconnu qu'une part importante de ces éléments sont dans une certaine mesure déjà présents dans la législation existante. Voir point 21 et suivants à cet égard.

[2] Voir articles 24 à 27 de la charte des droits fondamentaux de l'Union européenne (charte de l'UE), portant sur les droits de l'enfant et des personnes âgées, l'intégration des personnes handicapées et les droits des travailleurs. Voir également l'article 38 portant sur la protection des consommateurs.

II. S'appuyant sur le chapitre I, le chapitre II fournit des orientations sur la manière dont une IA digne de confiance peut être réalisée, en présentant **sept exigences** que tout système d'IA devrait respecter. Des méthodes tant techniques que non techniques peuvent être utilisées aux fins de leur mise en œuvre.

Orientations essentielles dérivées du chapitre II:

- ✓ Veiller à ce que la mise au point, le déploiement et l'utilisation de systèmes d'IA répondent aux exigences d'une IA digne de confiance: 1) action humaine et contrôle humain, 2) robustesse technique et sécurité, 3) respect de la vie privée et gouvernance des données, 4) transparence, 5) diversité, non-discrimination et équité, 6) bien-être sociétal et environnemental, et 7) responsabilité.
- ✓ Envisager des méthodes techniques et non techniques pour garantir la mise en œuvre de ces exigences.
- ✓ Encourager la recherche et l'innovation en vue de contribuer à l'évaluation des systèmes d'IA et de soutenir la mise en œuvre des exigences; diffuser les résultats et les questions ouvertes au grand public, et veiller à ce qu'une formation dans le domaine l'éthique en matière d'IA soit systématiquement dispensée à la nouvelle génération d'experts.
- ✓ Fournir de façon proactive des informations claires aux parties prenantes sur les capacités et les limites des systèmes d'IA, afin de leur permettre de formuler des attentes réalistes, ainsi que sur la manière dont les exigences sont mises en œuvre. Faire preuve de transparence sur le fait qu'elles interagissent avec un système d'IA.
- ✓ Faciliter la traçabilité et l'auditabilité des systèmes d'IA, en particulier dans les contextes ou situations critiques.
- ✓ Associer les parties prenantes tout au long du cycle de vie du système d'IA. Encourager la formation et l'éducation afin que toutes les parties prenantes soient renseignées sur l'IA digne de confiance et formées dans ce domaine.
- ✓ Savoir qu'il peut exister des tensions fondamentales entre différents principes et exigences. Recenser, évaluer, documenter et communiquer de manière continue ces arbitrages et leurs solutions.

III. Le chapitre III fournit une liste d'évaluation concrète mais non exhaustive pour une IA digne de confiance, qui vise à concrétiser les exigences définies au chapitre II. Cette **liste d'évaluation** devra être adaptée au cas d'utilisation spécifique du système d'IA.[3]

Orientations essentielles dérivées du chapitre III:

- ✓ Adopter une évaluation pour une IA digne de confiance lors de la mise au point, du déploiement ou de l'utilisation de systèmes d'IA, et l'adapter au cas d'utilisation spécifique du système.
- ✓ Garder à l'esprit qu'une liste d'évaluation de cette nature ne sera jamais exhaustive. Il ne suffit pas de cocher des cases pour garantir une IA digne de confiance. Il convient de déterminer des exigences et de les mettre en œuvre, d'évaluer des solutions et de veiller à améliorer les résultats tout au long du cycle de vie du système d'IA, et d'y associer les parties prenantes.

(3) La section finale du document vise à concrétiser certaines des questions abordées dans l'ensemble du cadre, en présentant des exemples de possibilités bénéfiques qu'il convient de mettre en œuvre, et les grandes préoccupations soulevées par les systèmes d'IA qu'il convient d'examiner avec soin.

(4) Si l'objectif des présentes lignes directrices est de proposer des orientations relatives aux applications de l'IA en général, en érigeant une base transversale pour parvenir à une IA digne de confiance, des situations différentes posent des défis différents. Il convient par conséquent d'examiner si, en plus de ce cadre

[3] Conformément au champ d'application du cadre établi au point 2, cette liste d'évaluation ne fournit aucun conseil pour veiller à la conformité juridique (IA licite), mais se limite à proposer des orientations pour réaliser les deuxième et troisième caractéristiques d'une IA digne de confiance (une IA éthique et robuste).

transversal, une approche sectorielle est nécessaire, étant donné la mesure dans laquelle les systèmes d'IA sont spécifiques à leurs contextes.

(5) Les présentes lignes directrices ne visent ni à remplacer toute forme actuelle ou future d'élaboration de politiques ou de réglementations ni à en décourager l'introduction. Il faut les considérer comme un document évolutif qu'il conviendra de réviser et mettre à jour au fil du temps afin d'en maintenir la pertinence, à mesure que la technologie, nos environnements sociaux et nos connaissances évolueront. Le présent document est conçu comme le point de départ de la discussion sur «Une IA digne de confiance pour l'Europe».[4] Au-delà de l'Europe, les présentes lignes directrices visent également à encourager la recherche, la réflexion et la discussion sur un cadre éthique pour les systèmes d'IA au niveau mondial.

[4] Cet idéal est destiné à être appliqué aux systèmes d'IA mis au point, déployés et utilisés dans les États membres de l'Union européenne, ainsi qu'aux systèmes mis au point ou produits ailleurs mais déployés et utilisés au sein de l'UE. Lorsqu'il est fait référence à l'«Europe» dans le présent document, ce sont les États membres de l'Union qui sont visés. Toutefois, les présentes lignes directrices aspirent également à être pertinentes en dehors de l'Union. À cet égard, il convient de noter que la Norvège et la Suisse font partie du plan coordonné dans le domaine de l'IA adopté et publié en décembre 2018 par la Commission et les États membres.

A. INTRODUCTION

(6) Dans ses communications du 25 avril 2018 et du 7 décembre 2018, la Commission européenne (ci-après la «Commission») définit sa vision pour l'intelligence artificielle (IA), qui préconise une «IA éthique, sûre et de pointe réalisée en Europe».[5] La vision de la Commission repose sur trois piliers: i) accroître les investissements publics et privés dans l'IA afin d'intensifier le recours à l'IA, ii) se préparer aux changements socioéconomiques, et iii) garantir un cadre éthique et juridique approprié afin de renforcer les valeurs européennes.

(7) Pour soutenir la mise en œuvre de cette vision, la Commission a mis sur pied le groupe d'experts de haut niveau sur l'intelligence artificielle (GEHN IA), un groupe indépendant chargé d'élaborer deux contributions: 1) des lignes directrices en matière d'éthique, et 2) des recommandations en matière de politique et d'investissement dans le domaine de l'IA.

(8) Le présent document contient les lignes directrices en matière d'éthique dans le domaine de l'IA, qui ont été révisées à la suite de nouvelles délibérations de notre groupe à la lumière des commentaires reçus dans le cadre de la consultation publique relative au projet publié le 18 décembre 2018. Il s'appuie en outre sur les travaux du Groupe européen d'éthique des sciences et des nouvelles technologies[6] et s'inspire d'autres efforts similaires.[7]

(9) Au cours des derniers mois, nos 52 membres se sont réunis, ont discuté et ont interagi, sans déroger à la devise européenne: «Unie dans la diversité». Nous sommes convaincus que l'IA est susceptible de transformer la société de manière significative. L'IA n'est pas une fin en soi, mais plutôt un moyen prometteur d'accroître la prospérité humaine, en renforçant ainsi le bien-être individuel et de la société ainsi que le bien commun, et en étant porteur de progrès et d'innovation. Les systèmes d'IA peuvent notamment contribuer à faciliter la réalisation des objectifs de développement durable des Nations unies, tels que promouvoir l'égalité entre les sexes et lutter contre le changement climatique, rationaliser notre utilisation des ressources naturelles, améliorer notre santé, notre mobilité et nos processus de production, et nous aider à surveiller nos progrès par rapport à des indicateurs de durabilité et de cohésion sociale.

(10) Pour parvenir à ces objectifs, les systèmes d'IA[8] doivent être **centrés sur l'humain**, en s'appuyant sur l'engagement de mettre leur utilisation au service de l'humanité et du bien commun, avec pour objectif d'améliorer le bien-être et la liberté des êtres humains. S'ils offrent de brillantes possibilités, les systèmes d'IA soulèvent également certains risques qui doivent être traités de manière appropriée et proportionnée. Nous sommes à présent face à une occasion unique de façonner leur élaboration. Nous voulons pouvoir nous fier aux environnements sociotechniques auxquels ils sont intégrés, et nous voulons que les concepteurs de systèmes d'IA obtiennent un avantage concurrentiel en intégrant une IA digne de confiance à leurs produits et services. Cet objectif nécessite de chercher à **optimiser les avantages offerts par les systèmes d'IA** tout en veillant **à prévenir et réduire le plus possible les risques qu'ils présentent.**

(11) Dans un contexte d'évolution technologique rapide, nous sommes convaincus qu'il est essentiel que la confiance reste le ciment des sociétés, des communautés, des économies et du développement durable. Nous

---

5 COM(2018) 237 et COM(2018) 795. Il convient de noter que le terme «made in Europe» est employé par la Commission dans sa communication. Le champ d'application des présentes lignes directrices englobe non seulement les systèmes d'IA réalisés en Europe, mais également ceux mis au point ailleurs et qui sont déployés ou utilisés en Europe. Tout au long de ce document, nous nous efforçons donc de promouvoir une IA digne de confiance «pour» l'Europe.

6 Le Groupe européen d'éthique des sciences et des nouvelles technologies (GEE) est un groupe consultatif de la Commission.

7 Voir section 3.3 du document COM(2018) 237.

8 Le glossaire figurant à la fin du présent document fournit une définition des systèmes d'IA aux fins de ce même document. Cette définition est davantage détaillée dans un document spécifique élaboré par le GEHN IA et accompagnant les présentes lignes directrices, intitulé «A definition of AI: Main capabilities and scientific disciplines» (Définition de l'IA: principales capacités et disciplines scientifiques).

avons ainsi fait de **l'IA digne de confiance notre ambition fondatrice**; étant donné que les êtres humains et les communautés ne pourront avoir confiance dans le développement de la technologie et dans ses applications que lorsqu'un cadre clair et exhaustif pour la rendre digne de confiance sera en place.

(12) Il s'agit, de notre point de vue, de la voie que devrait suivre l'Europe pour se positionner comme foyer et leader d'une technologie éthique et de pointe. C'est grâce à une IA digne de confiance que, en tant que citoyens européens, nous pourrons bénéficier de ses avantages d'une manière qui reflète nos valeurs fondamentales que sont le respect des droits de l'homme, la démocratie et l'état de droit.

IA digne de confiance

(13) La fiabilité est une condition préalable pour que les personnes et les sociétés mettent au point, déploient et utilisent des systèmes d'IA. S'ils ne démontrent pas qu'ils sont dignes de confiance, les systèmes d'IA – et les êtres humains qui les conçoivent – pourraient être à l'origine de conséquences indésirables susceptibles de nuire à leur utilisation, ce qui empêcherait la réalisation des avantages sociaux et économiques potentiellement vastes qu'apportent les systèmes d'IA. Pour aider l'Europe à obtenir la réalisation de ces avantages, notre vision consiste à faire de l'éthique un pilier essentiel pour garantir et développer une IA digne de confiance.

(14) La confiance dans la mise au point, le déploiement et l'utilisation de systèmes d'IA concerne non seulement les propriétés intrinsèques de la technologie, mais également les qualités des systèmes sociotechniques impliquant des applications d'IA.[9] De manière analogue à des questions de (perte de) confiance dans l'aviation, l'énergie nucléaire ou la sécurité alimentaire, ce ne sont pas uniquement les composantes des systèmes d'IA qui pourraient ou non susciter la confiance, mais le système dans son contexte global. La quête d'une IA digne de confiance concerne donc non seulement la fiabilité du système d'IA en tant que tel, mais requiert également une approche globale et systémique qui englobe la fiabilité de l'ensemble des acteurs et processus qui composent le contexte sociotechnique du système tout au long de son cycle de vie.

(15) Une IA digne de confiance comporte les **trois éléments** suivants, qui doivent être présents tout au long du cycle de vie du système:

1. elle doit être **licite**, en assurant le respect des législations et réglementations applicables;
2. elle doit être **éthique**, en assurant l'adhésion à des principes et valeurs éthiques, et
3. elle doit être **robuste**, sur le plan tant technique que social car, même avec de bonnes intentions, les systèmes d'IA peuvent causer des préjudices involontaires.

(16) Toutes ces caractéristiques sont nécessaires, mais elles ne sauraient suffire à la réalisation d'une IA digne de confiance[10]. L'idéal serait que ces trois caractéristiques fonctionnent en harmonie et se chevauchent. Toutefois, dans la pratique, des tensions peuvent survenir entre ces éléments (par exemple, dans certains cas, le champ d'application et le contenu de la législation existante pourraient ne pas correspondre à des normes éthiques). Il en va de notre responsabilité individuelle et collective en tant que société de veiller à ce que chacune de ces trois caractéristiques contribue à garantir l'avènement d'une IA digne de confiance.[11]

(17) Une approche digne de confiance est essentielle pour permettre une «compétitivité responsable», en établissant les bases sur lesquelles les personnes concernées par des systèmes d'IA peuvent se fier au caractère licite, éthique et robuste de leur conception, de leur mise au point et de leur utilisation. Les présentes lignes directrices visent à encourager une innovation responsable et durable dans le domaine de l'IA

[9] Ces systèmes se composent d'êtres humains, d'acteurs étatiques, d'entreprises, d'infrastructures, de logiciels, de protocoles, de normes, de gouvernance, de législations existantes, de mécanismes de contrôle, de structures d'incitation, de procédures d'audit, de meilleures pratiques, de documentation, et d'autres éléments.

[10] Cela n'exclut pas le fait que des conditions supplémentaires pourraient être (ou devenir) nécessaires.

[11] Cela signifie également que le législateur ou les décideurs politiques pourraient être amenés à revoir le caractère approprié de la législation en vigueur lorsque celle-ci pourrait ne pas correspondre à des principes éthiques.

en Europe. Elles cherchent à ériger l'éthique en pilier essentiel de la mise au point d'une approche unique de l'IA cherchant à favoriser, renforcer et protéger tant la prospérité individuelle des êtres humains que le bien commun de la société. Nous sommes convaincus que cela permettra à l'Europe de s'imposer comme leader mondial d'une IA de pointe, digne de notre confiance individuelle et collective. Ce n'est que si la fiabilité des systèmes d'IA est garantie que les citoyens européens pourront bénéficier pleinement de ses avantages, forts de la conviction que des mesures sont en place pour les protéger contre les risques potentiels.

(18) Tout comme l'utilisation de systèmes d'IA ne s'arrête pas aux frontières nationales, leurs incidences ne s'y arrêtent pas davantage. Des solutions mondiales sont par conséquent nécessaires face aux possibilités et aux défis mondiaux que présente l'IA. Nous encourageons par conséquent l'ensemble des parties prenantes à travailler à l'élaboration d'un cadre mondial pour une IA digne de confiance, en cherchant un consensus international tout en encourageant et en préservant notre approche fondée sur le respect des droits fondamentaux.

Public et champ d'application

(19) Les présentes lignes directrices sont destinées à l'ensemble des parties prenantes de l'IA qui conçoivent, mettent au point, déploient, mettent en œuvre, utilisent l'IA ou sont soumises à ses incidences, et notamment aux entreprises, aux organisations, aux chercheurs, aux services publics, organismes gouvernementaux, institutions, organisations de la société civile, particuliers, travailleurs et consommateurs. Les parties prenantes résolues à réaliser une IA digne de confiance peuvent librement décider d'utiliser les présentes lignes directrices comme méthode pour concrétiser leur engagement, notamment en ayant recours à la liste d'évaluation pratique du chapitre III dans leurs processus de mise au point et de déploiement de systèmes d'IA. Cette liste d'évaluation peut également compléter, et donc intégrer, les processus d'évaluation existants.

(20) L'objectif des présentes lignes directrices est de proposer des orientations relatives aux applications d'IA en général, en érigeant une base transversale pour parvenir à une IA digne de confiance. Toutefois, **des situations différentes posent des défis différents.** Les systèmes d'IA de recommandation musicale ne soulèvent pas les mêmes préoccupations éthiques que les systèmes d'IA proposant des traitements médicaux essentiels. De même, les systèmes d'IA utilisés dans le contexte des relations d'entreprise à consommateur, d'entreprise à entreprise, d'employeur à employé et de la sphère publique aux citoyens ou, plus généralement, dans différents secteurs ou cas d'utilisation, présentent des possibilités et des défis différents. Les systèmes d'IA étant propres à leur contexte, il est par conséquent reconnu que la mise en œuvre des présentes lignes directrices doit être adaptée à l'application spécifique de l'IA. Il convient en outre d'examiner la mesure dans laquelle une approche sectorielle supplémentaire pourrait être nécessaire pour compléter le cadre transversal plus général proposé dans le présent document.

Afin de mieux comprendre la manière dont ces orientations peuvent être mises en œuvre au niveau transversal, ainsi que les questions qui requièrent une approche sectorielle, nous invitons l'ensemble des parties prenantes à tester la liste d'évaluation pour une IA digne de confiance (chapitre III) concrétisant ce cadre et à nous communiquer leurs observations. Sur la base des commentaires recueillis lors de cette phase pilote, nous réviserons la liste d'évaluation des présentes lignes directrices d'ici le début de 2020. La phase pilote débutera d'ici l'été 2019 et se prolongera jusqu'à la fin de l'année. Toutes les parties prenantes intéressées auront la possibilité de participer en manifestant leur intérêt via l'Alliance européenne pour l'IA.

B. CADRE POUR UNE IA DIGNE DE CONFIANCE

(21) Les présentes lignes directrices définissent un cadre pour parvenir à la mise en œuvre d'une IA digne de confiance fondée sur les droits fondamentaux tels que consacrés dans la charte des droits fondamentaux de l'Union européenne (charte de l'UE), et dans le droit international pertinent en matière de droits de l'homme. Ci-dessous, nous abordons brièvement les trois caractéristiques d'une IA digne de confiance.

IA licite

(22) Les systèmes d'IA ne sont pas mis en œuvre dans un monde sans loi. Un ensemble de règles contraignantes aux niveaux européen, national et international s'appliquent déjà ou sont pertinentes dans le cadre de la mise au point, du déploiement et de l'utilisation de systèmes d'IA. Les sources de droit pertinentes comprennent, sans s'y limiter, le droit primaire de l'Union (les traités de l'Union européenne et sa charte des droits fondamentaux), le droit dérivé de l'Union (comme le règlement général sur la protection des données, les directives antidiscrimination, la directive «Machines», la directive sur la responsabilité du fait des produits, le règlement sur la libre circulation des données à caractère non personnel, les directives relatives au droit des consommateurs et à la sécurité et la santé au travail), mais également les traités en matière de droits de l'homme des Nations unies et les conventions du Conseil de l'Europe (telles que la Convention européenne des droits de l'homme) et de nombreuses autres législations des États membres de l'Union. Outre les règles applicables au niveau transversal, il existe différentes règles propres à un domaine donné qui s'appliquent à des applications d'IA particulières (comme le règlement relatif aux dispositifs médicaux dans le secteur des soins de santé).

(23) La législation prévoit des obligations tant positives que négatives; autrement dit, il convient de ne pas l'interpréter uniquement en lien avec ce qui ne *peut pas* être fait, mais aussi en lien avec ce qui *devrait* être fait. La législation ne se limite pas à interdire certaines actions mais en rend également d'autres possibles. À cet égard, il convient de noter que la charte de l'Union contient des articles relatifs à la «liberté d'entreprise» et à la «liberté des arts et des sciences», ainsi que des articles portant sur des domaines que nous connaissons mieux lorsqu'il s'agit de veiller à la fiabilité de l'IA, tels que la protection des données et la non-discrimination.

(24) Les lignes directrices ne traitent pas explicitement de la première caractéristique d'une IA digne de confiance (IA licite), mais visent plutôt à proposer des orientations pour encourager et garantir les deuxième et troisième caractéristiques (une IA éthique et robuste). Si ces deux dernières caractéristiques sont dans une certaine mesure déjà reflétées dans la législation existante, leur pleine réalisation pourrait aller au-delà des obligations juridiques existantes.

(25) Aucune partie du présent document ne peut s'entendre ou être interprétée comme fournissant des conseils ou orientations juridiques sur la manière de se mettre en conformité avec les normes et exigences juridiques existantes applicables. Aucun élément du présent document ne peut créer des droits ou imposer des obligations juridiques vis-à-vis de tiers. Nous rappelons toutefois que toute personne physique ou morale se doit de respecter la législation – qu'elle soit applicable aujourd'hui ou adoptée dans le futur en fonction de l'évolution de l'IA. Les présentes lignes directrices partent du principe que **l'ensemble des droits et obligations juridiques applicables aux processus et activités faisant partie de la mise au point, du déploiement et de l'utilisation de l'IA conservent un caractère obligatoire et doivent être dûment respectés.**

IA éthique

(26) Le respect du droit n'est qu'une des trois caractéristiques pour parvenir à la mise en œuvre d'une IA digne de confiance. La législation ne suit pas toujours le rythme des évolutions technologiques, ne correspond parfois pas à des normes éthiques ou peut simplement s'avérer inadaptée face à certaines questions. Pour être dignes de confiance, les systèmes d'IA devraient donc également être éthiques, en veillant à l'alignement sur les normes éthiques.

IA robuste

(27) Même lorsqu'une finalité éthique est garantie, les individus et la société doivent également être convaincus que les systèmes d'IA ne causeront pas de préjudice involontaire. Ces systèmes devraient être mis en œuvre de manière sûre, sécurisée et fiable, et il importe de prévoir des garanties pour éviter les incidences négatives involontaires. Il est par conséquent important de veiller à la robustesse des systèmes d'IA. Cet élément est nécessaire tant sur le plan technique (veiller à la robustesse technique du système selon les besoins dans un contexte donné, tel que le domaine d'application ou la phase du cycle de vie), que sur le plan social (en tenant

dûment compte du contexte et de l'environnement dans lesquels le système fonctionne). L'éthique et la robustesse de l'IA sont donc étroitement liées et se complètent mutuellement. Les principes mis en avant au chapitre I, ainsi que les exigences qui en découlent au chapitre II, portent sur ces deux caractéristiques.

Le cadre

(28) Les orientations du présent document se présentent sous la forme de trois niveaux d'abstraction, du plus abstrait, au chapitre I, au plus concret, au chapitre III:

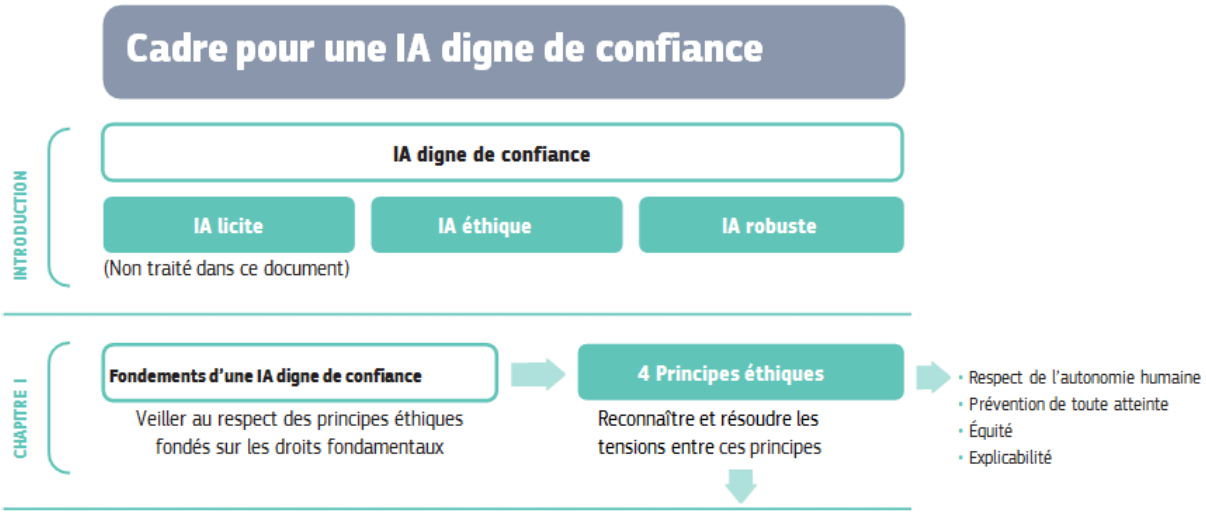
I) Fondements d'une IA digne de confiance. Le chapitre I établit les fondements d'une IA digne de confiance, en définissant son approche fondée sur le respect des droits fondamentaux[12]. Il recense et décrit les principes éthiques auxquels il convient d'adhérer afin de garantir une IA éthique et robuste.

II) Parvenir à une IA digne de confiance. Le chapitre II traduit ces principes éthiques en sept exigences que les systèmes d'IA devraient mettre en œuvre et respecter tout au long de leur cycle de vie. En outre, il propose des méthodes tant techniques que non techniques pouvant être appliquées aux fins de leur mise en œuvre.

III) Évaluer une IA digne de confiance. Les professionnels de l'IA attendent des orientations concrètes. Le chapitre III établit par conséquent une liste d'évaluation préliminaire et non exhaustive pour une IA digne de confiance afin de concrétiser les exigences du chapitre II. Cette évaluation devrait être adaptée à l'application spécifique du système.

(29) La dernière section du présent document expose des possibilités bénéfiques et des préoccupations importantes suscitées par les systèmes d'IA dont il convient de tenir compte et sur lesquelles nous souhaitons encourager de nouvelles discussions.

(30) La structure des présentes lignes directrices est illustrée à la *figure 1* ci-dessous.



[12] Les droits fondamentaux sont le fondement du droit tant international que de l'Union en matière de droits de l'homme et sous-tendent les droits opposables garantis par les traités de l'Union et par la charte des droits fondamentaux de l'Union européenne. Les droits fondamentaux étant juridiquement contraignants, leur respect relève donc de la première caractéristique d'une IA digne de confiance, à savoir une «IA licite». Les droits fondamentaux peuvent toutefois être interprétés comme reflétant aussi des droits moraux spéciaux reconnus à l'ensemble des individus en vertu de leur humanité, que ces droits soient ou non juridiquement contraignants. En ce sens, ils relèvent également de la deuxième caractéristique d'une IA digne de confiance, à savoir une «IA éthique».

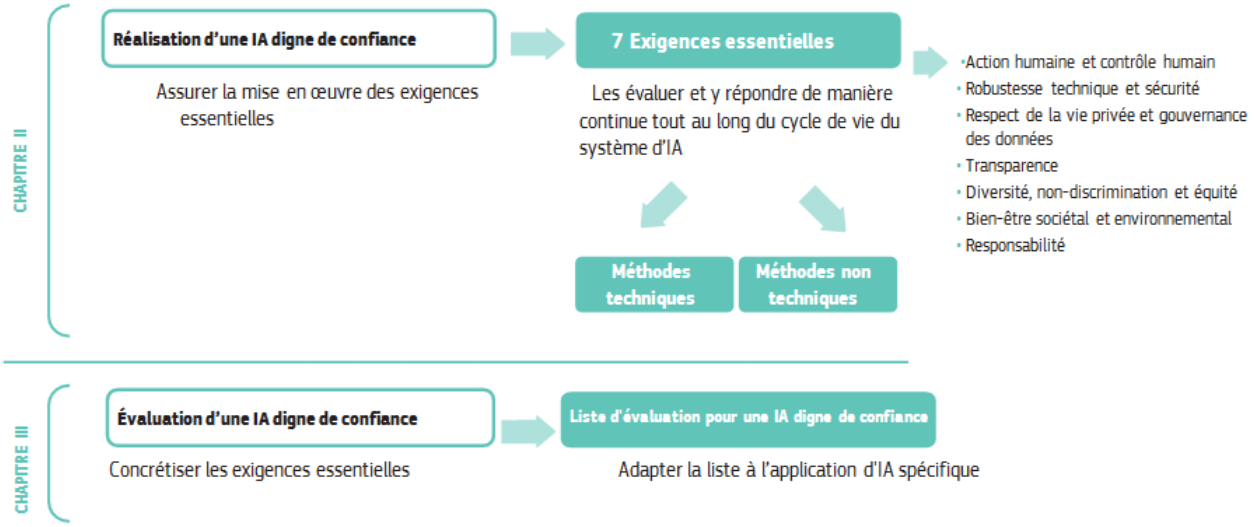


Figure 1: les lignes directrices en tant que cadre pour une IA digne de confiance

I. Chapitre I: Fondements d'une IA digne de confiance

(31) Ce chapitre établit les fondements d'une IA digne de confiance, reposant sur les droits fondamentaux et reflétée par quatre principes éthiques auxquels il convient d'adhérer afin de garantir une IA éthique et robuste. Ce chapitre s'appuie fortement sur le domaine de l'éthique.

(32) L'éthique en matière d'IA est un sous-domaine de l'éthique appliquée qui est axé sur les questions d'ordre éthique soulevées par la mise au point, le déploiement et l'utilisation de l'IA. Sa préoccupation centrale consiste à déterminer la manière dont l'IA peut soulever des préoccupations relatives au bien-être des individus ou y apporter des solutions, que ce soit du point de vue de la qualité de vie ou de l'autonomie humaine et de la liberté nécessaire pour une société démocratique.

(33) Une réflexion éthique sur la technologie de l'IA peut servir plusieurs objectifs. Premièrement, elle peut stimuler la réflexion sur la nécessité de protéger les individus et les groupes au niveau le plus élémentaire. Deuxièmement, elle peut stimuler de nouveaux genres d'innovation dont l'objectif est de promouvoir des valeurs éthiques, telles que celles contribuant à la réalisation des objectifs de développement durable des Nations unies[13], qui sont fermement ancrés dans le futur programme de l'Union européenne à l'horizon 2030[14]. Si le présent document porte principalement sur le premier objectif mentionné, il ne faut pas sous-estimer l'importance que pourrait revêtir l'éthique dans le cadre du deuxième objectif. Une IA digne de confiance peut renforcer la prospérité des individus et le bien-être collectif en générant de la prospérité, en créant de la valeur et en maximisant les richesses. Elle peut contribuer à la réalisation d'une société juste, en contribuant à l'amélioration de la santé et du bien-être des citoyens d'une manière qui renforce l'égalité dans la répartition des possibilités économiques, sociales et politiques.

(34) Il est par conséquent impératif que nous comprenions comment soutenir au mieux la mise au point, le déploiement et l'utilisation de l'IA pour faire en sorte que chacun puisse s'épanouir dans un monde fondé sur l'IA, et pour préparer un avenir meilleur tout en préservant la compétitivité au niveau mondial. Comme toute technologie puissante, l'utilisation de systèmes d'IA au sein de notre société soulève plusieurs problèmes éthiques, par exemple en ce qui concerne leur incidence sur les individus et la société, les capacités de prise de décision et la sécurité. Si nous prévoyons de nous faire assister par des systèmes d'IA ou de leur déléguer de plus en plus de décisions, nous devons veiller à ce que l'incidence de ces systèmes sur la vie des personnes soit équitable, à ce que ces systèmes soient conformes aux valeurs inaliénables et capables d'agir en ce sens, ainsi qu'à l'existence de processus adaptés en matière de responsabilisation pour y veiller.

(35) L'Europe doit définir la vision normative qu'elle souhaite mettre en œuvre pour un avenir marqué par l'omniprésence de l'IA et, par conséquent, comprendre quelle notion de l'IA devrait être étudiée, mise au point, déployée et utilisée en Europe pour réaliser cette vision. Avec ce document, nous souhaitons contribuer à cet effort en introduisant la notion d'IA digne de confiance qui est, selon nous, la manière adaptée de bâtir un avenir avec l'IA. Un avenir dans lequel la démocratie, l'état de droit et les droits fondamentaux sous-tendent les systèmes d'IA et dans lequel ces systèmes améliorent et défendent de manière continue la culture démocratique permettra également de mettre en place un environnement dans lequel l'innovation et la compétitivité responsable peuvent se développer.

(36) Un code de déontologie spécifique à un domaine donné – quel que soit le niveau de cohérence, d'élaboration et de détail de ses futures versions – ne pourra jamais se substituer à un raisonnement éthique en tant que tel, qui doit en toutes circonstances rester sensible aux éléments de contexte, qui ne peuvent jamais être rendus dans des lignes directrices générales. En plus d'élaborer un ensemble de règles, il convient de mettre sur pied et de conserver une culture et un état d'esprit éthiques dans le débat public, l'éducation et l'apprentissage pratique pour garantir une IA digne de confiance.

---

13 https://ec.europa.eu/commission/publications/reflection-paper-towards-sustainable-europe-2030_fr.

14 https://sustainabledevelopment.un.org/?menu=1300.

1. Les droits fondamentaux en tant que droits moraux et légaux

(37) Nous croyons en une approche de l'éthique en matière d'IA qui est fondée sur les droits fondamentaux consacrés par les traités de l'Union,[15] la charte des droits fondamentaux de l'Union européenne (charte de l'Union) et le droit international en matière de droits de l'homme.[16]Le respect des droits fondamentaux, dans le cadre de la démocratie et de l'état de droit, est le fondement le plus prometteur pour recenser les principes et les valeurs éthiques abstraits pouvant être concrétisés dans le contexte de l'IA.

(38) Les traités de l'Union et la charte de l'Union prescrivent un ensemble de droits fondamentaux que les États membres et les institutions de l'UE sont juridiquement tenus de respecter dans le cadre de la mise en œuvre du droit de l'Union. Ces droits sont décrits dans la charte de l'Union en référence à la dignité, aux libertés, à l'égalité et la solidarité, aux droits des citoyens et à la justice. Ces droits ont pour base commune un ancrage dans le respect de la dignité humaine, reflété par ce que nous décrivons comme une «approche centrée sur l'humain», dans laquelle l'être humain jouit d'un statut moral unique et inaliénable de primauté dans les domaines civil, politique, économique et social.[17]

(39) Alors que les droits établis dans la charte de l'Union sont juridiquement contraignants,[18] il est important de reconnaître que les droits fondamentaux n'assurent pas dans tous les cas une protection juridique complète. En ce qui concerne par exemple la charte de l'Union, il est important de souligner que son champ d'application se limite aux domaines couverts par le droit de l'Union. Le droit international en matière de droits de l'homme et notamment la Convention européenne des droits de l'homme sont juridiquement contraignants pour les États membres de l'UE, y compris dans les domaines qui sortent du champ d'application du droit de l'Union. Dans le même temps, il convient de souligner que des droits fondamentaux sont également conférés aux individus et (dans une certaine mesure) aux groupes en vertu de leur statut moral en tant qu'êtres humains, indépendamment de leur force juridique. Interprétés comme droits opposables, les droits fondamentaux relèvent par conséquent de la première caractéristique d'une IA digne de confiance (IA licite), qui garantit la conformité avec le droit. Interprétés comme les droits de chacun, ancrés dans le statut moral inhérent aux êtres humains, ils sous-tendent également la deuxième caractéristique d'une IA digne de confiance (IA éthique), qui porte sur des normes éthiques qui, sans être nécessairement contraignantes sur le plan juridique, sont pourtant essentielles pour parvenir à une IA digne de confiance. Étant donné que le présent document n'a pas vocation à fournir des orientations relatives à la première caractéristique, aux fins des présentes orientations non contraignantes, les références aux droits fondamentaux reflètent la deuxième caractéristique.

2. Des droits fondamentaux aux principes éthiques

2.1 Les droits fondamentaux comme base d'une IA digne de confiance

(40) Parmi l'éventail complet de droits indivisibles énoncés dans le droit international en matière de droits de l'homme, les traités de l'Union et la charte de l'Union, les familles de droits fondamentaux mentionnées ci-après sont particulièrement adaptées à une application aux systèmes d'IA. Une part importante de ces droits sont, dans des circonstances définies, opposables au sein de l'UE, ce qui rend juridiquement obligatoire la

[15] L'UE est fondée sur l'engagement constitutionnel de protéger les droits fondamentaux et indivisibles des êtres humains, de veiller au respect de l'état de droit, d'encourager la liberté démocratique et de promouvoir le bien commun. Ces droits sont reflétés aux articles 2 et 3 du traité sur l'Union européenne, ainsi que dans la charte des droits fondamentaux de l'Union européenne.

[16] D'autres instruments juridiques reflètent et précisent ces engagements, comme la charte sociale européenne du Conseil de l'Europe ou des législations spécifiques telles que le règlement général sur la protection des données de l'Union.

[17] Il convient de noter qu'un engagement envers une IA centrée sur l'humain et son ancrage dans les droits fondamentaux, plutôt que de supposer une valeur indûment individualiste de l'humain, nécessite des fondements sociétaux et constitutionnels collectifs dans lesquels la liberté individuelle et le respect de la dignité humaine sont à la fois possibles et pertinents.

[18] Conformément à l'article 51 de la charte, celle-ci s'applique aux institutions et aux États membres de l'Union lorsqu'ils mettent en œuvre le droit de l'Union.

conformité avec leurs exigences. Toutefois, même lorsque la conformité avec les droits fondamentaux opposables a été atteinte, une réflexion éthique peut nous aider à comprendre de quelle manière la mise au point, le déploiement et l'utilisation de l'IA peuvent mettre en jeu les droits fondamentaux et leurs valeurs sous-jacentes, et peuvent contribuer à des orientations plus précises lorsqu'il s'agit de déterminer ce que nous *devrions* faire plutôt que ce que nous *pouvons* faire (actuellement) à l'aide de la technologie.

(41) ***Respect de la dignité humaine***. La dignité humaine comprend l'idée que chaque être humain possède une «valeur intrinsèque», qui ne devrait jamais être diminuée, compromise ou réprimée par autrui – ni par de nouvelles technologies telles que des systèmes d'IA.[19] Dans le contexte de l'IA, le respect de la dignité humaine signifie que chaque personne est traitée avec respect du fait de son statut de *sujet* moral, plutôt que comme simple *objet* que l'on trie, classe, marque, régente, conditionne ou manipule. Les systèmes d'IA devraient donc être mis au point de manière à respecter, protéger et servir l'intégrité physique et mentale des êtres humains, leur sentiment d'identité personnel et culturel et la satisfaction de leurs besoins essentiels.[20]

(42) ***Liberté des individus.*** Les êtres humains devraient rester libres de faire leurs propres choix de vie. Cela suppose l'absence d'intrusion du pouvoir, mais requiert également l'intervention des pouvoirs publics et des organisations non gouvernementales pour faire en sorte que les individus exposés au risque d'exclusion jouissent d'une égalité d'accès aux avantages et aux possibilités que présente l'IA. Dans un contexte d'IA, la liberté des individus requiert l'atténuation des contraintes illégitimes, des menaces à l'encontre de l'autonomie et de la santé mentales, de la surveillance injustifiée, de la tromperie et de la manipulation injuste, que ces atteintes soient directes ou indirectes. En fait, la liberté des individus signifie un engagement visant à permettre aux individus d'exercer un contrôle accru sur leurs vies, y compris (entre autres droits) la protection de la liberté d'entreprise, la liberté des arts et des sciences, la liberté d'expression, le droit à la vie privée et à la confidentialité, et la liberté de réunion et d'association.

(43) ***Respect de la démocratie, de la justice et de l'état de droit.*** Dans les démocraties constitutionnelles, tout pouvoir gouvernemental doit être légalement autorisé et limité par la loi. Les systèmes d'IA devraient servir à conserver et à encourager les processus démocratiques et le respect de la pluralité des valeurs et des choix de vie des individus. Les systèmes d'IA ne doivent pas compromettre les processus démocratiques, la délibération humaine ou les systèmes de vote démocratiques. Les systèmes d'IA doivent également intégrer l'engagement de veiller à ce qu'ils ne soient pas mis en œuvre d'une manière qui compromette les engagements fondamentaux sur lesquels se fondent l'état de droit, les législations et règlements contraignants, et de garantir le droit à une procédure régulière et à l'égalité en droit.

(44) ***Égalité, non-discrimination et solidarité*** – *y compris le droit des personnes exposées au risque d'exclusion.* Il convient d'assurer un respect égal de la valeur morale et de la dignité de tous les êtres humains. Cela va au-delà de la non-discrimination, qui tolère le fait d'établir des distinctions entre des situations différentes sur la base de justifications objectives. Dans un contexte d'IA, l'égalité implique que le fonctionnement du système ne peut pas produire de résultats fondés sur des biais injustes (par exemple, les données utilisées pour entraîner les systèmes d'IA devraient être aussi inclusives que possible et représenter différents groupes de population), ce qui requiert également un respect approprié des personnes et des groupes potentiellement vulnérables[21], tels que les travailleurs, les femmes, les personnes handicapées, les minorités ethniques, les enfants, les consommateurs ou d'autres catégories de personnes exposées au risque d'exclusion.

(45) ***Droits des citoyens.*** Les citoyens bénéficient d'un large éventail de droits, dont le droit de vote, le droit à une bonne administration ou à l'accès aux documents publics, et le droit d'adresser des pétitions à l'administration. Grâce aux systèmes d'IA, les pouvoirs publics seront en mesure de fournir à la société des

[19] C. McCrudden, Human Dignity and Judicial Interpretation of Human Rights, *EJIL*, 19(4), 2008.

[20] Pour comprendre la «dignité humaine» en ce sens, voir E. Hilgendorf, Problem Areas in the Dignity Debate and the Ensemble Theory of Human Dignity, dans: D. Grimm, A. Kemmerer, C. Möllers (eds.), *Human Dignity in Context. Explorations of a Contested Concept*, 2018, pp. 325 et suiv.

[21] Pour une description de ce terme tel qu'il est employé dans le présent document, voir le glossaire.

biens et des services publics à une échelle et avec une efficience supérieures. Dans le même temps, les applications d'IA pourraient aussi avoir une incidence négative sur les droits des citoyens; il convient par conséquent de protéger ces droits. L'emploi du terme «droits des citoyens» dans le présent document ne signifie nullement que nous nions ou négligeons les droits des ressortissants de pays tiers et des personnes en situation irrégulière (ou illégale) sur le territoire de l'UE, qui jouissent également de droits au titre du droit international et, par conséquent, dans le domaine de l'IA.

2.2 Principes éthiques dans le contexte des systèmes d'IA[22]

(46) De nombreuses organisations publiques, privées et civiles se sont inspirées des droits fondamentaux pour élaborer des cadres éthiques pour les systèmes d'IA.[23] Dans l'UE, le Groupe européen d'éthique des sciences et des nouvelles technologies («GEE») a proposé un ensemble de neuf principes fondamentaux, reposant sur les valeurs fondamentales énoncées dans les traités de l'Union et dans la charte des droits fondamentaux de l'Union européenne.[24] Nous continuons à nous appuyer sur ces travaux, en reconnaissant la plupart des principes avancés jusqu'à présent par différents groupes, tout en précisant à quelles fins l'ensemble de ces principes cherchent à répondre et à apporter un soutien. Ces principes éthiques peuvent inspirer de nouveaux instruments réglementaires spécifiques, contribuer à l'interprétation des droits fondamentaux au fur et à mesure qu'évolue notre environnement sociotechnique et orienter les motifs justifiant la mise au point, l'utilisation et la mise en œuvre de systèmes d'IA – en s'adaptant de manière dynamique aux évolutions de la société elle-même.

(47) Les systèmes d'IA doivent améliorer le bien-être individuel et collectif. Cette section présente **quatre principes éthiques**, ancrés dans les droits fondamentaux, auxquels il convient d'adhérer pour faire en sorte que les systèmes d'IA soient mis au point, déployés et utilisés d'une manière digne de confiance. Ils sont présentés comme des **impératifs éthiques**, si bien que les professionnels de l'IA devraient en toutes circonstances s'efforcer d'y adhérer. Sans imposer de hiérarchie, nous présentons les principes ci-dessous de manière à refléter l'ordre d'apparition, dans la charte de l'Union, des droits fondamentaux sur lesquels ils se fondent.[25]

(48) Il s'agit des principes suivants:

(i) respect de l'autonomie humaine
(ii) prévention de toute atteinte
(iii) équité
(iv) explicabilité

(49) La plupart de ces principes sont dans une large mesure déjà reflétés dans les exigences juridiques contraignantes dont la mise en œuvre est obligatoire et relèvent donc également du champ d'application de l'«IA licite», soit la première caractéristique d'une IA digne de confiance.[26] Pourtant, comme indiqué plus haut, même si de nombreuses obligations juridiques reflètent des principes éthiques, l'adhésion à des principes

[22] Ces principes s'appliquent également à la mise au point, au déploiement et à l'utilisation d'autres technologies, et ne sont par conséquent pas spécifiques aux systèmes d'IA. Nous nous sommes efforcés ci-dessous d'établir leur pertinence dans un contexte spécifiquement lié à l'IA.

[23] Le recours aux droits fondamentaux contribue également à limiter l'insécurité réglementaire, car elle peut s'appuyer sur des décennies de pratique en matière de protection des droits fondamentaux dans l'UE, ce qui apporte de la clarté, de la lisibilité et de la prévisibilité.

[24] Plus récemment, le groupe de travail de AI4People a examiné les principes susmentionnés du GEE ainsi que 36 autres principes éthiques énoncés à ce jour et les a ramenés à quatre principes généraux. L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, E. J. M. Vayena (2018), «AI4People —An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations», Minds and Machines 28(4): p. 689-707.

[25] Le respect de l'autonomie humaine est fortement associé au droit à la dignité humaine et à la liberté (reflété aux articles 1 et 6 de la charte). La prévention de toute atteinte est fortement liée à la protection de l'intégrité physique ou mentale (reflétée à l'article 3). L'équité est étroitement liée aux droits à la non-discrimination, à la solidarité et à la justice (reflétés aux articles 21 et suivants). L'explicabilité et la responsabilité sont étroitement liées aux droits relatifs à la justice (tels que reflétés à l'article 47).

[26] On pense par exemple au RGPD ou aux règlements de l'Union relatifs à la protection des consommateurs.

éthiques dépasse le respect formel de la législation existante.[27]

- Le principe du respect de l'autonomie humaine

(50) Les droits fondamentaux sur lesquels l'UE est fondée ont vocation à garantir le respect de la liberté et de l'autonomie des êtres humains. Les êtres humains qui interagissent avec des systèmes d'IA doivent être en mesure de conserver leur autodétermination totale et effective et de prendre part au processus démocratique. En l'absence de justification, les systèmes d'IA ne devraient pas subordonner, contraindre, tromper, manipuler, conditionner ni régenter des êtres humains. Au contraire, les systèmes d'IA devraient être conçus afin d'augmenter, de compléter et de favoriser les compétences cognitives, sociales et culturelles. La répartition des tâches entre êtres humains et systèmes d'IA devrait suivre des principes de conception centrés sur l'humain et donner à l'être humain une possibilité réelle de poser des choix. En d'autres termes, il convient de veiller à la supervision[28] et au contrôle humains sur les processus de travail des systèmes d'IA. Les systèmes d'IA pourraient également modifier fondamentalement la sphère du travail. Ces systèmes devraient aider les êtres humains dans l'environnement de travail, et avoir pour objectif de créer des emplois qui aient du sens.

- Le principe de la prévention de toute atteinte

(51) Les systèmes d'IA ne devraient ni porter atteinte, ni aggraver toute atteinte portée[29], ni nuire aux êtres humains d'une quelconque autre manière.[30] Cela englobe la protection de la dignité humaine ainsi que de l'intégrité mentale et physique. Les systèmes d'IA et les environnements dans lesquels ils évoluent doivent être sûrs et sécurisés. Ils doivent être robustes sur le plan technique et il convient de veiller à ce qu'ils ne soient pas exposés à des utilisations malveillantes. Les personnes vulnérables devraient faire l'objet d'une attention accrue et être prises en compte dans la mise au point et le déploiement des systèmes d'IA. Il convient également d'accorder une attention particulière aux situations dans lesquelles les systèmes d'IA peuvent entraîner ou aggraver des incidences négatives du fait d'asymétries de pouvoir ou d'information, par exemple entre les employeurs et les travailleurs, entre les entreprises et les consommateurs ou entre les pouvoirs publics et les citoyens. La prévention de toute atteinte implique également la prise en compte de l'environnement naturel et de tous les êtres vivants.

- Le principe de l'équité

(52) La mise au point, le déploiement et l'utilisation de systèmes d'IA doivent être équitables. Si nous reconnaissons que l'équité peut s'interpréter de multiples manières, nous considérons que l'équité se caractérise à la fois par un volet matériel et un volet procédural. Le volet matériel suppose l'engagement de veiller à une répartition égale et juste des bénéfices et des coûts, et de veiller à ce que les individus et les groupes ne fassent pas l'objet de biais injustes, de discrimination et de stigmatisation. Si les biais injustes peuvent être évités, les systèmes d'IA pourraient même améliorer le caractère équitable de la société. Il convient également d'encourager l'égalité des chances en ce qui concerne l'accès à l'éducation, aux biens, aux services et à la technologie. En outre, l'utilisation de systèmes d'IA ne devrait jamais avoir pour conséquence de tromper les utilisateurs (finaux) ou de limiter leur liberté de choix. L'équité implique en outre que les professionnels de l'IA devraient respecter le principe de proportionnalité entre la fin et les moyens, et examiner de manière attentive la manière de trouver un équilibre entre des intérêts et des objectifs en

---

27 Pour d'autres références sur le sujet, voir par exemple L. Floridi, Soft Ethics and the Governance of the Digital, *Philosophy & Technology,* March 2018, Volume 31, Issue 1, pp 1–8.

28 Le concept du contrôle humain est approfondi au point 65 ci-dessous.

29 Une atteinte portée peut être individuelle ou collective, et peut comprendre une atteinte immatérielle aux environnements sociaux, culturels et politiques.

30 Les atteintes au mode de vie des individus et des groupes sociaux peuvent être qualifiées d'atteintes culturelles et doivent être évitées.

concurrence.[31] Le volet procédural de l'équité suppose la capacité de contester les décisions prises par des systèmes d'IA et par les êtres humains qui les utilisent, ainsi que celle d'introduire un recours efficace à l'encontre de ces décisions[32]. Pour ce faire, l'entité responsable de la décision doit pouvoir être identifiée, et le processus de prise de décisions devrait pouvoir être expliqué.

- Le principe de l'explicabilité

(53) L'explicabilité est essentielle pour renforcer et conserver la confiance des utilisateurs envers les systèmes d'IA. Cela signifie que les processus doivent être transparents, que les capacités et la finalité des systèmes d'IA doivent être communiquées ouvertement, et que les décisions – dans la mesure du possible – doivent pouvoir être expliquées aux personnes directement et indirectement concernées. Sans ces informations, une décision ne peut être dûment contestée. Il n'est pas toujours possible d'expliquer pour quelle raison un modèle a généré un résultat ou une décision en particulier (et quelle combinaison de facteurs d'entrée y a contribué). On parle d'algorithmes à effet «boîte noire». Ceux-ci doivent faire l'objet d'une attention particulière. Dans de telles circonstances, d'autres mesures d'explicabilité (par exemple la traçabilité, l'auditabilité et la communication transparente concernant les capacités du système) pourraient être requises, pour autant que le système dans son ensemble respecte les droits fondamentaux. La mesure dans laquelle l'explicabilité est nécessaire dépend fortement du contexte et de la gravité des conséquences si ce résultat est erroné ou imprécis d'une autre manière.[33]

2.3 Tensions entre ces principes

(54) Des tensions pourraient survenir entre les principes susmentionnés, pour lesquelles il n'existe pas de solution unique. En vertu de l'engagement fondamental de l'UE envers l'engagement démocratique, le droit à une procédure régulière et la participation politique ouverte, des méthodes de délibération responsable devraient être établies pour faire face à ces tensions. Par exemple, dans divers domaines d'application, *le principe de la prévention de toute atteinte* et *le principe de l'autonomie humaine* peuvent entrer en conflit. Ainsi, l'utilisation de systèmes d'IA aux fins d'une «police prédictive» pourrait contribuer à réduire la criminalité, mais d'une manière impliquant des activités de surveillance qui portent atteinte à la liberté individuelle et à la vie privée. En outre, la somme des avantages liés aux systèmes d'IA doit être sensiblement supérieure aux risques individuels prévisibles. Si ces principes fournissent clairement des orientations destinées à trouver des solutions, ils n'en demeurent pas moins des prescriptions éthiques abstraites. On ne peut attendre des professionnels de l'IA qu'ils trouvent la solution adaptée sur la base des principes ci-dessus. Il leur faut toutefois aborder les dilemmes et arbitrages éthiques selon une réflexion raisonnée et fondée sur des éléments probants, plutôt que sur la base de l'intuition ou d'un jugement aléatoire. Il pourrait toutefois exister des situations dans lesquelles aucun arbitrage acceptable du point de vue éthique ne peut être déterminé. Certains droits fondamentaux et principes connexes sont absolus et ne peuvent dépendre d'un exercice de mise en balance (par exemple, la dignité humaine).

Orientations essentielles dérivées du chapitre I:

[31] Cette exigence est liée au principe de la proportionnalité (reflété par la maxime selon laquelle «on ne tue pas une mouche avec un bazooka»). Les mesures prises pour parvenir à une fin (par exemple, l'extraction de données en vue d'optimiser l'IA) devraient être limitées au strict nécessaire. Cela implique également que lorsque plusieurs mesures sont en concurrence pour la réalisation d'un même but, la préférence devrait être accordée à celle qui est la moins défavorable aux droits fondamentaux et aux normes éthiques (par exemple, les développeurs d'IA devraient toujours accorder la préférence à des données du secteur public par rapport aux données à caractère personnel). Il convient également de faire référence à la proportionnalité entre l'utilisateur et le prestataire du déploiement, en tenant compte des droits des entreprises (y compris de propriété intellectuelle et de confidentialité), d'une part, et des droits de l'utilisateur, d'autre part.

[32] Notamment en invoquant leur droit d'association et d'adhérer à un syndicat dans un environnement de travail, comme le prévoit l'article 12 de la charte des droits fondamentaux de l'Union européenne.

[33] Par exemple, les préoccupations éthiques résultant de recommandations d'achat imprécises générées par un système d'IA ne pourraient être que limitées, contrairement à celles résultant de systèmes d'IA évaluant si un individu reconnu coupable d'une infraction pénale devrait être mis en liberté conditionnelle.

- ✓ Mettre au point, déployer et utiliser des systèmes d'IA en respectant les principes éthiques suivants: *respect de l'autonomie humaine, prévention de toute atteinte, équité et explicabilité.* Reconnaître et résoudre les tensions potentielles entre ces principes.
- ✓ Accorder une attention particulière aux situations concernant des groupes plus vulnérables tels que les enfants, les personnes handicapées et d'autres groupes historiquement défavorisés, exposés au risque d'exclusion, et/ou aux situations caractérisées par des asymétries de pouvoir ou d'information, par exemple entre les employeurs et les travailleurs, ou entre les entreprises et les consommateurs.[34]
- ✓ Reconnaître et être conscient que certaines applications d'IA sont certes susceptibles d'apporter des avantages considérables aux individus et à la société, mais qu'elles peuvent également avoir des incidences négatives, y compris des incidences pouvant s'avérer difficiles à anticiper, reconnaître ou mesurer (par exemple, en matière de démocratie, d'état de droit et de justice distributive, ou sur l'esprit humain lui-même). Adopter des mesures appropriées pour atténuer ces risques le cas échéant, de manière proportionnée à l'ampleur du risque.

II. Chapitre II: Parvenir à une IA digne de confiance

(55) Ce chapitre fournit des orientations relatives à la mise en œuvre et à la réalisation d'une IA digne de confiance, au moyen d'une liste de sept exigences qui devraient être respectées, s'appuyant sur les principes énoncés au chapitre I. En outre, des méthodes tant techniques que non techniques actuellement disponibles sont présentées aux fins de l'application de ces exigences tout au long du cycle de vie du système d'IA.

1. Exigences d'une IA digne de confiance

(56) Pour parvenir à une IA digne de confiance, il faut que les principes énoncés au chapitre I soient traduits en exigences concrètes. Ces exigences s'appliquent aux différentes parties prenantes participant au cycle de vie des systèmes d'IA: développeurs, prestataires et utilisateurs finaux, ainsi que la société au sens large. Le terme «développeurs» désigne les personnes qui effectuent des recherches sur les systèmes d'IA, et qui conçoivent et/ou mettent au point ces systèmes. Le terme «prestataires» désigne les organismes publics ou privés qui utilisent des systèmes d'IA dans leurs processus opérationnels pour proposer des produits et services à des tiers. Les utilisateurs finaux sont les personnes qui interagissent directement ou indirectement avec le système d'IA. Enfin, la société au sens large englobe tous les autres acteurs qui sont directement ou indirectement concernés par les systèmes d'IA.

(57) Différentes catégories de parties prenantes ont différents rôles à jouer pour veiller au respect des exigences:

a. Les développeurs devraient mettre en œuvre et appliquer les exigences aux processus de conception et de mise au point;
b. Les prestataires devraient veiller à ce que les systèmes qu'ils utilisent et les produits et services qu'ils proposent respectent les exigences;
c. Les utilisateurs finaux et la société au sens large devraient être informés de ces exigences et être en mesure de demander qu'elles soient respectées.

(58) La liste des exigences ci-dessous n'est pas exhaustive.[35] Elle comprend des aspects systémiques, individuels et sociétaux:

1 Action humaine et contrôle humain

Comprend les droits fondamentaux, l'action humaine et le contrôle humain

[34] Voir articles 24 à 27 de la charte l'UE, portant sur les droits de l'enfant et des personnes âgées, l'intégration des personnes handicapées et le droit des travailleurs. Voir également l'article 38 portant sur la protection des consommateurs.

[35] Sans imposer de hiérarchie, nous présentons les principes ci-dessous de manière à refléter l'ordre d'apparition, dans la charte de l'Union, des principes et des droits auxquels ils se rapportent.

2 **Robustesse technique et sécurité**

Comprend la résilience aux attaques et la sécurité, les plans de secours et la sécurité générale, la précision, la fiabilité et la reproductibilité

3 **Respect de la vie privée et gouvernance des données**

Comprend le respect de la vie privée, la qualité et l'intégrité des données, et l'accès aux données

4 **Transparence**

Comprend la traçabilité, l'explicabilité et la communication

5 **Diversité, non-discrimination et équité**

Comprend l'absence de biais injustes, l'accessibilité et la conception universelle, et la participation des parties prenantes

6 **Bien-être sociétal et environnemental**

Comprend la durabilité et le respect de l'environnement, l'impact social, la société et la démocratie

7 **Responsabilité**

Comprend l'auditabilité, la réduction au minimum des incidences négatives et la communication à leur sujet, les arbitrages et les recours.

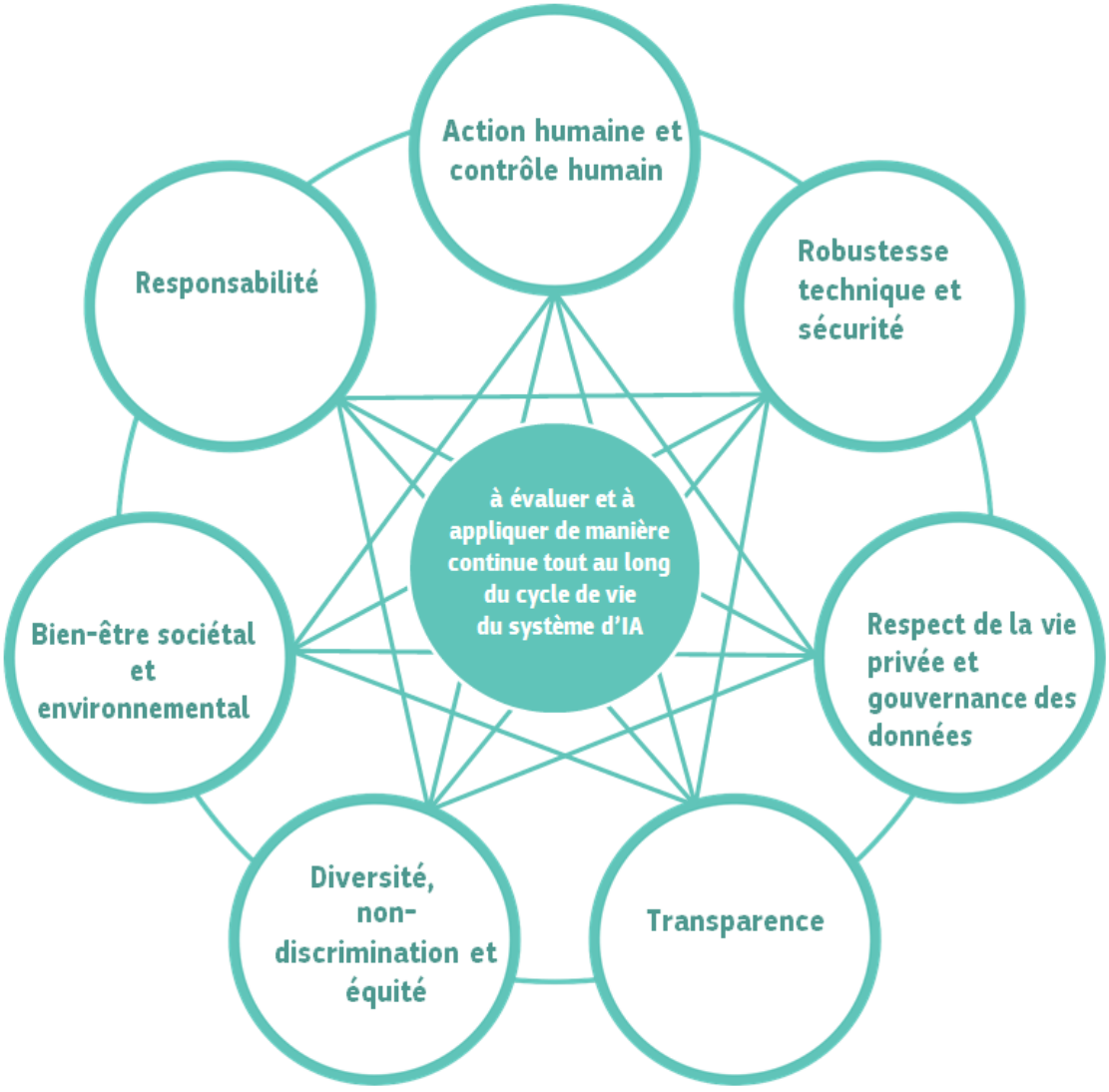


Figure 2: interrelation des sept exigences: elles revêtent toutes une importance égale, elles se soutiennent mutuellement et devraient être appliquées et évaluées tout au long du cycle de vie d'un système d'IA.

(59) Si toutes ces exigences revêtent une importance égale, le contexte et les tensions s'exerçant potentiellement entre elles devront être pris en compte lors de leur application à différents domaines et secteurs d'activité. La mise en œuvre de ces exigences devrait se faire tout au long du cycle de vie d'un système d'IA, et dépend de l'application spécifique. Si la plupart des exigences s'appliquent à l'ensemble des systèmes d'IA, une attention spécifique est accordée à celles qui ont des effets directs ou indirects sur les personnes. Par conséquent, pour certaines applications (par exemple, dans des contextes industriels), elles peuvent s'avérer moins pertinentes.

(60) Les exigences ci-dessus comprennent des éléments qui, dans certains cas, sont déjà reflétés dans la législation existante. Nous rappelons que – conformément à la première caractéristique d'une IA digne de confiance – les développeurs et prestataires de systèmes d'IA ont la responsabilité de faire en sorte qu'ils respectent leurs obligations juridiques, tant en ce qui concerne les règles applicables au niveau transversal que les règles spécifiques à un domaine donné.

(61) Dans les paragraphes qui suivent, chaque exigence fait l'objet d'un examen plus approfondi.

1. Action humaine et contrôle humain

(62) Les systèmes d'IA devraient soutenir l'autonomie et la prise de décisions humaines, conformément au principe du *respect de l'autonomie humaine*, en vertu duquel les systèmes d'IA devraient être à la fois les vecteurs d'une société démocratique, prospère et équitable en se mettant au service de l'utilisateur et favoriser les droits fondamentaux, ainsi que permettre un contrôle humain.

(63) ***Droits fondamentaux.*** À l'instar de nombreuses technologies, les systèmes d'IA peuvent autant favoriser qu'entraver les droits fondamentaux. Ils peuvent par exemple servir les particuliers en les aidant à suivre leurs données à caractère personnel ou en améliorant l'accès à l'éducation, en soutenant ainsi leur droit à l'éducation. Toutefois, étant donné la portée et la capacité des systèmes d'IA, ils peuvent également avoir une incidence négative sur les droits fondamentaux. Dans les situations où de tels risques existent, il convient d'entreprendre une analyse d'impact relative aux droits fondamentaux. Cette analyse devrait être menée préalablement à leur mise au point et comprendre une évaluation destinée à déterminer si ces risques peuvent être réduits ou justifiés comme nécessaires dans une société démocratique afin de respecter les droits et les libertés d'autrui. Il convient en outre de mettre sur pied des mécanismes permettant de recevoir des commentaires externes concernant les systèmes d'IA susceptibles de nuire aux droits fondamentaux.

(64) ***Action humaine.*** Les utilisateurs devraient être en mesure de prendre des décisions autonomes éclairées à l'égard des systèmes d'IA. Ils devraient recevoir les connaissances et les outils pour comprendre les systèmes d'IA et interagir avec eux dans une mesure satisfaisante et, autant que possible, être à même de procéder à une autoévaluation du système ou de le contester d'une manière appropriée. Les systèmes d'IA devraient aider les individus à prendre de meilleures décisions et à faire des choix plus éclairés en rapport avec leurs objectifs. Les systèmes d'IA peuvent parfois être déployés pour modeler et influencer le comportement humain à travers des mécanismes parfois difficiles à détecter, du fait qu'ils peuvent exploiter des processus subconscients, y compris différentes formes de manipulation déloyale, de tromperie, d'asservissement et de conditionnement, chacune étant susceptible de menacer l'autonomie individuelle. Le principe général d'autonomie des utilisateurs doit être au cœur des fonctionnalités du système. À cet égard, le droit des utilisateurs de ne pas faire l'objet d'une décision fondée exclusivement sur un traitement automatisé lorsque cela produit sur eux des effets juridiques ou d'autres effets d'importance comparable [36] revêt un caractère essentiel.

(65) ***Contrôle humain***. Le contrôle humain contribue à éviter qu'un système d'IA ne mette en péril l'autonomie humaine ou ne provoque d'autres effets néfastes. Le contrôle peut être assuré en recourant à des mécanismes de gouvernance tels que les approches dites «human-in-the-loop» (l'humain intervient dans le processus),

[36] Il peut être fait référence à l'article 22 du RGPD qui consacre déjà ce droit.

«human-on-the-loop» (l'humain supervise le processus) ou «human-in-command» (l'humain reste aux commandes). L'approche «human-in-the-loop» (HITL) désigne la capacité d'intervention humaine dans chaque cycle de décision du système, ce qui, dans de nombreux cas, n'est ni possible ni souhaitable. L'approche «human-on-the-loop» (HOTL) désigne une capacité d'intervention humaine dans le cycle de conception du système et la surveillance du fonctionnement du système. L'approche «human-in-command» (HIC) désigne une 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. Cette faculté peut comprendre la décision de ne pas utiliser un système d'IA dans une situation donnée, de définir des marges d'appréciation pour les interventions humaines lors de l'utilisation du système ou d'ignorer une décision prise par un système. Il convient en outre de veiller à ce que les autorités publiques soient en mesure d'exercer un contrôle conformément à leur mandat. Des mécanismes de contrôle peuvent être requis à des degrés divers pour soutenir d'autres mesures de sécurité et de contrôle, en fonction du domaine d'application du système d'IA et du risque potentiel. Toutes choses étant égales par ailleurs, moins un être humain peut exercer de contrôle sur un système d'IA, plus il faut approfondir les essais et renforcer la gouvernance.

2. Robustesse technique et sécurité

(66) Une caractéristique essentielle pour parvenir à une IA digne de confiance est la robustesse technique, qui est étroitement liée au *principe de la prévention de toute atteinte*. La robustesse technique passe par la mise au point de systèmes d'IA selon une approche de prévention des risques, et de telle manière que ces systèmes se comportent, de manière fiable, conformément aux attentes, tout en réduisant le plus possible les atteintes involontaires et inattendues, et en empêchant toute atteinte inacceptable. Cette exigence également s'appliquer aux modifications potentielles de l'environnement dans lequel ils sont exploités ou à la présence d'autres agents (humains et artificiels) pouvant avoir des interactions antagonistes avec le système. Il convient en outre de garantir l'intégrité physique et mentale des êtres humains.

(67) ***Résilience aux attaques et sécurité***. Les systèmes d'IA, à l'instar de tous les systèmes logiciels, devraient être protégés face aux vulnérabilités qui pourraient permettre à des adversaires de les exploiter (par exemple, piratage). Des attaques pourraient cibler les données (empoisonnement des données), le modèle (fuite de modèle) ou l'infrastructure sous-jacente, tant matérielle que logicielle. Lorsqu'un système d'IA fait l'objet d'une attaque, par exemple d'une attaque antagoniste, le comportement des données ainsi que du système peut être modifié, ce qui conduit le système à prendre des décisions différentes voire à s'arrêter. Les systèmes et les données peuvent également être corrompus en raison d'interventions malveillantes ou de l'exposition à des situations imprévues. Des procédures de sécurité insuffisantes peuvent également mener à des décisions erronées ou même entraîner des préjudices physiques. Pour que les systèmes d'IA soient considérés comme sûrs,[37] il convient de prendre en compte les applications involontaires potentielles de l'IA (par exemple, applications à double usage) et l'utilisation potentiellement abusive d'un système d'IA par des acteurs malveillants et de prendre des mesures pour les empêcher et les atténuer.[38]

(68) ***Plans de secours et sécurité générale***. Les systèmes d'IA devraient comporter des garanties permettant le déclenchement de plans de secours en cas de problèmes. Un système d'IA pourrait ainsi être amené à passer d'une procédure statistique à une procédure fondée sur des règles, ou à demander les instructions d'un

[37] Voir par exemple les considérations au point 2.7 du plan coordonné de l'Union européenne dans le domaine de l'intelligence artificielle.

[38] Pour assurer la sécurité des systèmes d'IA, il pourrait être indispensable de mettre en place un cercle vertueux en matière de recherche et de développement entre la compréhension des attaques, la mise au point de protections appropriées et l'amélioration des méthodes d'évaluation. Pour y parvenir, il convient de promouvoir une convergence entre la communauté de l'IA et la communauté de la sécurité. Il incombe en outre à l'ensemble des acteurs concernés de définir des normes communes de sûreté et de sécurité transfrontières et de mettre en place un environnement de confiance mutuelle, encourageant la collaboration internationale. Pour des mesures possibles, voir Malicious Use of AI (Avin S., Brundage M., et. al., 2018).

opérateur humain avant de poursuivre son action.[39] Il convient de s'assurer que le système fera ce qui est attendu de lui sans porter atteinte à des êtres vivants ou à l'environnement. Cela comprend la nécessité de réduire le plus possible les effets non désirés et les dysfonctionnements. En outre, des processus devraient être mis en place pour clarifier et évaluer les risques potentiels liés à l'utilisation de systèmes d'IA pour un éventail de domaines d'application. Le niveau des mesures de sécurité nécessaires dépend de l'ampleur du risque que présente un système d'IA, qui dépend en retour des capacités du système. Lorsqu'il apparaît prévisible que le processus de mise au point ou le système même présenteront des risques particulièrement élevés, il est essentiel de mettre au point et de tester de manière proactive des mesures de sécurité.

(69) ***Précision.*** La précision est fonction de la capacité d'un système d'IA à poser un jugement correct, par exemple en classant correctement des informations dans les bonnes catégories, ou de sa capacité à réaliser des prévisions, des recommandations ou des décisions correctes sur la base de données ou de modèles. Un processus de mise au point et d'évaluation explicite et bien formé peut, en plus d'apporter le soutien nécessaire, atténuer et corriger les risques imprévus découlant de prévisions inexactes. Lorsqu'il n'est pas possible d'éviter des prévisions inexactes occasionnelles, il est important que le système puisse indiquer le niveau de probabilité de ces erreurs. Un niveau élevé de précision est particulièrement essentiel dans les situations où le système d'IA a une incidence directe sur des vies humaines.

(70) ***Fiabilité et reproductibilité.*** Il est essentiel que les résultats des systèmes d'IA soient à la fois reproductibles et fiables. Un système d'IA fiable est un système qui fonctionne correctement avec toute une gamme de données d'entrée et dans un ensemble de situations. Ces caractéristiques sont nécessaires pour qu'un système d'IA puisse être soumis à un examen attentif et éviter tout préjudice involontaire. La reproductibilité est une indication de la mesure dans laquelle un système d'IA, dans le cadre d'essais répétés dans les mêmes conditions, produit un comportement similaire. Cela permet aux scientifiques et aux décideurs politiques de décrire avec précision ce que font les systèmes d'IA. Les fichiers de reproduction[40] peuvent faciliter le processus d'essai et de reproduction des comportements.

3. Respect de la vie privée et gouvernance des données

(71) Étroitement lié au *principe de la prévention de toute atteinte*, le respect de la vie privée est un droit fondamental particulièrement sensible aux incidences des systèmes d'IA. La prévention de toute atteinte au respect de la vie privée requiert également une gouvernance appropriée des données qui porte sur la qualité et l'intégrité des données utilisées, leur pertinence par rapport au domaine dans lequel les systèmes d'IA seront déployés, leurs protocoles d'accès et la capacité à traiter les données d'une manière qui protège la vie privée.

(72) ***Respect de la vie privée et protection des données***. Les systèmes d'IA doivent garantir le respect de la vie privée et la protection des données tout au long du cycle de vie d'un système.[41] Cela couvre les informations initialement fournies par l'utilisateur, ainsi que les informations générées au sujet de l'utilisateur au cours de ses interactions avec le système (par exemple, des résultats générés par le système d'IA pour des utilisateurs spécifiques, ou la manière dont les utilisateurs ont répondu à des recommandations spécifiques). La numérisation des comportements humains peut permettre aux systèmes d'IA de déduire non seulement les préférences d'une personne, mais aussi son orientation sexuelle, son âge, son sexe, ses convictions religieuses ou ses opinions politiques. Pour que les citoyens aient confiance dans le processus de collecte des données, ils doivent avoir la garantie que les données recueillies les concernant ne seront pas utilisées à leur encontre à des fins discriminatoires, de manière illicite ou injuste.

---

[39] Il convient également d'envisager des scénarios dans lesquels une intervention humaine ne serait pas immédiatement possible.

[40] Il s'agit de fichiers qui reproduiront chaque étape du processus de mise au point du système d'IA, du stade de la recherche et de la collecte initiale des données jusqu'au stade des résultats.

[41] Il peut être fait référence à la législation existante en matière de respect de la vie privée, telle que le RGPD ou le futur règlement «vie privée et communications électroniques».

(73) ***Qualité et intégrité des données***. La qualité des ensembles de données utilisés est essentielle au bon fonctionnement des systèmes d'IA. La collecte de données peut être entachée de biais d'ordre social, d'imprécisions, de fautes et d'erreurs. Il faut tenir compte de cet élément avant d'utiliser un ensemble de données pour entraîner un système d'IA. Par ailleurs, l'intégrité des données doit être assurée. Alimenter un système d'IA avec des données malveillantes peut modifier son comportement, notamment avec les systèmes d'autoapprentissage. Les processus et ensembles de données utilisés doivent être testés et documentés à chaque étape (planification, entraînement, essais et déploiement). Ce principe devrait s'appliquer également aux systèmes d'IA qui n'ont pas été développés en interne mais qui ont été acquis à l'extérieur.

(74) ***Accès aux données****.* Dans toute organisation traitant les données relatives à des personnes (qu'il s'agisse ou non d'utilisateurs du système), des protocoles de données régissant l'accès aux données devraient être mis en place. Ces protocoles devraient indiquer qui peut avoir accès aux données et dans quelles circonstances. Seul le personnel dûment qualifié ayant les compétences nécessaires et justifiant du besoin d'accéder à des données à caractère personnel devrait y être autorisé.

4. Transparence

(75) Cette exigence est étroitement liée au *principe de l'explicabilité* et comprend la transparence des éléments pertinents d'un système d'IA: les données, le système et les modèles économiques.

(76) ***Traçabilité****.* Les ensembles de données et les processus permettant au système d'IA de rendre une décision, y compris les processus de collecte et d'étiquetage de données, ainsi que les algorithmes utilisés, devraient être documentés selon les normes les plus strictes afin de permettre la traçabilité ainsi qu'une amélioration de la transparence. Ce principe s'applique également aux décisions rendues par le système d'IA. Cela permet de déterminer les raisons pour lesquelles une décision d'IA était erronée ce qui, en retour, pourrait contribuer à éviter de futures erreurs. La traçabilité facilite donc l'auditabilité et l'explicabilité.

(77) ***Explicabilité.*** L'explicabilité concerne la capacité d'expliquer à la fois les processus techniques d'un système d'IA et les décisions humaines qui s'y rapportent (par exemple, domaines d'application d'un système d'IA). L'explicabilité technique suppose que les décisions prises par un système d'IA puissent être comprises et retracées par des êtres humains. Par ailleurs, des arbitrages peuvent s'avérer nécessaires entre le renforcement de l'explicabilité d'un système (qui pourrait réduire sa précision) et l'amélioration de sa précision (au détriment de l'explicabilité). Dès qu'un système d'IA a une incidence importante sur la vie des personnes, il devrait être possible d'exiger une explication appropriée du processus de décision du système d'IA. Ces explications devraient être présentées en temps opportun et adaptées à l'expertise de la partie prenante concernée (par exemple, non-spécialiste, autorité de réglementation ou chercheur). Des explications devraient également être fournies sur la mesure dans laquelle un système d'IA influence et façonne le processus de prise de décisions organisationnel, les choix opérés dans la conception du système, et la justification de son déploiement (de manière à assurer la transparence du modèle économique).

(78) ***Communication.*** Les systèmes d'IA ne devraient pas se présenter comme des êtres humains auprès des utilisateurs; lorsqu'ils interagissent avec un système d'IA, les êtres humains ont le droit d'en être informés. Cet aspect implique que les systèmes d'IA doivent être identifiables en tant que tels. Qui plus est, la possibilité de s'opposer à cette interaction au profit d'une interaction humaine devrait être proposée le cas échéant afin de garantir le respect des droits fondamentaux. Outre cet aspect, il convient de communiquer aux professionnels de l'IA ou aux utilisateurs finaux des informations appropriées sur les capacités et les limites du système d'IA, selon des modalités adaptées au contexte d'utilisation concerné. Ces informations pourraient comprendre le degré de précision du système d'IA, ainsi que ses limites.

5. Diversité, non-discrimination et équité

(79) Pour parvenir à une IA digne de confiance, il est nécessaire de favoriser l'inclusion et la diversité tout au long du cycle de vie du système d'IA. Outre la prise en compte et la participation de l'ensemble des parties prenantes concernées tout au long du processus, cela implique également de veiller à l'égalité d'accès au moyen de processus conçus de manière inclusive, ainsi qu'à l'égalité de traitement. Cette exigence est étroitement liée au *principe de l'équité*.

(80) ***Absence de biais injustes***. Les ensembles de données utilisés par les systèmes d'IA (tant pour leur entraînement que pour leur exploitation) peuvent être biaisés par des partis pris historiques accidentels, des omissions et des modèles de gouvernance défectueux. La persistance de ces biais pourrait être source de discrimination et de préjudice (in)directs[42] involontaires à l'encontre de certains groupes de personnes, aggravant potentiellement le préjudice et la marginalisation. Des préjudices peuvent également résulter de l'exploitation intentionnelle de préjugés (des consommateurs) ou d'une concurrence déloyale, comme l'homogénéisation des prix par le biais d'une collusion ou l'opacité d'un marché.[43] Dans la mesure du possible, les biais détectables et discriminatoires devraient être supprimés lors de la phase de collecte. La manière dont les systèmes d'IA sont mis au point (par exemple la programmation des algorithmes) peut également être entachée de biais. On peut contrer cette tendance en mettant en place des procédures de contrôle pour analyser de manière claire et transparente la finalité, les contraintes, les exigences et les décisions du système. En outre, le recrutement de personnes issues de contextes, de cultures et de disciplines différents peut garantir la diversité des opinions et devrait être encouragé.

(81) ***Accessibilité et conception universelle***. Dans le contexte des relations d'entreprise à consommateur, notamment, les systèmes devraient être centrés sur l'utilisateur et conçus de manière à permettre à toute personne d'utiliser des produits ou services d'IA, quels que soient son âge, son sexe, ses capacités ou ses caractéristiques. L'accessibilité de cette technologie aux personnes atteintes de handicaps, qui sont présentes dans tous les segments de la société, revêt une importance particulière. Les systèmes d'IA ne devraient pas adopter une approche uniforme et devraient envisager des principes de conception universelle[44] répondant aux besoins du plus large éventail possible d'utilisateurs, en suivant des normes d'accessibilité pertinentes.[45] Ce principe permettra un accès équitable et la participation active de chacun aux activités humaines informatisées existantes et émergentes, ainsi qu'aux technologies d'assistance.[46]

(82) ***Participation des parties prenantes***. Pour mettre au point des systèmes d'IA dignes de confiance, il est souhaitable de consulter les parties prenantes sur lesquelles le système est susceptible d'avoir des effets directs ou indirects tout au long de son cycle de vie. Il est bénéfique de solliciter régulièrement des commentaires, même après le déploiement, et de mettre en place des mécanismes à plus long terme de participation des parties prenantes, en veillant par exemple à l'information, la consultation et la participation des travailleurs à travers tout le processus de mise en œuvre de systèmes d'IA au sein d'organisations.

6. Bien-être sociétal et environnemental

(83) Tout comme pour les *principes de l'équité* et *de la prévention de toute atteinte*, il convient de considérer également la société au sens large, les autres êtres sensibles et l'environnement comme des parties prenantes tout au long du cycle de vie de l'IA. La durabilité et la responsabilité écologique des systèmes d'IA devraient être encouragées, et il convient de promouvoir la recherche de solutions d'IA répondant à des préoccupations

---

42 Pour une définition des formes directes et indirectes de discrimination, voir par exemple l'article 2 de la directive 2000/78/CE du Conseil du 27 novembre 2000 portant création d'un cadre général en faveur de l'égalité de traitement en matière d'emploi et de travail. Voir également l'article 21 de la charte des droits fondamentaux de l'Union européenne.

43 Voir document de l'Agence des droits fondamentaux de l'Union européenne: «BigData: Discrimination in data-supported decision making (2018)» http://fra.europa.eu/en/publication/2018/big-data-discrimination.

44 L'article 42 de la directive relative aux marchés publics prévoit que les spécifications techniques doivent prendre en compte l'accessibilité et la conception pour tous.

45 Par exemple EN 301 549.

46 Cette exigence est liée à la Convention des Nations unies relative aux droits des personnes handicapées.

de portée mondiale, par exemple les objectifs de développement durable. Idéalement, tous les êtres humains devraient bénéficier de l'IA, y compris les générations futures.

(84) ***IA durable et respectueuse de l'environnement***. Les systèmes d'IA promettent de contribuer à répondre à certaines des plus vives préoccupations de la société; il faut cependant veiller à ce que les réponses apportées soient aussi respectueuses de l'environnement que possible. Il convient, à cet égard, d'évaluer le processus de mise au point, de déploiement et d'utilisation du système, ainsi que toute sa chaîne d'approvisionnement, par exemple au moyen d'un examen critique de l'utilisation des ressources et de la consommation d'énergie au cours de l'entraînement, en réalisant les choix les moins préjudiciables. Il convient d'encourager les mesures permettant de garantir que l'ensemble de la chaîne d'approvisionnement du système d'IA respecte l'environnement.

(85) ***Incidences sociales***. L'omniprésence des systèmes d'IA sociaux[47] dans tous les domaines de notre vie (qu'il s'agisse de l'enseignement, du travail, des soins ou des loisirs) peut altérer notre conception de l'action sociale ou avoir une incidence sur nos relations et nos liens sociaux. Si les systèmes d'IA peuvent être utilisés pour renforcer les compétences sociales[48], ils peuvent également contribuer à leur détérioration. Cela pourrait également nuire au bien-être physique ou mental des personnes. Les effets de ces systèmes doivent par conséquent faire l'objet d'un contrôle et d'un examen minutieux.

(86) ***Société et démocratie.*** En plus d'évaluer l'incidence de la mise au point, du déploiement et de l'utilisation d'un système d'IA sur les individus, il convient également d'évaluer cette incidence d'un point de vue sociétal, en tenant compte de son effet sur les institutions, la démocratie et la société au sens large. L'utilisation des systèmes d'IA devrait faire l'objet d'une attention particulière dans les situations mettant en jeu le processus démocratique, non seulement la prise de décisions politiques, mais aussi les contextes électoraux.

7. Responsabilité

(87) L'exigence de la responsabilité complète les exigences susmentionnées, étroitement liées au *principe de l'équité*. Elle requiert la mise en place de mécanismes permettant de garantir l'autonomie et la responsabilité à l'égard des systèmes d'IA et de leurs résultats, tant avant qu'après leur mise en œuvre.

(88) ***Auditabilité.*** L'auditabilité implique la possibilité d'évaluer les algorithmes, les données et les processus de conception. Elle n'implique pas nécessairement que les informations sur les modèles économiques et la propriété intellectuelle en lien avec le système d'IA doivent toujours être librement accessibles. L'évaluation par des auditeurs internes et externes, ainsi que la disponibilité des rapports de ces évaluations, peuvent contribuer à la fiabilité de la technologie. Pour les applications mettant en jeu les droits fondamentaux, notamment les applications critiques pour la sécurité, les systèmes d'IA devraient pouvoir faire l'objet d'audits indépendants.

(89) ***Réduction au minimum et documentation des incidences négatives***. Il convient de garantir la capacité aussi bien de documenter les actions ou décisions contribuant à un certain résultat du système que de répondre aux conséquences d'un tel résultat. Il est particulièrement important pour les personnes touchées directement ou indirectement que les effets négatifs potentiels des systèmes d'IA soient répertoriés, analysés, documentés et réduits le plus possible. Il convient d'assurer un niveau de protection approprié aux lanceurs d'alertes, aux ONG, aux syndicats ou à d'autres entités lorsqu'ils font état de préoccupations légitimes au sujet d'un système

[47] Sont visés ici les systèmes d'IA qui communiquent et interagissent avec les êtres humains en simulant un comportement social dans les interactions entre robots et humains (IA embarquée) ou comme avatars dans la réalité virtuelle. Ce faisant, ces systèmes ont le potentiel de modifier nos pratiques socioculturelles et le tissu de notre vie sociale.

[48] Voir par exemple le projet financé par l'UE en vue de la mise au point d'un logiciel fondé sur l'IA permettant à des robots d'interagir plus efficacement avec des enfants autistes lors de sessions thérapeutiques dirigées par des êtres humains, contribuant à améliorer leurs compétences sociales et de communication: http://ec.europa.eu/research/infocentre/article_en.cfm?id=/research/headlines/news/article_19_03_12_en.html?infocentre&item=Infocentre&artid=49968.

fondé sur l'IA. Le recours aux analyses d'impact (par exemple, le «red teaming» ou certaines formes d'analyse d'impact algorithmique), tant avant que pendant la mise au point, le déploiement et l'utilisation de systèmes d'IA, peut contribuer à réduire le plus possible les effets négatifs. Ces analyses doivent être proportionnées au risque associé aux systèmes d'IA.

(90) ***Arbitrages.*** Lors de la mise en œuvre des exigences ci-dessus, des tensions pourraient survenir entre elles, ce qui pourrait rendre inévitables certains arbitrages. Ces arbitrages devraient être effectués avec raison et méthode, conformément à l'état actuel de la technique. Cela implique qu'il convient de recenser les intérêts et valeurs pertinents concernés par le système d'IA et que, en cas de conflit, les arbitrages entre eux devraient être explicitement reconnus et évalués du point de vue du risque qu'ils posent pour les principes éthiques, y compris les droits fondamentaux. 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. Toute décision concernant un arbitrage à faire devrait être raisonnée et correctement documentée. La personne chargée de prendre la décision doit être tenue responsable de la manière dont l'arbitrage pertinent est effectué et devrait en permanence reconsidérer le caractère approprié de la décision résultante, pour veiller à ce que les modifications nécessaires soient apportées au système en cas de besoin. [49]

(91) ***Recours***. Lorsqu'une incidence négative injuste se produit, il convient de prévoir des mécanismes accessibles assurant une voie de recours adéquate.[50] Savoir qu'un recours est possible lorsque les choses se passent mal est essentiel pour garantir la confiance. Il convient d'accorder une attention particulière aux personnes ou groupes vulnérables.

2. Méthodes techniques et non techniques pour parvenir à une IA digne de confiance

(92) Pour mettre en œuvre les exigences susmentionnées, des méthodes tant techniques que non techniques peuvent être appliquées. Ces méthodes englobent toutes les phases du cycle de vie d'un système d'IA. Il convient de procéder de manière continue à une évaluation des méthodes employées pour mettre en œuvre les exigences, ainsi qu'à la communication et à la justification[51] des modifications apportées aux processus de mise en œuvre. Étant donné que les systèmes d'IA évoluent et agissent de manière continue dans un environnement dynamique, la réalisation d'une IA digne de confiance est un processus continu, illustré à la figure 3 ci-dessous.

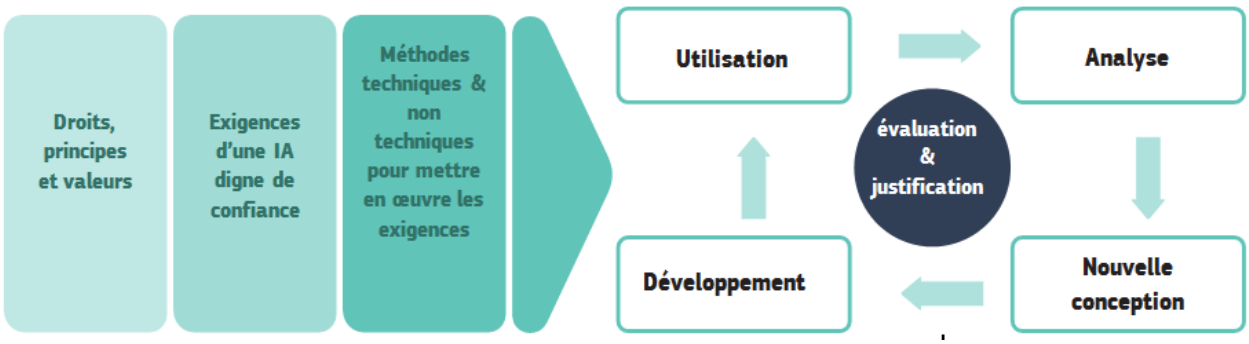


Figure 3: parvenir à une IA digne de confiance tout au long du cycle de vie du système

[49] Différents modèles de gouvernance peuvent contribuer à cet objectif. Par exemple, la présence d'un expert ou conseil éthique (et sectoriel) interne et/ou externe pourrait être utile pour mettre en évidence des domaines de conflit potentiel et proposer les manières les plus adaptées de résoudre ce conflit. Il est également utile de procéder à une consultation et à une discussion concrètes avec les parties prenantes, y compris celles susceptibles de subir les incidences négatives d'un système d'IA. Les universités européennes devraient jouer un rôle de premier plan dans la formation d'experts nécessaires dans le domaine de l'éthique.

[50] Voir également l'avis de l'Agence des droits fondamentaux de l'Union européenne sur l'amélioration de l'accès aux voies de recours dans les domaines des droits de l'homme et des entreprises au niveau de l'Union (2017), https://fra.europa.eu/en/opinion/2017/business-human-rights.

[51] Cela implique, par exemple, la justification des choix réalisés dans la conception, la mise au point et le déploiement du système pour intégrer les exigences susmentionnées.

(93) Les méthodes suivantes peuvent être considérées comme mutuellement complémentaires ou alternatives, car des exigences différentes – et des sensibilités différentes – peuvent justifier des méthodes de mise en œuvre différentes. Cet aperçu n'a pas vocation à être exhaustif ou obligatoire. Il vise plutôt à proposer une liste de méthodes possibles susceptibles de contribuer à la mise en œuvre d'une IA digne de confiance.

1. Méthodes techniques

(94) Cette section décrit les méthodes techniques pour garantir une IA digne de confiance qui puisse être incorporée aux phases de conception, de mise au point et d'utilisation d'un système d'IA. Le niveau de maturité des méthodes présentées ci-dessous varie.[52]

- *Architectures pour une IA digne de confiance*

(95) Les exigences d'une IA digne de confiance devraient être «traduites» en procédures et/ou en contraintes imposées aux procédures, qui devraient être ancrées dans l'architecture du système d'IA. Cela pourrait être accompli au moyen d'un ensemble de règles dites «listes blanches» (comportements ou états) que le système devrait toujours suivre, de restrictions dites «listes noires» relatives aux comportements ou états que le système ne devrait jamais transgresser, et de combinaisons des deux ou de garanties démontrables plus complexes concernant le comportement du système. Un processus distinct pourrait servir à contrôler le respect de ces restrictions par le système, pendant son fonctionnement.

(96) Les systèmes d'IA dotés de capacités d'apprentissage et capables d'adapter leur comportement de façon dynamique peuvent être perçus comme des systèmes non déterministes susceptibles d'afficher un comportement inattendu. Ces systèmes sont souvent considérés à travers le prisme théorique d'un cycle «sense-plan-act» (détection-planification-action). Pour que cette architecture soit adaptée à une IA digne de confiance, il convient d'intégrer les exigences à chacune des trois étapes du cycle: i) à l'étape de la «détection», le système devrait être mis au point de telle sorte qu'il reconnaisse l'ensemble des éléments présents dans l'environnement qui sont nécessaires en vue de garantir l'adhésion à ces exigences; ii) à l'étape de la «planification», le système devrait uniquement envisager des plans adhérant aux exigences, et; iii) à l'étape de l'action, les actions du système devraient être limitées aux comportements correspondant à ces exigences.

(97) L'architecture telle qu'illustrée ci-dessus est générique et ne constitue qu'une description imparfaite pour la plupart des systèmes d'IA. Elle présente toutefois des points d'ancrage pour les contraintes et les règles qui devraient être reflétées dans des modules spécifiques aux fins de la mise au point d'un système digne de confiance et perçu comme tel.

- *Éthique et état de droit dès la conception (X dès la conception)*

(98) Les méthodes destinées à garantir les valeurs dès la conception établissent des liens précis et explicites entre les principes abstraits auxquels le système doit adhérer et les décisions spécifiques de mise en œuvre. L'idée selon laquelle la conformité aux normes peut être incorporée dans la conception du système d'IA est essentielle pour cette méthode. Les entreprises ont la responsabilité de recenser les effets de leurs systèmes d'IA dès le tout début, ainsi que les normes auxquelles ces systèmes doivent se conformer pour éviter les répercussions négatives. Différentes approches «dès la conception» sont déjà largement utilisées, comme le *respect de la vie privée dès la conception* et la *sécurité dès la conception*. Comme indiqué plus haut, pour susciter la confiance, les processus, données et résultats de l'IA doivent être sûrs, et le système devrait être conçu de manière à être robuste face aux données et attaques antagonistes. Un mécanisme d'arrêt, assurant la sûreté après défaillance, devrait être mis en œuvre et le redémarrage à la suite d'un arrêt forcé (par

[52] Alors que certaines de ces méthodes sont déjà disponibles aujourd'hui, d'autres doivent encore faire l'objet de davantage de recherches. Le GEHN IA s'appuiera également sur les domaines devant faire l'objet de recherches supplémentaires aux fins de sa deuxième contribution, à savoir les recommandations en matière de politique et d'investissement.

exemple, une attaque) devrait être rendu possible.

- *Méthodes d'explication*

(99) Pour qu'un système soit digne de confiance, nous devons être en mesure de comprendre pourquoi il s'est comporté d'une certaine manière et pourquoi il a fourni une interprétation donnée. Un domaine de recherche à part entière, l'IA explicable (*Explainable AI* – XAI), essaie de répondre à cette question pour mieux comprendre les mécanismes sous-jacents du système et trouver des solutions. Il s'agit encore à ce jour d'un défi à relever pour les systèmes d'IA fondés sur des réseaux neuronaux. Les processus d'entraînement avec réseaux neuronaux peuvent déboucher sur des paramètres de réseau réglés sur des valeurs numériques difficiles à mettre en lien avec des résultats. En outre, de légères modifications apportées aux valeurs des données peuvent parfois modifier de façon spectaculaire l'interprétation, menant par exemple le système à confondre un bus scolaire avec une autruche. Cette vulnérabilité peut également être exploitée au cours d'attaques contre le système. Les méthodes recourant à l'IA explicable sont non seulement essentielles pour expliquer le comportement du système aux utilisateurs, mais également pour déployer une technologie fiable.

- *Essais et validations*

(100) Étant donné la nature non déterministe des systèmes d'IA et la mesure dans laquelle ils sont spécifiques à leurs contextes, les essais traditionnels ne sont pas suffisants. Les défaillances des concepts et des représentations utilisés par le système ne sont susceptibles de se manifester que lorsqu'un programme est appliqué à des données suffisamment réalistes. Par conséquent, pour vérifier et valider le traitement des données, le modèle sous-jacent doit faire l'objet d'un contrôle attentif tant au cours de l'entraînement que du déploiement en ce qui concerne sa stabilité, sa robustesse et son fonctionnement dans des limites bien comprises et prévisibles. Il convient de veiller à ce que le résultat du processus de planification corresponde aux données d'entrée, et que les décisions soient prises d'une façon qui permette la validation du processus sous-jacent.

(101) Les essais et la validation du système devraient avoir lieu le plus tôt possible, pour veiller à ce que le système se comporte de la manière prévue tout au long de son cycle de vie et notamment après son déploiement. Ils devraient porter sur l'ensemble des éléments d'un système d'IA, y compris les données, les modèles pré-entraînés, les environnements et le comportement du système dans son ensemble, et être conçus et mis en œuvre par un groupe de personnes le plus divers possible. Plusieurs indicateurs devraient être définis pour couvrir les catégories faisant l'objet d'essais dans différentes perspectives. Des essais antagonistes réalisés par des «équipes rouges» de confiance et diversifiées, tentant délibérément de «pénétrer» le système à la recherche de failles, ainsi que des «primes au bogue» incitant les utilisateurs externes à détecter et signaler de manière responsable les erreurs et les faiblesses du système, peuvent être envisagés. Enfin, il convient de veiller à ce que les résultats ou actions correspondent aux résultats des processus préalables, en les comparant aux règles définies antérieurement pour faire en sorte que celles-ci ne soient pas violées.

- *Qualité des indicateurs de service*

(102) Un niveau de qualité approprié des indicateurs de service peut être défini pour les systèmes d'IA, afin de faire en sorte qu'un point de comparaison existe pour déterminer s'ils ont été testés et mis au point en tenant compte de la sécurité et de la sûreté. Ces indicateurs pourraient comprendre des mesures pour évaluer les essais et l'entraînement des algorithmes ainsi que des indicateurs logiciels traditionnels de la fonctionnalité, de la performance, de la facilité d'utilisation, de la fiabilité, de la sécurité et de la maintenabilité.

2. Méthodes non techniques

(103) Cette section décrit un éventail de méthodes non techniques susceptibles de jouer un rôle important pour obtenir et préserver une IA digne de confiance. Ces méthodes devraient également faire l'objet d'une **évaluation constante**.

- *Réglementation*

(104) Comme indiqué plus haut, une réglementation est déjà en place pour soutenir la fiabilité de l'IA, comme la législation relative à la sécurité des produits et les cadres applicables en matière de responsabilité. Dans la mesure où nous considérons qu'il pourrait être nécessaire de réviser ou d'adapter la réglementation, ou d'en adopter de nouvelles – pour servir tant de garanties que de catalyseurs – cet aspect sera abordé dans le cadre de notre deuxième contribution, qui consistera à formuler des recommandations en matière de politique et d'investissement.

- *Codes de conduite*

(105) Les organisations et les parties prenantes peuvent adopter les lignes directrices et adapter leurs chartes de responsabilité de l'organisation, leurs indicateurs de performances clés («KPI»), leurs codes de conduite ou règles internes pour y ajouter l'objectif de parvenir à une IA digne de confiance. Une organisation travaillant à la mise au point d'un système d'IA peut, de manière plus générale, documenter ses intentions, ainsi que les appuyer sur des normes relatives à certaines valeurs souhaitables, telles que les droits fondamentaux, la transparence et la prévention des préjudices.

- *Normalisation*

(106) Les normes, en matière par exemple de conception, de fabrication et de pratiques commerciales, peuvent fonctionner en tant que système de gestion de la qualité pour les utilisateurs de l'IA, consommateurs, organisations, instituts de recherche et pouvoirs publics, en offrant la possibilité de reconnaître et d'encourager un comportement éthique par leurs décisions d'achats. Outre les normes conventionnelles, il existe des approches de corégulation: systèmes d'agrément, codes de déontologie professionnels ou normes relatives à une conception conforme aux droits fondamentaux. On compte notamment parmi les exemples actuels les normes ISO ou les séries de normes IEEE P7000. Toutefois, un futur label «IA digne de confiance» pourrait être approprié, qui confirmerait par référence à des normes techniques spécifiques que le système est conforme, par exemple, en matière de sûreté, de robustesse technique et d'explicabilité.

- *Certification*

(107) Si on ne peut pas s'attendre à ce que tout le monde soit capable de comprendre totalement le fonctionnement et les effets des systèmes d'IA, on pourrait imaginer des organisations qui soient en mesure d'attester auprès du grand public qu'un système d'IA est transparent, responsable et juste.[53] Ces certifications appliqueraient des normes définies pour différents domaines d'application et techniques d'IA, dûment alignées sur les normes industrielles et sociétales des différents contextes. Une certification ne pourra toutefois jamais remplacer la responsabilité. Elle devrait par conséquent s'accompagner de cadres de responsabilité, y compris de clauses de non-responsabilité ainsi que de mécanismes de révision et de correction[54].

- *La responsabilité au moyen de cadres de gouvernance*

(108) Les organisations devraient définir des cadres de gouvernance, tant internes qu'externes, garantissant la responsabilité à l'égard des dimensions éthiques des décisions associées à la mise au point, au déploiement et à l'utilisation de l'IA. Cela pourrait, par exemple, comprendre la nomination d'une personne chargée des questions d'éthique en lien avec l'IA, ou d'un groupe ou conseil interne/externe traitant de ces questions. Ces personnes, groupes ou conseils pourraient être chargés d'assurer une supervision et de formuler des conseils. Comme indiqué plus haut, des spécifications et/ou organismes de certification peuvent jouer un rôle à cet

[53] Comme le préconise par exemple l'IEEE dans son initiative relative à une conception alignée sur le plan éthique: https://standards.ieee.org/industry-connections/ec/autonomous-systems.html.

[54] Pour plus d'informations sur les limites de la certification, voir: https://ainowinstitute.org/AI_Now_2018_Report.pdf.

effet. Des canaux de communication devraient être mis en place avec des groupes de supervision issus des secteurs public et/ou privé, pour partager les bonnes pratiques, discuter des dilemmes ou signaler des problèmes émergents liés à des préoccupations éthiques. De tels mécanismes peuvent compléter mais pas remplacer le contrôle juridique (par exemple, via la nomination d'un responsable de la protection des données ou des mesures équivalentes, qui sont juridiquement requises par la législation sur la protection des données).

- *Éducation et sensibilisation pour encourager un état d'esprit éthique*

(109) Une IA digne de confiance encourage la participation informée de toutes les parties prenantes. La communication, l'éducation et la formation jouent un rôle important, tant pour veiller à la diffusion des connaissances sur les incidences potentielles des systèmes d'IA que pour informer la population qu'elle peut participer à l'orientation du développement de la société. Cela concerne l'ensemble des parties prenantes, par exemple celles qui sont impliquées dans la création de produits (concepteurs et développeurs), les utilisateurs (entreprises ou personnes) et d'autres groupes concernés (ceux qui n'achèteront ou n'utiliseront pas nécessairement un système d'IA mais au nom de qui des décisions sont prises par un système d'IA, et la société au sens large). L'acquisition de connaissances de base en matière d'IA devrait être encouragée dans toute la société. Une condition préalable pour éduquer le public est de veiller à ce que les éthiciens possèdent les compétences et la formation requises dans ce domaine.

- *Participation des parties prenantes et dialogue social*

(110) Les avantages de l'IA sont nombreux, et l'Europe doit veiller à ce qu'ils soient à la disposition de chacun. Cela nécessite une discussion ouverte et la participation des partenaires sociaux, des parties prenantes ainsi que du grand public. De nombreuses organisations s'appuient déjà sur des groupes de parties prenantes pour discuter de l'utilisation des systèmes d'IA et de l'analyse des données. Ces groupes sont composés de différents membres, tels que des experts juridiques, des experts techniques, des éthiciens, des représentants de consommateurs et des travailleurs. La recherche active d'une participation et d'un dialogue concernant l'utilisation et les incidences des systèmes d'IA contribue à l'évaluation des résultats et des approches, et peut s'avérer particulièrement utile dans les cas complexes.

- *Diversité et équipes de conception inclusives*

(111) La diversité et l'inclusion jouent un rôle essentiel dans la mise au point de systèmes d'IA destinés à être utilisés dans le monde réel. Alors que les systèmes d'IA réalisent davantage de tâches de manière autonome, il est essentiel que les équipes qui conçoivent, mettent au point, testent, entretiennent, déploient et/ou achètent ces systèmes reflètent la diversité des utilisateurs et de la société en général. Cela contribue à l'objectivité et à la prise en compte de différents points de vue, besoins et objectifs. Idéalement, les équipes doivent non seulement être diversifiées en ce qui concerne le genre, la culture et l'âge, mais également du point de vue du parcours professionnel et des compétences.

Orientations essentielles dérivées du chapitre II:

- ✓ Veiller à ce que l'ensemble du cycle de vie du système d'IA réponde aux exigences d'une IA digne de confiance: 1) action humaine et contrôle humain, 2) robustesse technique et sécurité, 3) respect de la vie privée et gouvernance des données, 4) transparence, 5) diversité, non-discrimination et équité, 6) bien-être sociétal et environnemental, et 7) responsabilité.
- ✓ Envisager des méthodes techniques et non techniques afin de garantir la mise en œuvre de ces exigences.
- ✓ Encourager la recherche et l'innovation en vue de contribuer à l'évaluation des systèmes d'IA et de soutenir la mise en œuvre des exigences; diffuser les résultats et adresser les questions au grand public, et veiller à ce qu'une formation dans le domaine l'éthique en matière d'IA soit systématiquement dispensée à la nouvelle génération d'experts.

- ✓ Fournir, clairement et de façon proactive, des informations aux parties prenantes sur les capacités et les limites des systèmes d'IA, afin de leur permettre de formuler des attentes réalistes, ainsi que sur la manière dont les exigences sont mises en œuvre. Faire preuve de transparence sur le fait qu'elles interagissent avec un système d'IA.
- ✓ Faciliter la traçabilité et l'auditabilité des systèmes d'IA, en particulier dans les contextes et situations critiques.
- ✓ Mobiliser les parties prenantes tout au long du cycle de vie des systèmes d'IA. Encourager la formation et l'éducation afin que toutes les parties prenantes soient renseignées sur l'IA digne de confiance et formées dans ce domaine.
- ✓ Avoir conscience qu'il peut exister des tensions fondamentales entre différents principes et exigences. Recenser, évaluer, documenter et communiquer de manière continue ces arbitrages et leurs solutions.

III. Chapitre III: évaluation d'une IA digne de confiance

(112) Sur la base des exigences essentielles du chapitre II, ce chapitre établit une **liste d'évaluation** non exhaustive **pour une IA digne de confiance** (version pilote) permettant de **concrétiser une IA digne de confiance.** Cette liste s'applique notamment aux systèmes d'IA qui interagissent directement avec les utilisateurs et est avant tout destinée aux développeurs et aux prestataires chargés du déploiement de systèmes d'IA (qu'ils aient été mis au point en interne ou obtenus auprès de tiers). Elle ne porte pas sur la concrétisation du premier élément caractéristique d'une IA digne de confiance (la licéité). Cette liste d'évaluation, dont le respect ne constitue pas une preuve de conformité à la législation, n'a pas vocation à fournir des orientations visant à garantir le respect de la législation applicable. Vu la spécificité des applications propres aux systèmes d'IA, cette liste d'évaluation devra être adaptée aux cas d'utilisation et contextes spécifiques dans lesquels le système fonctionne. En outre, ce chapitre propose une recommandation générale sur la manière de mettre en œuvre la liste d'évaluation pour une IA digne de confiance au moyen d'une structure de gouvernance englobant tant le niveau opérationnel que le niveau de l'encadrement.

(113) La liste d'évaluation et la structure de gouvernance seront élaborées en étroite collaboration avec les parties prenantes des secteurs public et privé. Ce processus sera mené en tant que processus «pilote» et permettra de recueillir de nombreuses réactions dans le cadre de deux processus parallèles:

a. un processus qualitatif garantissant la représentation, auquel un nombre limité d'entreprises, d'organisations et d'institutions (de différents secteurs et de différentes tailles) adhérera pour tester la liste d'évaluation et la structure de gouvernance dans la pratique et fournir un retour d'information détaillé;

b. un processus quantitatif auquel toutes les parties prenantes pourront adhérer pour tester la liste d'évaluation et formuler des commentaires au moyen d'une consultation ouverte.

(114) À la suite de la phase pilote, nous intégrerons les résultats de ces processus à la liste d'évaluation et préparerons une version révisée début 2020. L'objectif est d'obtenir un cadre pouvant être appliqué de manière transversale à l'ensemble des applications et donc de jeter les bases d'une IA digne de confiance dans tous les domaines. Une fois ces bases établies, un cadre sectoriel ou spécifique aux applications pourrait être défini.

Gouvernance

(115) Des entreprises, organisations et institutions pourraient s'intéresser à la manière de mettre en œuvre la liste d'évaluation pour une IA digne de confiance au sein de leur structure. Pour ce faire, elles pourraient inclure le processus d'évaluation à leurs mécanismes de gouvernance existants, ou mettre en œuvre de

nouveaux processus. Ce choix dépendra de la structure interne de l'organisation ainsi que de sa taille et des ressources dont elle dispose.

(116) Des recherches[55] indiquent que tout changement requiert nécessairement l'attention des cadres supérieurs. Elles indiquent aussi que mobiliser l'ensemble des parties prenantes d'une entreprise, organisation ou institution accroît l'acceptation et la pertinence de l'introduction d'un nouveau processus (technologique ou non)[56]. Nous recommandons par conséquent la mise en œuvre d'un processus comprenant la participation tant au niveau opérationnel qu'au niveau des cadres supérieurs.

Niveau	Rôles pertinents (en fonction de l'organisation)
Encadrement et organe supérieur	L'encadrement supérieur étudie et évalue la mise au point, le déploiement ou l'achat de l'IA, en tant qu'échelon supérieur pour l'évaluation de l'ensemble des innovations et utilisations de l'IA, lorsque des préoccupations majeures sont détectées. Il y associe les personnes concernées par l'éventuelle introduction de systèmes d'IA (par exemple, des travailleurs) et leurs représentants tout au long du processus via des procédures d'information, de consultation et de participation.
Service chargé des questions de conformité, de licéité et de responsabilité de l'organisation	Ce service contrôle l'utilisation de la liste d'évaluation et sa nécessaire évolution pour répondre aux changements technologiques ou réglementaires. Il met à jour les normes ou règles internes relatives aux systèmes d'IA et veille à ce que l'utilisation de ces systèmes respecte le cadre juridique et réglementaire en vigueur et les valeurs de l'organisation.
Service chargé de la mise au point des produits et services ou équivalent	Le service chargé de la mise au point des produits et services utilise la liste d'évaluation pour évaluer les produits et services fondés sur l'IA et assure la journalisation de tous les résultats. Ces résultats font l'objet de discussions au niveau de l'encadrement, qui approuve en fin de compte les applications nouvelles ou révisées fondées sur l'IA.
Assurance qualité	Le service d'assurance qualité (ou équivalent) contrôle les résultats de la liste d'évaluation et prend des mesures pour faire remonter le problème à un échelon supérieur lorsque le résultat n'est pas satisfaisant ou que des résultats imprévus sont détectés.
RH	Le service RH veille à ce que les développeurs de systèmes d'IA possèdent un bon éventail de compétences et présentent une diversité de profils. Il fait en sorte qu'un niveau approprié de formation soit dispensé au sein de l'organisation au sujet de l'IA digne de confiance.
Achats	Le service des achats veille à ce que la procédure d'achat de produits ou services dotés fondés sur l'IA prévoie un contrôle de leur fiabilité.
Activités quotidiennes	Les développeurs et gestionnaires de projets utilisent la liste d'évaluation dans leur travail quotidien et documentent les résultats et les conséquences de l'évaluation.

[55] https://www.mckinsey.com/business-functions/operations/our-insights/secrets-of-successful-change-implementation.

[56] Voir par exemple A. Bryson, E. Barth et H. Dale-Olsen, T*he Effects of Organisational change on worker well-being and the moderating role of trade unions*, *ILRReview*, 66(4), juillet 2013; Jirjahn, U. et Smith, S.C. (2006). *What Factors Lead Management to Support or Oppose Employee Participation—With and Without Works Councils? Hypotheses and Evidence from Germany's Industrial Relations*, 45(4), 650–680; Michie, J. et Sheehan, M. (2003). *Labour market deregulation, "flexibility" and innovation*, *Cambridge Journal of Economics*, 27(1), 123–143.

Utilisation de la liste d'évaluation pour une IA digne de confiance

(117) Lorsque la liste d'évaluation est utilisée en pratique, nous recommandons de prêter attention non seulement aux sujets de préoccupation, mais également aux questions auxquelles aucune réponse ne peut (facilement) être apportée. Le manque de diversité dans les aptitudes et les compétences de l'équipe chargée de la mise au point et des essais du système d'IA pourrait être un problème potentiel, et il pourrait par conséquent être nécessaire de mobiliser d'autres parties prenantes à l'intérieur ou à l'extérieur de l'organisation. Il est fortement recommandé de consigner tous les résultats d'un point de vue tant technique que managérial, en veillant à ce que la résolution du problème puisse être comprise à tous les niveaux de la structure de gouvernance.

(118) La liste d'évaluation a vocation à guider les professionnels de l'IA dans la mise au point, le déploiement et l'utilisation d'une IA digne de confiance. L'évaluation devrait être adaptée au cas d'utilisation spécifique de manière proportionnée. Au cours de la phase pilote, des domaines sensibles spécifiques pourraient être révélés et le besoin d'autres dispositions particulières dans ce genre de cas sera évalué à la prochaine étape. Si cette liste d'évaluation n'apporte pas de réponses concrètes aux questions soulevées, elle encourage la réflexion sur les démarches susceptibles de contribuer à la fiabilité des systèmes d'IA et sur les démarches potentielles à adopter à cet égard.

Lien avec la législation et les processus existants

(119) Il est également important pour les personnes participant à la mise au point, au déploiement et à l'utilisation de l'IA de reconnaître que différentes législations existantes, portant sur l'application de processus spécifiques et l'interdiction de résultats particuliers, pourraient se chevaucher et coïncider avec certaines des mesures figurant dans la liste d'évaluation. Par exemple, la législation sur la protection des données définit un ensemble d'exigences juridiques que les personnes mobilisées pour la collecte et le traitement de données à caractère personnel doivent appliquer. Pourtant, étant donné qu'une IA digne de confiance passe aussi par un traitement éthique des données, les procédures et règles internes visant à assurer la conformité avec la législation sur la protection des données pourraient également contribuer à faciliter le traitement éthique des données et peuvent donc compléter les processus juridiques existants. Cette liste d'évaluation, dont le respect *ne constitue pas* une preuve de conformité à la législation, n'a pourtant pas vocation à fournir des orientations visant à garantir le respect de la législation applicable. Elle vise plutôt à présenter un ensemble de questions spécifiques aux destinataires, pour veiller à ce que leur approche de la mise au point et du déploiement de l'IA soit orientée vers l'obtention d'une IA digne de confiance.

(120) De la même manière, de nombreux professionnels de l'IA disposent déjà d'outils d'évaluation et de processus de développement de logiciels pour veiller également à la conformité avec des normes non juridiques. L'évaluation ci-dessous ne devrait pas nécessairement être réalisée de manière isolée, mais peut être incorporée à de telles pratiques existantes.

LISTE D'EVALUATION POUR UNE IA DIGNE DE CONFIANCE (VERSION PILOTE)

1. Action humaine et contrôle humain

Droits fondamentaux:

✓ Dans les cas d'utilisation susceptibles d'entraîner des effets négatifs sur les droits fondamentaux,

avez-vous réalisé une analyse d'impact sur les droits fondamentaux? Avez-vous déterminé et documenté le recours potentiel à des arbitrages entre les différents principes et droits?

- ✓ Le système d'IA interagit-il avec la prise de décision par un utilisateur final humain (par exemple, en recommandant des mesures ou décisions à prendre, ou en présentant des choix possibles)?
  - Dans de tels cas, existe-t-il un risque que le système d'IA affecte l'autonomie humaine en interférant de manière involontaire avec le processus décisionnel de l'utilisateur final?
  - Estimez-vous qu'un système d'IA devrait communiquer aux utilisateurs qu'une décision, un contenu, un conseil ou un résultat découlent d'une décision algorithmique?
  - Lorsque le système d'IA comporte un robot ou système conversationnel, les utilisateurs humains sont-ils informés du fait qu'ils interagissent avec un agent virtuel?

Action humaine:

- ✓ Lorsque le système d'IA est intégré dans un processus de travail, avez-vous réfléchi à la répartition des tâches entre le système d'IA et les travailleurs humains pour permettre des interactions constructives ainsi qu'une supervision et un contrôle humains appropriés?
  - Le système d'IA renforce-t-il ou augmente-t-il les capacités humaines?
  - Avez-vous prévu des garanties pour empêcher toute confiance ou dépendance excessives envers le système d'IA dans les processus de travail?

Contrôle humain:

- ✓ Avez-vous réfléchi au niveau approprié de contrôle humain pour le système d'IA et le cas d'utilisation en question?
  - Pouvez-vous décrire le niveau de contrôle ou de participation humains, le cas échéant? Qui est «l'humain aux manettes» et à quel moment y a-t-il intervention humaine, ou avec quels outils?
  - Avez-vous mis en place des mécanismes et des mesures pour garantir un contrôle ou une supervision humains potentiels de cette nature, ou pour veiller à ce que les décisions soient prises sous la responsabilité globale d'êtres humains?
  - Avez-vous pris des mesures pour permettre la réalisation d'audits et résoudre des questions liées à la gouvernance de l'autonomie de l'IA?
- ✓ Dans le cas d'un système d'IA ou d'une utilisation capables d'autoapprentissage ou autonomes, avez-vous mis en place des mécanismes plus spécifiques de contrôle et de supervision?
  - Quel type de mécanismes de détection et de réponse avez-vous mis sur pied pour évaluer le risque que des problèmes surviennent?
  - Avez-vous veillé à la présence d'un «bouton d'arrêt» ou à l'existence d'une procédure pour suspendre, en cas de besoin, une opération en toute sécurité? Cette procédure suspend-elle le processus dans sa totalité, en partie, ou délègue-t-elle le contrôle à un être humain?

2. Robustesse technique et sécurité

Résilience aux attaques et sécurité:

- ✓ Avez-vous évalué des formes d'attaques potentielles auxquelles le système d'IA pourrait être vulnérable?
  - Avez-vous en particulier envisagé différents types et différentes natures de vulnérabilités, comme la pollution des données, l'infrastructure physique, les cyberattaques?
- ✓ Avez-vous prévu des mesures ou systèmes pour veiller à l'intégrité et à la résilience du système d'IA face à de potentielles attaques?
- ✓ Avez-vous évalué le comportement de votre système dans des situations ou des environnements imprévus?
- ✓ Avez-vous envisagé si, et dans quelle mesure, votre système pourrait avoir un double usage? Le cas échéant, avez-vous pris des mesures préventives appropriées contre un tel cas de figure (y compris, par exemple, ne pas publier la recherche ou ne pas déployer le système)?

Solutions de secours et sécurité générale:

- ✓ Avez-vous veillé à ce que votre système dispose de suffisamment de solutions de secours pour faire face à d'éventuelles attaques antagonistes ou autres situations imprévues (par exemple, procédures de relais technique ou demande de communication avec un opérateur humain avant d'agir)?
- ✓ Avez-vous envisagé le niveau de risque posé par le système d'IA dans ce cas d'utilisation spécifique?
  - Avez-vous mis en place un processus pour mesurer et évaluer les risques et la sécurité?
  - Avez-vous fourni les informations nécessaires en cas de risque pour l'intégrité physique humaine?
  - Avez-vous réfléchi à une politique d'assurance pour couvrir les dégâts potentiels provoqués par le système d'IA?
  - Avez-vous recensé les risques potentiels en matière de sécurité d'(autres) utilisations prévisibles de la technologie, y compris d'utilisation abusive accidentelle ou malveillante? Existe-t-il un plan pour atténuer ou gérer ces risques?
- ✓ Avez-vous évalué s'il est probable que le système d'IA cause des dommages ou préjudices aux utilisateurs ou à des tiers? Le cas échéant, avez-vous évalué la probabilité, les dommages potentiels, le public concerné et la gravité?
  - En cas de risque qu'un système d'IA cause des dommages, avez-vous réfléchi à des règles de responsabilité et de protection des consommateurs, et de quelle manière en avez-vous tenu compte?
  - Avez-vous réfléchi à l'incidence potentielle ou au risque en matière de sécurité sur l'environnement ou les animaux?
  - Vous êtes-vous demandé, dans le cadre de votre analyse des risques, si des problèmes de sécurité ou de réseau (par exemple, des menaces pesant sur la cybersécurité) pourraient mettre en péril la sécurité ou entraîner des préjudices du fait d'un comportement involontaire du

système d'IA?

- ✓ Avez-vous évalué l'incidence probable d'une défaillance de votre système d'IA entraînant la production de résultats erronés, l'indisponibilité de votre système, ou la production de résultats inacceptables pour la société (par exemple, pratiques discriminatoires)?
  - Avez-vous mis en place des seuils et une gouvernance pour les scénarios ci-dessus afin de déclencher d'autres plans/solutions de secours?
  - Avez-vous défini et testé des solutions de secours?

Précision

- ✓ Avez-vous évalué le niveau de précision et la définition de la précision nécessaires dans le contexte du système d'IA et du cas d'utilisation concerné?
  - Avez-vous réfléchi à la manière dont la précision est mesurée et assurée?
  - Avez-vous mis en place des mesures pour veiller à ce que les données utilisées soient exhaustives et à jour?
  - Avez-vous mis en place des mesures pour évaluer si des données supplémentaires sont nécessaires, par exemple pour améliorer la précision et éliminer les biais?
- ✓ Avez-vous évalué le préjudice que causeraient des prédictions inexactes du système d'IA?
- ✓ Avez-vous prévu des moyens de mesurer si votre système produit un nombre inacceptable de prédictions inexactes?
- ✓ En cas de prédictions inexactes, avez-vous mis en place une série d'étapes pour résoudre le problème?

Fiabilité et reproductibilité:

- ✓ Avez-vous mis en place une stratégie afin de contrôler le système d'IA et de vous assurer qu'il répond aux objectifs, aux finalités et aux applications prévues?
  - Avez-vous vérifié si des contextes spécifiques ou conditions particulières doivent être pris en compte pour garantir la reproductibilité?
  - Avez-vous mis en place des processus ou méthodes de vérification pour mesurer et garantir les différents aspects de la fiabilité et de la reproductibilité?
  - Avez-vous mis en place des processus visant à décrire certains réglages susceptibles d'entraîner une défaillance du système d'IA?
  - Avez-vous clairement documenté et appliqué ces processus aux fins des essais et de la vérification de la fiabilité du système d'IA?

    Avez-vous mis en place un mécanisme ou une communication pour garantir aux utilisateurs (finaux) la fiabilité du système d'IA?

3. <u>Respect de la vie privée et gouvernance des données</u>

Respect de la vie privée et protection des données:

- ✓ En fonction du cas d'utilisation, avez-vous mis sur pied un mécanisme permettant à autrui de signaler des problèmes en rapport avec le respect de la vie privée et la protection des données durant les processus suivis par le système d'IA pour la collecte des données (aux fins de l'entraînement et du fonctionnement) et le traitement des données?
- ✓ Avez-vous évalué le type et la portée des données constituant vos ensembles de données (par exemple, si elles contiennent des données à caractère personnel)?
- ✓ Avez-vous réfléchi à des manières de mettre au point le système d'IA ou d'entraîner le modèle sans utiliser (ou en utilisant de manière limitée) des données potentiellement sensibles ou à caractère personnel?
- ✓ Avez-vous intégré des mécanismes de notification et de contrôle concernant les données à caractère personnel en fonction du cas d'utilisation (comme un consentement valable et la possibilité de révoquer le consentement, le cas échéant)?
- ✓ Avez-vous pris des mesures pour renforcer le respect de la vie privée, par exemple des mesures de cryptage, d'anonymisation et d'agrégation?
- ✓ Lorsqu'il existe un responsable de la protection des données, avez-vous mobilisé cette personne à un stade précoce dans le processus?

Qualité et intégrité des données:

- ✓ Avez-vous aligné votre système sur d'éventuelles normes pertinentes (par exemple, ISO, IEEE) ou des protocoles largement adoptés dans le cadre de votre gestion et de votre gouvernance quotidiennes des données?
- ✓ Avez-vous mis sur pied des mécanismes de contrôle pour la collecte, le stockage, le traitement et l'utilisation des données?
- ✓ Avez-vous évalué la mesure dans laquelle vous contrôlez la qualité des sources externes des données utilisées?
- ✓ Avez-vous mis en place des processus pour garantir la qualité et l'intégrité de vos données? Avez-vous envisagé d'autres processus? De quelle manière vérifiez-vous que vos ensembles de données n'ont pas été compromis ou piratés?

Accès aux données:

- ✓ Quels protocoles, processus et procédures ont été suivis pour gérer et garantir la gouvernance appropriée des données?
  - Avez-vous analysé qui peut accéder aux données des utilisateurs et dans quelles circonstances?
  - Avez-vous veillé à ce que ces personnes soient qualifiées, qu'elles aient effectivement besoin d'accéder aux données et à ce qu'elles disposent des compétences nécessaires pour comprendre précisément la politique de protection des données?
  - Avez-vous prévu un mécanisme de contrôle pour consigner quand, où, comment, par qui et dans quel but les données ont été consultées?

4. Transparence

Traçabilité:

- ✓ Avez-vous mis des mesures en place susceptibles de garantir la traçabilité? Cela pourrait consister à documenter:
  - ➢ Les méthodes appliquées aux fins de la conception et de la mise au point du système algorithmique:
    - o dans le cas d'un système d'IA fondé sur des règles, la méthode de programmation ou la manière dont le modèle a été mis au point devraient être documentées;
    - o dans le cas d'un système d'IA fondé sur l'apprentissage, la méthode d'entraînement de l'algorithme, y compris quelles données d'entrée ont été collectées et sélectionnées, et dans quelles conditions, devrait être documentée.
  - ➢ Les méthodes appliquées pour tester et valider le système algorithmique:
    - o dans le cas d'un système d'IA fondé sur des règles, les scénarios ou cas utilisés pour tester et valider devraient être documentés;
    - o dans le cas d'un système d'IA fondé sur l'apprentissage, les informations relatives aux données utilisées pour tester et valider devraient être documentées.
  - ➢ Les résultats du système algorithmique:
    - o les résultats d'un algorithme ou les décisions qu'il prend, ainsi que les éventuelles autres décisions qui résulteraient de différents cas (par exemple, pour d'autres sous-groupes d'utilisateurs) devraient être documentés.

Explicabilité:

- ✓ Avez-vous évalué la mesure dans laquelle les décisions prises, et donc les résultats obtenus, par le système d'IA peuvent être compris?
- ✓ Avez-vous veillé à ce qu'une explication de la raison pour laquelle un système a procédé à un certain choix entraînant un certain résultat puisse être rendue compréhensible pour l'ensemble des utilisateurs qui pourraient souhaiter obtenir une explication?
- ✓ Avez-vous évalué la mesure dans laquelle la décision du système influence les processus décisionnels de l'organisation?
- ✓ Avez-vous évalué pourquoi ce système particulier a été déployé dans ce domaine spécifique?
- ✓ Avez-vous évalué le modèle économique concernant ce système (par exemple, en quoi crée-t-il de la valeur pour l'organisation)?
- ✓ Avez-vous conçu le système d'IA en ayant dès le départ l'interprétation à l'esprit?
  - ▪ Avez-vous cherché à utiliser le modèle le plus simple et le plus facile à interpréter pour l'application en question?
  - ▪ Avez-vous évalué si vous êtes en mesure d'analyser les données que vous avez utilisées aux fins de l'entraînement et des essais? Cela peut-il être modifié et actualisé au fil du temps?

- Avez-vous évalué si des solutions s'offrent à vous suite à l'entraînement et à la mise au point du modèle pour examiner l'interprétation ou si vous avez accès à la séquence des opérations du modèle?

Communication:

✓ Avez-vous informé les utilisateurs (finaux) – au moyen d'une clause de non-responsabilité ou de tout autre moyen – qu'ils interagissent avec un système d'IA et pas avec un autre être humain? Avez-vous indiqué clairement que votre système est doté de l'IA?

✓ Avez-vous mis en place des mécanismes pour informer les utilisateurs des raisons et critères expliquant les résultats du système d'IA?

- Les utilisateurs visés en sont-ils informés de manière claire et intelligible?
- Avez-vous établi des processus pour tenir compte des commentaires des utilisateurs et utiliser ces commentaires pour adapter le système?
- Avez-vous également communiqué les risques potentiels ou perçus, tels que les biais?
- En fonction du cas d'utilisation, avez-vous également réfléchi à la communication et à la transparence envers d'autres publics, des tiers ou le grand public?

✓ Avez-vous clairement indiqué la finalité du système d'IA et qui ou ce qui pourrait bénéficier du produit/service?

- Les scénarios d'utilisation du produit ont-ils été définis et clairement expliqués, en envisageant également d'autres moyens de communication pour veiller à ce qu'ils soient compréhensibles et appropriés pour le destinataire?
- En fonction du cas d'utilisation, avez-vous réfléchi à la psychologie humaine et aux potentielles limites humaines, comme le risque de confusion, les biais de confirmation ou la fatigue cognitive?

✓ Avez-vous clairement expliqué les caractéristiques, les limites et les éventuelles lacunes du système d'IA:

- s'agissant de la mise au point: à toute personne chargée de son déploiement pour en faire un produit ou service?
- s'agissant du déploiement: à l'utilisateur final ou au consommateur?

5. <u>Diversité, non-discrimination et équité</u>

Éviter les biais injustes:

✓ Avez-vous prévu une stratégie ou un ensemble de procédures pour éviter de créer ou de renforcer des biais injustes dans le système d'IA, en ce qui concerne tant l'utilisation des données d'entrée que la conception de l'algorithme?

- Avez-vous évalué et reconnu les éventuelles limites provenant de la composition des ensembles de données utilisés?
- Avez-vous réfléchi à la diversité et à la représentativité des utilisateurs dans les données? Avez-

vous procédé à des essais portant sur des populations spécifiques ou des cas d'utilisation problématiques?

- Avez-vous recherché et utilisé les outils techniques disponibles pour améliorer votre compréhension des données, du modèle et de la performance?
- Avez-vous mis en place des processus pour tester et contrôler les biais éventuels au cours de la phase de mise au point, de déploiement et d'utilisation du système?

✓ En fonction du cas d'utilisation, avez-vous prévu un mécanisme permettant à autrui de signaler des problèmes liés aux biais, à la discrimination ou aux mauvaises performances du système d'IA?

- Avez-vous envisagé des mesures et des moyens de communication clairs pour savoir comment et à qui ces problèmes peuvent être signalés?
- Avez-vous tenu compte non seulement des utilisateurs (finaux) mais également des autres personnes susceptibles d'être indirectement affectées par le système d'IA?

✓ Avez-vous évalué si, dans des conditions identiques, une éventuelle variabilité des décisions est possible?

- Le cas échéant, avez-vous réfléchi aux causes probables?
- Concernant la variabilité, avez-vous mis sur pied un mécanisme de mesure ou d'évaluation de l'incidence potentielle de cette variabilité sur les droits fondamentaux?

✓ Avez-vous prévu une définition appropriée de l'«équité» que vous appliquez dans la conception des systèmes d'IA?

- Votre définition est-elle couramment utilisée? Avez-vous envisagé d'autres définitions avant de choisir celle-ci?
- Avez-vous prévu une analyse quantitative ou des indicateurs pour mesurer et tester la définition appliquée de l'équité?
- Avez-vous mis sur pied des mécanismes visant à garantir l'équité dans vos systèmes d'IA? Avez-vous envisagé d'autres mécanismes potentiels?

Accessibilité et conception universelle:

✓ Avez-vous veillé à ce que le système d'IA réponde aux besoins d'un large ensemble de préférences et de capacités individuelles?

- Avez-vous évalué si le système d'IA peut être utilisé par les personnes présentant des besoins spécifiques ou un handicap ou qui sont exposées au risque d'exclusion? Comment cet aspect a-t-il été intégré à la conception du système et comment est-il vérifié?
- Avez-vous veillé à ce que les informations relatives au système d'IA soient également accessibles aux utilisateurs de technologies d'assistance?
- Avez-vous mobilisé ou consulté cette communauté d'utilisateurs au cours de la phase de mise au point du système d'IA?

✓ Avez-vous tenu compte de l'incidence de votre système d'IA sur le groupe d'utilisateurs potentiels?

- L'équipe participant à la mise au point du système d'IA est-elle représentative de votre groupe cible d'utilisateurs? Est-elle représentative de la population au sens large, compte tenu également d'autres groupes susceptibles d'être indirectement concernés?
- Avez-vous évalué si certaines personnes ou certains groupes pourraient subir de manière disproportionnée des effets négatifs?
- D'autres équipes ou groupes présentant différents parcours professionnels et expériences vous ont-ils fait parvenir des réactions?

Participation des parties prenantes:

✓ Avez-vous réfléchi à un mécanisme pour inclure la participation de différentes parties prenantes dans la mise au point et l'utilisation du système d'IA?

✓ Avez-vous préparé la voie à l'introduction du système d'IA au sein de votre organisation en informant et en mobilisant au préalable les travailleurs concernés et leurs représentants?

6. <u>Bien-être sociétal et environnemental</u>

IA durable et respectueuse de l'environnement:

✓ Avez-vous mis en place des mécanismes pour mesurer l'impact environnemental de la mise au point, du déploiement et de l'utilisation du système d'IA (par exemple, énergie consommée par les centres de données, type d'énergie consommée par les centres de données, etc.)?

✓ Avez-vous prévu des mesures pour réduire l'impact environnemental du cycle de vie de votre système d'IA?

Incidence sociale:

✓ Lorsque le système d'IA interagit directement avec des êtres humains:

- Avez-vous évalué si le système d'IA encourage les êtres humains à développer de l'attachement et de l'empathie pour le système?
- Avez-vous veillé à ce que le système d'IA indique clairement que son interaction sociale est simulée et qu'il n'a nullement la capacité de «comprendre» et de «ressentir»?

✓ Avez-vous veillé à ce que les incidences sociales du système d'IA soient bien comprises? Par exemple, vous êtes-vous demandé s'il existe un risque de perte d'emplois et de perte de compétences de la main-d'œuvre? Quelles mesures ont été prises pour contrer ces risques?

Société et démocratie:

✓ Avez-vous évalué l'incidence plus large de l'utilisation du système d'IA sur la société, au-delà de l'utilisateur (final) individuel, par exemple les parties prenantes susceptibles d'être indirectement concernées?

7. <u>Responsabilité</u>

Auditabilité:

- ✓ Avez-vous mis en place des mécanismes pour faciliter l'auditabilité du système par des acteurs internes et/ou indépendants, en veillant pas exemple à la traçabilité et à la journalisation des processus et des résultats du système d'IA?

Minimisation et documentation des incidences négatives:

- ✓ Avez-vous réalisé une analyse des risques ou de l'impact du système d'IA qui tienne compte de différentes parties prenantes qui sont directement et indirectement concernées?
- ✓ Avez-vous mis en place des cadres de formation et d'éducation pour définir des pratiques en matière de responsabilité?
  - Quels travailleurs ou branches de travailleurs sont concernés? Le sont-ils au-delà de la phase de mise au point?
  - Ces formations portent-elles également sur le cadre juridique potentiellement applicable au système d'IA?
  - Avez-vous envisagé la mise sur pied d'un «comité d'examen pour l'IA éthique» ou d'un mécanisme similaire pour discuter des pratiques globales en matière de responsabilité et d'éthique, y compris des zones grises potentiellement floues?
- ✓ Outre les initiatives ou cadres internes destinés à contrôler l'éthique et la responsabilité, des orientations externes ou des processus d'audit ont-ils également été mis en place?
- ✓ Existe-t-il des processus permettant aux tiers (par exemple, fournisseurs, consommateurs, distributeurs/vendeurs) ou aux travailleurs de signaler de possibles vulnérabilités, risques ou biais dans le système/l'application d'IA?

Documentation des arbitrages:

- ✓ Avez-vous mis sur pied un mécanisme permettant de recenser les intérêts et les valeurs pertinents concernés par le système d'IA et les éventuels arbitrages entre eux?
- ✓ Quel processus utilisez-vous pour prendre des décisions relatives à ces arbitrages? Avez-vous veillé à ce que les décisions d'arbitrage soient documentées?

Voies de recours:

- ✓ Avez-vous mis en place un ensemble approprié de mécanismes permettant un recours en cas de préjudice ou d'effet néfaste?
- ✓ Avez-vous mis en place des mécanismes pour fournir des informations aux utilisateurs (finaux)/tiers à propos des possibilités de recours?

Nous invitons l'ensemble des parties prenantes à tester la liste d'évaluation en pratique et à nous faire parvenir leurs réactions concernant son potentiel de mise en œuvre, son exhaustivité, sa pertinence à l'égard de l'application ou du domaine d'IA spécifique, et concernant tout chevauchement ou toute complémentarité avec des processus existants en matière de conformité ou d'évaluation. Sur la base de ces commentaires, une version révisée de la liste d'évaluation pour une IA digne de confiance sera proposée à la Commission début 2020

Orientations essentielles dérivées du chapitre III:

- ✓ Adopter une **liste d'évaluation** pour une IA digne de confiance au stade de la mise au point, du déploiement ou de l'utilisation de systèmes d'IA, et l'adapter au cas d'utilisation spécifique du système.
- ✓ Garder à l'esprit qu'une liste d'évaluation de cette nature **ne sera jamais exhaustive**. Il ne suffit pas de cocher des cases pour garantir une IA digne de confiance. Il convient de déterminer des exigences, d'évaluer des solutions et de garantir l'amélioration des résultats de manière continue tout au long du cycle de vie du système d'IA, et de mobiliser les parties prenantes.

C. EXEMPLES DE POSSIBILITES ET DE PREOCCUPATIONS MAJEURES SOULEVEES PAR L'IA

(121) Dans la section qui suit, nous présentons des exemples de mise au point et d'utilisation d'une IA qui devraient être encouragés, ainsi que des exemples de situations dans lesquelles la mise au point, le déploiement ou l'utilisation d'une IA peuvent nuire à nos valeurs et peuvent soulever des préoccupations spécifiques. Un équilibre doit être trouvé entre ce qui devrait et ce qui peut être fait avec l'IA. Il convient en outre de rester vigilant à ce qui doit être évité avec l'IA.

1. Exemples de possibilités offertes par une d'IA digne de confiance

(122) L'IA digne de confiance peut représenter un formidable potentiel pour contribuer à l'atténuation des problèmes urgents auxquels notre société est confrontée, tels que le vieillissement de la population, l'accroissement des inégalités sociales et la pollution de l'environnement. Ce potentiel est également reflété au niveau mondial, par exemple avec les objectifs de développement durable des Nations unies.[57] La section qui traite de la manière d'encourager une stratégie européenne dans le domaine de l'IA qui réponde à certains de ces problèmes.

a. Action pour le climat et infrastructures durables

(123) S'il est vrai que la lutte contre le changement climatique devrait être une priorité absolue pour les décideurs politiques du monde entier, la transformation numérique et une IA digne de confiance présentent un énorme potentiel pour réduire l'incidence humaine sur l'environnement et permettre une utilisation efficiente et efficace de l'énergie et des ressources naturelles[58]. Une IA digne de confiance peut par exemple être combinée aux mégadonnées pour détecter les besoins en énergie de manière plus efficace, ce qui donnerait lieu à des infrastructures et à une consommation énergétiques plus efficaces[59].

(124) Dans les secteurs tels que les transports publics, des systèmes d'IA appliqués à des systèmes de transport intelligents[60] peuvent être utilisés pour réduire le plus possible les files, optimiser l'itinéraire, aider les personnes souffrant de problèmes de vue à être plus indépendantes[61], optimiser les moteurs efficaces d'un point de vue énergétique et renforcer ainsi les efforts de décarbonation et réduire l'empreinte environnementale, pour une société plus verte. Aujourd'hui, à l'échelle mondiale, une personne meurt toutes

57 https://sustainabledevelopment.un.org/?menu=1300.

58 Un certain nombre de projets de l'UE visent à développer les réseaux intelligents et le stockage de l'énergie, qui peuvent potentiellement contribuer au succès d'une transition énergétique soutenue par les technologies numériques, y compris via des solutions fondées sur l'IA et d'autres solutions numériques. Pour compléter le travail de ces projets individuels, la Commission a lancé l'initiative BRIDGE, qui permet aux projets de réseaux intelligents et de stockage de l'énergie qui sont en cours dans le cadre d'Horizon 2020 d'établir une vision commune sur des questions transversales: https://www.h2020-bridge.eu/.

59 Voir par exemple le projet Encompass: http://www.encompass-project.eu/.

60 De nouvelles solutions fondés sur l'IA contribuent à préparer les villes à la mobilité de demain. Voir par exemple le projet financé par l'UE appelé Fabulos: https://fabulos.eu/.

61 Voir par exemple le projet PRO4VIP, qui fait partie de la stratégie européenne Vision 2020 pour combattre la cécité évitable, principalement due au vieillissement. La mobilité et l'orientation faisaient partie des domaines prioritaires du projet.

les 23 secondes dans un accident de voiture[62]. Les systèmes d'IA pourraient contribuer à réduire considérablement le nombre de victimes, par exemple grâce à une amélioration des temps de réaction et à un meilleur respect des règles[63].

b. Santé et bien-être

(125) Les technologies relevant de l'IA digne de confiance peuvent être utilisées – et le sont déjà – pour rendre les traitements plus intelligents et mieux ciblés, et contribuer à la prévention des maladies mortelles[64]. Les médecins et professionnels de la santé peuvent potentiellement réaliser un examen plus précis et détaillé de données de santé complexes relatives à un patient, avant même l'apparition d'une maladie, et administrer un traitement préventif sur mesure[65]. Dans le contexte du vieillissement de la population de l'Europe, l'IA et la robotique peuvent être des outils précieux pour assister les prestataires de soins et favoriser les soins prodigués aux personnes âgées[66], ainsi que pour surveiller en temps réel l'état de santé des patients, et donc sauver des vies[67].

(126) Une IA digne de confiance peut également être utile à une plus grande échelle. Par exemple, elle peut examiner et déterminer des tendances générales dans le secteur des soins de santé et des traitements[68], ce qui mène à une détection plus précoce des maladies, à un développement plus efficace des médicaments, à des traitements davantage ciblés[69] et, en fin de compte, à un plus grand nombre de vies sauvées.

c. Éducation de qualité et transformation numérique

(127) Les nouveaux changements technologiques, économiques et environnementaux impliquent que la société doit devenir plus proactive. Les pouvoirs publics, les leaders des secteurs concernés, les établissements d'enseignement et les syndicats ont la responsabilité de faire entrer les citoyens dans la nouvelle ère numérique en veillant à ce qu'ils disposent des compétences requises pour occuper les emplois de demain. Les technologies relevant de l'IA digne de confiance pourraient contribuer à améliorer la précision des prévisions relatives aux emplois et aux professions qui seront le plus perturbés par la technologie, aux nouveaux rôles qui

62 https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries.

63 Le projet européen UP-Drive vise par exemple à apporter des solutions aux problèmes signalés en matière de transport, par des contributions permettant l'automatisation et la collaboration progressives des véhicules entre eux, facilitant ainsi un système de transports plus sûr, plus inclusif et plus abordable. https://up-drive.eu/.

64 Voir par exemple le projet REVOLVER (Repeated Evolution of Cancer): https://www.healtheuropa.eu/personalised-cancer-treatment/87958/, ou le projet Murab qui réalise des biopsies plus précises, et qui vise à diagnostiquer plus rapidement le cancer et d'autres maladies: https://ec.europa.eu/digital-single-market/en/news/murab-eu-funded-project-success-story.

65 Voir par exemple le projet Live INCITE: www.karolinska.se/en/live-incite. Ce consortium d'acteurs du secteur de la santé incite ce secteur à mettre au point des solutions d'IA et d'autres solutions TIC intelligentes qui permettent des modifications du mode de vie dans le processus peropératoire. L'objectif concerne de nouvelles solutions innovantes de santé en ligne capables d'influencer les patients de façon personnalisée, afin qu'ils prennent les mesures nécessaires, tant avant qu'après la chirurgie, dans leur mode de vie pour optimiser le résultat des soins.

66 Le projet CARESSES financé par l'UE concerne des robots destinés aux soins aux personnes âgées, centrés sur leur sensibilité culturelle: ils adaptent leur manière d'agir et de parler pour correspondre à la culture et aux habitudes des personnes âgées qu'ils aident: http://caressesrobot.org/en/project/. Voir également l'application d'IA appelée Alfred, un assistant virtuel aidant les personnes âgées à rester actives: https://ec.europa.eu/digital-single-market/en/news/alfred-virtual-assistant-helping-older-people-stay-active. En outre, le projet EMPATTICS (EMpowering PAtients for a BeTTer Information and improvement of the Communication Systems) effectuera des recherches pour définir la manière dont les professionnels de la santé et les patients utilisent les technologies TIC, y compris les systèmes d'IA, pour planifier les interventions avec les patients et surveiller la progression de leur état physique et mental: www.empattics.eu.

67 Voir par exemple MyHealth Avatar (www.myhealthavatar.eu), qui propose une représentation numérique de l'état de santé d'un patient. Le projet de recherche a lancé une application et une plateforme en ligne qui collecte les informations concernant votre état de santé numérique à long terme, et vous y donne accès. Cela se présente sous la forme d'un compagnon de santé pour toute la vie («avatar»). MyHealthAvatar prédit également les risques d'accident vasculaire cérébral, de diabète, de maladie cardiovasculaire et d'hypertension.

68 Voir par exemple le projet ENRICHME (www.enrichme.eu), qui lutte contre la perte progressive des capacités cognitives au sein de la population vieillissante. Une plateforme intégrée d'assistance à l'autonomie à domicile (AAD) et un robot de service mobile pour un suivi et des interactions à long terme aideront les personnes âgées à rester plus longtemps actives et indépendantes.

69 Voir par exemple l'utilisation de l'IA par Sophia Genetics, qui tire parti de l'inférence statistique, de la reconnaissance de modèles et de l'apprentissage automatique pour optimiser la valeur de la génomique et des données d'imagerie médicale: https://www.sophiagenetics.com/home.html.

seront créés et aux compétences qui seront nécessaires. Elles pourraient aider les pouvoirs publics, les syndicats et les secteurs concernés à planifier la (re)qualification des travailleurs, ainsi qu'offrir aux citoyens craignant un licenciement une voie de développement dans un nouveau rôle.

(128) En outre, l'IA peut s'avérer être un excellent outil pour combattre les inégalités en matière d'éducation et créer des programmes de formation personnalisés et adaptables qui pourraient aider chacun à obtenir de nouvelles qualifications, aptitudes et compétences en fonction de ses propres capacités d'apprentissage[70]. Cela pourrait augmenter la vitesse d'apprentissage et améliorer la qualité de l'enseignement – de l'école primaire à l'université.

2. Exemples de préoccupations majeures soulevées par l'IA

(129) La violation de l'un des éléments constitutifs d'une IA digne de confiance est de nature à susciter une préoccupation majeure. Un nombre important des préoccupations présentées ci-dessous tiennent déjà aux exigences juridiques en vigueur qui, étant contraignantes, doivent par conséquent être respectées. Toutefois, même dans les cas où la conformité avec les exigences juridiques a été démontrée, celles-ci pourraient ne pas répondre à l'éventail complet de préoccupations éthiques susceptibles d'être soulevées. Étant donné que notre conception de l'adéquation des règles et principes éthiques est en constante évolution et peut changer au fil du temps, la liste non exhaustive suivante de préoccupations pourrait à l'avenir être raccourcie, élargie, modifiée ou actualisée.

a. Identifier et suivre des individus avec l'IA

(130) L'IA permet une identification encore plus efficace de personnes par des entités tant publiques que privées. La reconnaissance faciale et d'autres méthodes involontaires d'identification utilisant des données biométriques (à savoir, détecteur de mensonges, évaluation de la personnalité au moyen de micro-expressions et détection vocale automatique) sont des exemples notoires de technologies évolutives d'identification fondées sur l'IA. L'identification d'individus est parfois le résultat souhaitable, lorsqu'elle s'accompagne de principes éthiques (par exemple pour la détection de fraude, de blanchiment de capitaux, ou de financement du terrorisme). Toutefois, l'identification automatique soulève de fortes préoccupations de nature tant juridique qu'éthique, car elle peut avoir des effets inattendus à de nombreux niveaux psychologiques et socioculturels. Pour préserver l'autonomie des citoyens européens, il est nécessaire d'utiliser de manière proportionnée les techniques de contrôle dans le domaine de l'IA. Pour parvenir à mettre en œuvre une IA digne de confiance, il sera essentiel de définir clairement si, quand et comment l'IA peut être utilisée aux fins de l'identification automatique de personnes, ainsi que de faire la distinction entre l'identification d'un individu et le fait de le suivre à la trace, et entre une surveillance ciblée et une surveillance de masse. L'application de telles technologies doit être clairement justifiée dans la législation applicable[71]. Lorsque la base juridique pour une activité de cette nature est le «consentement», des moyens pratiques[72] doivent être développés pour permettre qu'un consentement éclairé et vérifié soit automatiquement recensé par une IA ou des technologies équivalentes. Cela s'applique également à l'utilisation de données à caractère personnel «anonymes» pouvant être re-personnalisées.

b. Systèmes d'IA cachés

[70] Voir par exemple le projet MaTHiSiS, visant à offrir une solution pour l'apprentissage fondé sur les affects dans un environnement d'apprentissage confortable, comprenant des dispositifs technologiques et des algorithmes de pointe: (http://mathisis-project.eu/). Voir également IBM Watson Classroom ou la plateforme Century Tech.

[71] Il convient à cet égard de rappeler l'article 6 du RGPD, qui prévoit notamment que le traitement de données n'est licite que s'il s'appuie sur une base juridique valable.

[72] Comme le montrent les mécanismes actuellement utilisés sur l'internet pour donner un consentement éclairé, les consommateurs accordent en général leur consentement sans véritable considération. Ces mécanismes ne peuvent dès lors que difficilement être qualifiés de pratiques.

(131) Les êtres humains devraient toujours savoir s'ils interagissent directement avec un autre être humain ou une machine, et les professionnels de l'IA ont la responsabilité de veiller à ce que ce soit effectivement le cas. Les professionnels de l'IA doivent par conséquent veiller à ce que les êtres humains soient informés du fait qu'ils interagissent avec un système d'IA – ou soient en mesure de le demander et de le confirmer – (par exemple, au moyen de clauses de non-responsabilité claires et transparentes). Il convient de noter que des cas limites existent et sont de nature à compliquer la question (par exemple, une voix filtrée par IA appartenant à un être humain). Il convient de garder à l'esprit que la confusion entre êtres humains et machines pourrait entraîner de multiples conséquences, telles qu'un attachement, une certaine influence, ou une réduction de la valeur accordée à la qualité d'être humain[73]. Le développement de robots humanoïdes[74] doit par conséquent faire l'objet d'une évaluation éthique minutieuse.

c. Notation des citoyens assistée par l'IA en violation des droits fondamentaux

(132) Les sociétés doivent s'efforcer de protéger la liberté et l'autonomie de tous les citoyens. Toute forme de notation des citoyens peut entraîner la perte de cette autonomie et mettre en péril le principe de non-discrimination. La notation ne doit être utilisée que si elle se justifie clairement et lorsque les mesures sont proportionnées et équitables. La notation normative de citoyens (évaluation générale de la «personnalité morale» ou de l'«intégrité éthique») dans *tous* les aspects et à grande échelle de la part des autorités publiques ou d'acteurs privés menace ces valeurs, notamment lorsqu'il y est recouru de manière non conforme aux droits fondamentaux et de manière disproportionnée sans objectif légitime délimité et communiqué.

(133) De nos jours, la notation des citoyens – à grande ou plus petite échelle – est déjà souvent utilisée dans le cadre de notations purement descriptives et spécifiques à un domaine (par exemple, systèmes scolaires, apprentissage en ligne et permis de conduire). Même dans ces applications plus étroites, une procédure totalement transparente doit être mise à la disposition des citoyens et comprendre des informations sur le processus, l'objectif et la méthodologie de la notation. Il convient de souligner que la transparence ne peut prévenir la discrimination ou garantir l'équité. Il ne s'agit en outre pas de la panacée face au problème de la notation. Idéalement, il devrait être possible de retirer sa participation au mécanisme de notation sans préjudice – dans le cas contraire, des mécanismes pour contester et rectifier les notes devraient être disponibles. Cela est particulièrement important dans les situations présentant des asymétries de pouvoir entre les parties. De telles options de retrait, qui sont nécessaires dans une société démocratique, doivent être garanties au stade de la conception de la technologie dans les cas où cela est nécessaire pour veiller au respect des droits fondamentaux.

d. Systèmes d'armes létales autonomes (SALA)

(134) À l'heure actuelle, un nombre inconnu de pays et d'industries recherchent et développent des systèmes d'armes létales autonomes, allant de missiles capables de sélectionner des cibles à des calculateurs autoadaptatifs dotés de compétences cognitives pour décider par qui, quand et où des combats peuvent être menés sans intervention humaine. Cela suscite des préoccupations éthiques fondamentales, comme le fait qu'une course à l'armement incontrôlable à un niveau jamais égalé dans l'histoire pourrait en résulter, ainsi que des contextes militaires dans lesquels le contrôle humain a presque totalement été abandonné et les risques de défaillances ne sont pas éliminés. Le Parlement européen a appelé à l'élaboration urgente d'une position commune contraignante pour régir les questions éthiques et juridiques du contrôle humain, de la surveillance, de la responsabilité et de la mise en œuvre du droit international relatif aux droits de l'homme, du droit humanitaire international et des stratégies militaires[75]. Rappelant l'objectif de l'Union européenne de

[73] Madary et Metzinger (2016). Real Virtuality: A Code of Ethical Conduct. Recommendations for Good Scientific Practice and the Consumers of VR-Technology. Frontiers in Robotics and AI, 3(3).

[74] Cela s'applique également aux avatars fondés sur l'IA.

[75] Résolution 2018/2752(RSP) du Parlement européen.

promouvoir la paix, tel que consacré à l'article 3 du traité sur l'Union européenne, nous saluons la résolution du Parlement du 12 septembre 2018 et tous les efforts y relatifs dans le domaine des SALA, que nous aspirons à soutenir.

e. Préoccupations potentielles à plus long terme

(135) La mise au point de l'IA reste spécifique à un domaine et requiert des scientifiques et ingénieurs humains correctement formés pour définir ses objectifs avec précision. Toutefois, en extrapolant à un horizon plus lointain, certaines grandes préoccupations à long terme peuvent faire l'objet d'hypothèses[76]. Une approche fondée sur les risques indique qu'il convient de continuer à tenir compte de ces préoccupations dans la perspective de possibles «inconnues inconnues» et «cygnes noirs»[77]. Les fortes incidences inhérentes à ces préoccupations, combinées à l'incertitude actuelle des évolutions correspondantes, requièrent des évaluations régulières de ces sujets.

D. CONCLUSION

(136) Le présent document constitue les lignes directrices en matière d'éthique dans le domaine de l'IA, élaborées par le groupe d'experts de haut niveau sur l'intelligence artificielle (GEHN IA).

(137) Nous reconnaissons les effets positifs que les systèmes d'IA ont déjà et continuerons à avoir, tant d'un point de vue commercial que pour la société. Nous sommes toutefois tout aussi soucieux de faire en sorte que les risques et autres effets néfastes auxquels sont associées ces technologies soient gérés de manière adéquate et proportionnée eu égard à l'application de l'IA concernée. L'IA est une technologie à la fois transformatrice et perturbatrice, et son évolution au cours des dernières années a été facilitée par la disponibilité de quantités gigantesques de données numériques, des avancées technologiques majeures en matière de puissance de calcul et de capacité de stockage, ainsi que par d'importantes innovations scientifiques et en ingénierie concernant les méthodes et outils de l'IA. L'incidence des systèmes d'IA sur la société et les citoyens se poursuivra sous des formes que nous ne pouvons pas encore imaginer.

(138) Dans ce contexte, il est important de mettre au point des systèmes d'IA dignes de confiance, car les êtres humains ne seront en mesure de tirer pleinement parti des avantages de l'IA en toute sérénité que s'ils peuvent se fier à la technologie, y compris aux processus et aux personnes qui la soutiennent. Lors de l'élaboration des présentes lignes directrices, notre ambition fondamentale a par conséquent consisté à tendre vers une IA digne de confiance.

(139) Une IA digne de confiance comporte trois éléments: 1) elle doit être licite, en assurant le respect des législations et réglementations applicables, 2) elle doit être éthique, en assurant l'adhésion à des principes et valeurs éthiques, et 3) elle doit être robuste, sur le plan tant technique que social, pour faire en sorte que, même avec de bonnes intentions, les systèmes d'IA ne causent pas de préjudices involontaires. Chaque élément est nécessaire mais ne suffit pas pour parvenir à une IA digne de confiance. L'idéal serait que ces trois éléments fonctionnent en harmonie et se chevauchent. Lorsque des tensions apparaissent, nous devrions nous efforcer de les aligner.

(140) Au chapitre I, nous avons articulé les droits fondamentaux et un ensemble de principes éthiques correspondants qui sont essentiels dans le contexte d'une IA. Au chapitre II, nous avons présenté sept exigences essentielles que doivent respecter les systèmes d'IA pour parvenir à une IA digne de confiance. Nous avons proposé des méthodes tant techniques que non techniques pouvant être appliquées aux fins de leur

[76] Si certains estiment que l'intelligence artificielle générale, la conscience artificielle, les agents moraux artificiels, la superintelligence ou l'IA transformatrice peuvent être des exemples de telles préoccupations à long terme (qui n'existent actuellement pas), celles-ci sont considérées par beaucoup comme irréalistes.

[77] Un événement «cygne noir» est un événement très rare dont l'impact est pourtant important – rare à tel point qu'il pourrait ne jamais avoir été observé. De ce fait, la probabilité qu'un tel événement survienne ne peut en général être estimée que de manière très incertaine.

mise en œuvre. Enfin, au chapitre III, nous avons proposé une liste d'évaluation pour une IA digne de confiance qui peut contribuer à concrétiser les sept exigences. Dans la section finale, nous avons présenté des exemples de potentiels bénéfiques et de préoccupations majeures soulevées par les systèmes d'IA, à propos desquels nous espérons encourager la poursuite des discussions.

(141) La perspective unique de l'Europe tient au fait qu'elle s'efforce de placer les citoyens au cœur de son action. Ce centrage sur les citoyens fait partie de l'ADN de l'Union européenne grâce aux traités sur lesquels elle s'est construite. Le présent document s'inscrit dans une vision qui encourage une IA digne de confiance qui, selon nous, doit être le fondement sur lequel l'Europe peut s'imposer comme leader du secteur des systèmes d'IA de pointe et innovants. Cette vision ambitieuse contribuera à garantir la prospérité des citoyens européens, sur le plan tant individuel que collectif. Notre but est de mettre en place une culture de l'«IA digne de confiance pour l'Europe», au moyen de laquelle chacun pourra récolter les fruits de l'IA dans le respect de nos valeurs fondatrices: les droits fondamentaux, la démocratie et l'état de droit.

GLOSSAIRE

(142) Le présent glossaire fait partie intégrante des lignes directrices et sert à comprendre les termes employés dans celles-ci.

Systèmes d'intelligence artificielle ou IA

(143) Les systèmes d'intelligence artificielle (IA) sont des systèmes logiciels (et éventuellement matériels) conçus par des êtres humains[78] et qui, ayant reçu un objectif complexe, agissent dans le monde réel ou numérique en percevant leur environnement par l'acquisition de données, en interprétant les données structurées ou non structurées collectées, en appliquant un raisonnement aux connaissances, ou en traitant les informations, dérivées de ces données et en décidant de la (des) meilleure(s) action(s) à prendre pour atteindre l'objectif donné. Les systèmes d'IA peuvent soit utiliser des règles symboliques ou apprendre un modèle numérique, et peuvent également adapter leur comportement en analysant la mesure dans laquelle l'environnement est affecté par leurs actions préalables.

(144) En tant que discipline scientifique, l'IA comprend plusieurs approches et techniques, telles que l'apprentissage automatique (dont l'apprentissage profond et l'apprentissage par renforcement sont des exemples spécifiques), le raisonnement automatique (qui comprend la planification, la programmation, la représentation des connaissances et le raisonnement, la recherche et l'optimisation) et la robotique (qui comprend le contrôle, la perception, les capteurs et les actionneurs, ainsi que l'intégration de toutes les autres techniques dans des systèmes cyberphysiques).

(145) Un document distinct élaboré par le GEHN IA et développant la définition des *systèmes d'IA* utilisée aux fins du présent document est publié parallèlement et s'intitule «Définition de l'IA: principales capacités et disciplines scientifiques».

Professionnels de l'IA

(146) On entend par professionnels de l'IA l'ensemble des personnes et organisations qui mettent au point (ce qui comprend la recherche, la conception et la fourniture de données), déploient (ce qui comprend la mise en œuvre) ou utilisent des systèmes d'IA, à l'exception de celles qui utilisent des systèmes d'IA en tant qu'utilisateurs finaux ou consommateurs.

Cycle de vie du système d'IA

(147) Le cycle de vie d'un système d'IA comprend sa phase de mise au point (dont la recherche, la conception, la fourniture de données et des essais limités), de déploiement (dont la mise en œuvre) et d'utilisation.

Auditabilité

(148) L'auditabilité désigne la capacité d'un système d'IA à faire l'objet d'une évaluation de ses algorithmes, de ses données et de ses processus de conception. Elle constitue une des sept exigences que doit respecter un système d'IA digne de confiance. Cela ne signifie pas nécessairement que les informations sur les modèles économiques et la propriété intellectuelle en lien avec le système d'IA doivent toujours être librement accessibles. Prévoir des mécanismes de traçabilité et de journalisation dès la phase de conception initiale du système d'IA peut contribuer à l'auditabilité du système.

Biais

(149) Le biais est une inclination au préjugé envers ou contre une personne, un objet ou un point de vue. Des biais peuvent se manifester de multiples manières dans les systèmes d'IA. Par exemple, dans les systèmes d'IA fondés sur les données, tels que ceux produits par apprentissage automatique, des biais présents dans la

[78] Les êtres humains conçoivent des systèmes d'IA directement, mais peuvent également avoir recours à des techniques d'IA pour optimiser leur conception.

collecte de données et l'entraînement peuvent être à l'origine de la présence de biais dans le système d'IA. Dans l'IA fondée sur la logique, comme les systèmes fondés sur des règles, le biais peut résulter de la manière dont un ingénieur des connaissances envisage les règles s'appliquant à un contexte particulier. Le biais peut également résulter de l'apprentissage en ligne et de l'adaptation par interaction. Il peut également se manifester à travers la personnalisation, par laquelle les utilisateurs reçoivent des recommandations ou des informations correspondant à leurs préférences. Il n'est pas nécessairement le résultat d'un biais humain et de la collecte de données par des êtres humains. Le biais peut, par exemple, se manifester dans les circonstances des contextes limités dans lesquels le système est utilisé, auquel cas il n'est pas possible de le généraliser à d'autres contextes. Un biais peut être positif ou négatif, intentionnel ou involontaire. Dans certains cas, le biais peut entraîner des résultats discriminatoires et/ou injustes, qualifiés de biais injustes dans le présent document.

Éthique

(150) L'éthique est une discipline qui est un sous-champ de la philosophie. De manière générale, elle traite de questions telles que «Qu'est-ce qu'une bonne action?», «Quelle est la valeur d'une vie humaine?», «Qu'est-ce que la justice?», ou «Qu'est-ce qu'une bonne vie?». L'éthique théorique comprend quatre principaux domaines de recherche: i) la méta-éthique, qui concerne principalement la signification et la référence d'un énoncé normatif, et la question de savoir comment leurs valeurs de vérité peuvent être déterminées (le cas échéant); ii) l'éthique normative, les moyens pratiques de déterminer une conduite morale, en examinant les normes applicables aux bonnes et mauvaises actions et en attribuant une valeur aux actions spécifiques; iii) l'éthique descriptive, visant une analyse empirique des comportements moraux et croyances morales des personnes, et; iv) l'éthique appliquée, qui concerne ce que nous sommes obligés de (ou autorisés à) faire dans une situation spécifique (souvent une première) ou dans un domaine particulier de possibilités d'action (souvent sans précédent). L'éthique appliquée traite de situations réelles, dans lesquelles des décisions doivent être prises dans des délais limités, et souvent avec peu de rationalité. L'éthique en matière d'IA est en général considérée comme un exemple de l'éthique appliquée et se concentre sur les questions normatives soulevées par la conception, la mise au point, la mise en œuvre et l'utilisation de l'IA.

(151) Dans les débats d'éthique, les termes «morale» et «éthique» sont souvent employés. Le terme «morale» renvoie aux modèles concrets et factuels de comportements, d'habitudes et de conventions qui peuvent s'observer au sein de cultures et de groupes ou auprès d'individus spécifiques à un moment donné. Le terme «éthique» désigne une appréciation évaluative de ces actions et comportements concrets d'un point de vue systématique et théorique.

IA éthique

(152) Dans le présent document, le terme «IA éthique» désigne la mise au point, le déploiement et l'utilisation d'une IA qui garantit une conformité avec les normes éthiques, y compris les droits fondamentaux en tant que droits moraux spéciaux, les principes éthiques et les valeurs essentielles qui s'y rapportent. Il s'agit du deuxième des trois éléments essentiels pour parvenir à une IA digne de confiance.

IA centrée sur l'humain

(153) L'approche de l'IA centrée sur l'humain s'efforce de garantir que les valeurs humaines soient un élément central de la mise au point, du déploiement, de l'utilisation et du contrôle des systèmes d'IA, en veillant au respect des droits fondamentaux, y compris ceux consacrés par les traités de l'Union européenne et la charte des droits fondamentaux de l'Union européenne, qui se rejoignent tous dans un fondement commun ancré dans le respect de la dignité humaine, en vertu duquel l'être humain jouit d'un statut moral unique et inaliénable. Cela implique également la prise en compte de l'environnement naturel et des autres êtres vivants qui font partie de l'écosystème humain, ainsi qu'une approche durable permettant l'épanouissement des générations à venir.

Méthode de l'équipe rouge («red teaming»

(154) La méthode de l'équipe rouge («red teaming») est une pratique par laquelle une «équipe rouge», c'est-à-dire un groupe indépendant, met au défi une organisation d'améliorer son efficacité en assumant un rôle ou un point de vue antagoniste. Cette pratique sert notamment à l'identification et à la résolution des failles potentielles en matière de sécurité.

Reproductibilité

(155) La reproductibilité est une indication de la mesure dans laquelle une expérience en matière d'IA, dans le cadre d'un essai, présente un comportement identique lorsqu'elle est répétée dans les mêmes conditions.

IA robuste

(156) La robustesse d'un système d'IA englobe tant sa robustesse technique (adaptation à un contexte donné, tel que le domaine d'application ou la phase du cycle de vie) que sa robustesse d'un point de vue social (le système d'IA tient dûment compte du contexte et de l'environnement dans lesquels il fonctionne). Cela est essentiel pour garantir que, même avec de bonnes intentions, aucun préjudice involontaire ne puisse survenir. La robustesse est le troisième des trois éléments nécessaires pour parvenir à une IA digne de confiance.

Parties prenantes

(157) On entend par parties prenantes tous ceux qui mettent au point, conçoivent, déploient, utilisent l'IA ou mènent des recherches dans le domaine, ainsi que ceux qui sont (directement ou indirectement) concernés par l'IA – y compris, mais sans s'y limiter, les entreprises, organisations, chercheurs, services publics, institutions, organisations de la société civile, pouvoirs publics, régulateurs, partenaires sociaux, particuliers, citoyens, travailleurs et consommateurs.

Traçabilité

(158) La traçabilité d'un système d'IA désigne la capacité de suivre les données du système et les processus de mise au point et de déploiement du système, en général au moyen d'une identification documentée.

Confiance

(159) Nous reprenons la définition suivante tirée de la littérature: «La confiance se définit comme: 1) un ensemble de convictions spécifiques portant sur la bienveillance, la compétence, l'intégrité et la prévisibilité (convictions en matière de confiance); 2) la volonté d'une partie de dépendre d'une autre partie dans une situation risquée (intention de faire confiance), ou 3) la combinaison de ces éléments.»[79] Si la «confiance» n'est en général pas une propriété que l'on prête aux machines, le présent document vise à souligner l'importance d'être capable de faire confiance non seulement dans le fait que les systèmes d'IA sont conformes sur le plan juridique, respectent l'éthique et sont robustes, mais aussi que cette confiance peut être accordée à l'ensemble des personnes et des processus impliqués dans le cycle de vie du système d'IA.

IA digne de confiance

(160) Une IA digne de confiance comporte trois éléments: 1) elle doit être licite, en assurant le respect des législations et réglementations applicables, 2) elle doit être éthique, en assurant le respect de principes et de valeurs éthiques, ainsi qu'en assurant l'adhésion à ces principes et valeurs, et 3) elle doit être robuste, sur le plan tant technique que social, pour faire en sorte que, même avec de bonnes intentions, les systèmes d'IA ne causent pas de préjudices involontaires. Une IA digne de confiance concerne non seulement la fiabilité du système d'IA en tant que tel, mais comprend également la fiabilité de l'ensemble des processus et acteurs qui font partie du cycle de vie du système.

Personnes et groupes vulnérables

[79] Siau, K., Wang, W. (2018), Building Trust in Artificial Intelligence, Machine Learning, and Robotics, *CUTTER BUSINESS TECHNOLOGY JOURNAL* (31), p. 47–53.

(161) Il n'existe pas de définition juridique communément ou largement admise de la notion de «personnes vulnérables», étant donné leur hétérogénéité. La raison pour laquelle une personne ou un groupe est vulnérable dépend souvent du contexte. Les événements temporaires de la vie (comme l'enfance ou la maladie), les facteurs de marché (comme l'asymétrie de l'information ou le pouvoir de marché), les facteurs économiques (comme la pauvreté), les facteurs identitaires (comme le sexe, la religion ou la culture) ou d'autres facteurs peuvent jouer un rôle. La charte des droits fondamentaux de l'Union européenne prévoit les motifs suivants à son article 21 sur la non-discrimination, qui peuvent constituer un point de référence parmi d'autres: le sexe, la race, la couleur, les origines ethniques ou sociales, les caractéristiques génétiques, la langue, la religion ou les convictions, les opinions politiques ou toute autre opinion, l'appartenance à une minorité nationale, la fortune, la naissance, un handicap, l'âge et l'orientation sexuelle. D'autres articles de la législation régissent les droits de groupes spécifiques, outre ceux énumérés ci-dessus. Toute liste de cette nature ne saurait être exhaustive et peut varier au fil du temps. Un groupe vulnérable est un groupe de personnes partageant une ou plusieurs caractéristiques de vulnérabilité.

Le présent document a été élaboré par les membres du groupe d'experts de haut niveau sur l'IA

énumérés ci-dessous par ordre alphabétique

Pekka Ala-Pietilä, président du GEHN IA
AI Finland, Huhtamaki, Sanoma
Wilhelm Bauer
Fraunhofer
Urs Bergmann – Co-rapporteur
Zalando

Mária Bieliková
Université technique de Slovaquie à Bratislava
Cecilia Bonefeld-Dahl – Co-rapporteure
DigitalEurope
Yann Bonnet
ANSSI
Loubna Bouarfa
OKRA
Stéphan Brunessaux
Airbus
Raja Chatila
Initiative de l'IEEE pour l'éthique des systèmes intelligents/autonomes et Université de la Sorbonne
Mark Coeckelbergh
Université de Vienne
Virginia Dignum – Co-rapporteure
Université d'Umeå
Luciano Floridi
Université d'Oxford
Jean-Francois Gagné – Co-rapporteur
Element AI
Chiara Giovannini
ANEC
Joanna Goodey
Agence des droits fondamentaux
Sami Haddadin
École de robotique et IA de Munich
Gry Hasselbalch
The thinkdotank DataEthics et Université de Copenhague
Fredrik Heintz
Université de Linköping
Fanny Hidvegi
Access Now
Eric Hilgendorf
Université de Würzburg
Klaus Höckner
Hilfsgemeinschaft der Blinden und Sehschwachen
Mari-Noëlle Jégo-Laveissière
Orange
Leo Kärkkäinen
Nokia Bell Labs
Sabine Theresia Köszegi
Université technique de Vienne
Robert Kroplewski
Avocat et conseiller du gouvernement polonais
Elisabeth Ling
RELX

Pierre Lucas
Orgalim – Europe's technology industries
Ieva Martinkenaite
Telenor
Thomas Metzinger – Co-rapporteur
Université Johannes Gutenberg de Mayence et Association des universités d'Europe
Catelijne Muller
ALLAI Netherlands et CESE
Markus Noga
SAP
Barry O'Sullivan, vice-président du GEHN IA
University College de Cork
Ursula Pachl
BEUC
Nicolas Petit – Co-rapporteur
Université de Liège
Christoph Peylo
Bosch

Iris Plöger
BDI (Fédération allemande de l'industrie)
Stefano Quintarelli
Garden Ventures
Andrea Renda
Collège d'Europe et CEPS
Francesca Rossi
IBM
Cristina San José
Fédération bancaire de l'Union européenne
George Sharkov
Digital SME Alliance
Philipp Slusallek
Centre allemand de recherche en IA (DFKI)
Françoise Soulié Fogelman
Consultante en IA
Saskia Steinacker – Co-rapporteure
Bayer
Jaan Tallinn
Ambient Sound Investment
Thierry Tingaud
STMicroelectronics
Jakob Uszkoreit
Google
Aimee Van Wynsberghe – Co-rapporteure
Université technique de Delft
Thiébaut Weber
CES
Cecile Wendling
AXA
Karen Yeung – Co-rapporteure
Université de Birmingham

Aimee Van Wynsberghe et Karen Yeung ont été rapporteurs pour le présent document.

Pekka Ala-Pietilä préside le GEHN IA. Barry O'Sullivan est vice-président et coordonne la deuxième contribution du GEHN IA. Nozha Boujemaa, vice-présidente jusqu'au 1er février 2019, chargée de la coordination de la première contribution, a également contribué au contenu du présent document.

Nathalie Smuha a fourni un soutien rédactionnel.

BERKMAN KLEIN CENTER
FOR INTERNET & SOCIETY
AT HARVARD UNIVERSITY

Research Publication No. 2020-1
January 15, 2020

Principled Artificial Intelligence:
Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI

Jessica Fjeld
Nele Achten
Hannah Hilligoss
Adam Christopher Nagy
Madhulika Srikumar

23 Everett Street • Second Floor • Cambridge, Massachusetts 02138
+1 617.495.7547 • +1 617.495.7641 (fax) • http://cyber.law.harvard.edu/ •
cyber@law.harvard.edu

PRINCIPLED ARTIFICIAL INTELLIGENCE:

Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI

Jessica Fjeld, Nele Achten, Hannah Hilligoss,
Adam Christopher Nagy, Madhulika Srikumar

BERKMAN
KLEIN CENTER
FOR INTERNET & SOCIETY
AT HARVARD UNIVERSITY

Table of Contents

	2 Acknowledgements
1.	**3 Introduction** 4 Executive Summary 7 How to Use these Materials 8 Data Visualization
2.	**11 Definitions and Methodology** 11 Definition of Artificial Intelligence 12 Definition of Relevant Documents 14 Document Search Methodology 15 Principle and Theme Selection Methodology 18 Timeline Visualization
3.	**20 Themes among AI Principles**
3.1	21 Privacy
3.2	28 Accountability
3.3	37 Safety and Security
3.4	41 Transparency and Explainability
3.5	47 Fairness and Non-discrimination
3.6	53 Human Control of Technology
3.7	56 Professional Responsibility
3.8	60 Promotion of Human Values
4.	**64 International Human Rights**
5.	**66 Conclusion**
6.	**68 Bibliography**

Acknowledgements

This report, *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI*, is part of the Berkman Klein Center for Internet & Society research publication series. The report and its authors have significantly benefitted from more than two years of research, capacity building, and advocacy executed by the Berkman Klein Center as co-leader of the Ethics and Governance of Artificial Intelligence Initiative in partnership with the MIT Media Lab. Through the Initiative, the Center has worked on pressing issues raised by the development and deployment of AI-based systems – including groundbreaking research on automation's impact on the online media landscape; fostering the growth of inclusive AI research and expertise in academic centers across the globe; and advocating for increased transparency surrounding the use of algorithms in criminal justice decision-making. Notably, the Center engages both private-sector firms and public policymakers in troubleshooting related legal, social, and technical challenges, and has consulted on the development of AI governance frameworks with national governments and intergovernmental bodies, such as the UN High-Level Committee on Programmes and the OECD's Expert Group on AI. *Principled Artificial Intelligence* has been influenced by these collaborative, multistakeholder efforts, and hopes to contribute to their continued success.

The report's lead author, Jessica Fjeld, is the assistant director of the Harvard Law School Cyberlaw Clinic, based at the Center; Nele Achten is an affiliate; Hannah Hilligoss a former Senior Project Coordinator and current HLS JD candidate; Adam Nagy a current Project Coordinator; and Madhulika Srikumar a former Cyberlaw Clinic student and current HLS LLM candidate. Thank you to Amar Ashar, Christopher Bavitz, Ryan Budish, Rob Faris, Urs Gasser, Vivek Krishnamurthy, Momin Malik, Jonathan Zittrain, and others in the Berkman Klein community for their generous input, support, and questions.

Thank you to Maia Levy Daniel, Joshua Feldman, Justina He, and Sally Kagay for indispensable research assistance, and to Jessica Dheere, Fanny Hidvégi, Susan Hough, K.S. Park, and Eren Sozuer for their collegial engagement on these issues and this project, as well as to all the individuals and organizations who contributed comments on the draft data visualization we released in summer 2019. An updated and final data visualization accompanies this report: thank you to Melissa Axelrod and Arushi Singh for their thoughtfulness and significant skill in its production.

1. Introduction

Alongside the rapid development of artificial intelligence (AI) technology, we have witnessed a proliferation of "principles" documents aimed at providing normative guidance regarding AI-based systems. Our desire for a way to compare these documents – and the individual principles they contain – side by side, to assess them and identify trends, and to uncover the hidden momentum in a fractured, global conversation around the future of AI, resulted in this white paper and the associated data visualization.

It is our hope that the Principled Artificial Intelligence project will be of use to policymakers, advocates, scholars, and others working on the frontlines to capture the benefits and reduce the harms of AI technology as it continues to be developed and deployed around the globe.

Executive Summary

In the past several years, seemingly every organization with a connection to technology policy has authored or endorsed a set of principles for AI. As guidelines for ethical, rights-respecting, and socially beneficial AI develop in tandem with – and as rapidly as – the underlying technology, there is an urgent need to understand them, individually and in context. To that end, we analyzed the contents of thirty-six prominent AI principles documents, and in the process, discovered thematic trends that suggest the earliest emergence of sectoral norms.

While each set of principles serves the same basic purpose, to present a vision for the governance of AI, the documents in our dataset are diverse. They vary in their intended audience, composition, scope, and depth. They come from Latin America, East and South Asia, the Middle East, North America, and Europe, and cultural differences doubtless impact their contents. Perhaps most saliently, though, they are authored by different actors: governments and intergovernmental organizations, companies, professional associations, advocacy groups, and multi-stakeholder initiatives. Civil society and multistakeholder documents may serve to set an advocacy agenda or establish a floor for ongoing discussions. National governments' principles are often presented as part of an overall national AI strategy. Many private sector principles appear intended to govern the authoring organization's internal development and use of AI technology, as well as to communicate its goals to other relevant stakeholders including customers and regulators. Given the range of variation across numerous axes, it's all the more surprising that our close study of AI principles documents revealed common themes.

The first substantial aspect of our findings are the **eight key themes** themselves:

- **Privacy.** Principles under this theme stand for the idea that AI systems should respect individuals' privacy, both in the use of data for the development of technological systems and by providing impacted people with agency over their data and decisions made with it. Privacy principles are present in 97% of documents in the dataset.
- **Accountability.** This theme includes principles concerning the importance of mechanisms to ensure that accountability for the impacts of AI systems is appropriately distributed, and that adequate remedies are provided. Accountability principles are present in 97% of documents in the dataset.
- **Safety and Security.** These principles express requirements that AI systems be safe, performing as intended, and also secure, resistant to being compromised by unauthorized parties. Safety and Security principles are present in 81% of documents in the dataset.
- **Transparency and Explainability.** Principles under this theme articulate requirements that AI systems be designed and implemented to allow for oversight, including through translation of their operations into intelligible outputs and the provision of information about where, when, and how they are being used. Transparency and Explainability principles are present in 94% of documents in the dataset.
- **Fairness and Non-discrimination.** With concerns about AI bias already impacting individuals globally, Fairness and Non-discrimination principles call for AI systems to be designed and used to maximize fairness and promote inclusivity. Fairness and Non-discrimination principles are present in 100% of documents in the dataset.
- **Human Control of Technology.** The principles under this theme require that important decisions remain subject to human review. Human Control of Technology principles are present in 69% of documents in the dataset.
- **Professional Responsibility.** These principles recognize the vital role that individuals involved in the development and deployment of AI systems play in the systems' impacts, and call on their professionalism and integrity in ensuring that the appropriate stakeholders are consulted and long-term effects are planned for. Professional Responsibility principles are present in 78% of documents in the dataset.
- **Promotion of Human Values.** Finally, Human Values principles state that the ends to which AI is devoted, and the means by which it is implemented, should correspond with our core values and generally promote humanity's well-being. Promotion of Human Values principles are present in 69% of documents in the dataset.

The second, and perhaps even more striking, side of our findings is **that more recent documents tend to cover all eight of these themes**, suggesting that the conversation around principled AI is beginning to converge, at least among the communities responsible for the development of these documents. Thus, these themes may represent the "normative core" of a principle-based approach to AI ethics and governance.[1]

However, we caution readers against inferring that, in any individual principles document, broader coverage of the key themes is necessarily better. Context matters. Principles should be understood in their cultural, linguistic, geographic, and organizational context, and some themes will be more relevant to a particular context and audience than others. Moreover, principles are a starting place for governance, not an end. On its own, a set of principles is unlikely to be more than gently persuasive. Its impact is likely to depend on how it is embedded in a larger governance ecosystem, including for instance relevant policies (e.g. AI national plans), laws, regulations, but also professional practices and everyday routines.

[1] Both aspects of our findings are observable in the data visualization (p. 8-9) that accompanies this paper.

One existing governance regime with significant potential relevance to the impacts of AI systems is international human rights law. Scholars, advocates, and professionals have increasingly been attentive to the connection between AI governance and human rights laws and norms,[2] and we observed the impacts of this attention among the principles documents we studied. 64% of our documents contained a reference to human rights, and five documents took international human rights as a framework for their overall effort. Existing mechanisms for the interpretation and protection of human rights may well provide useful input as principles documents are brought to bear on individuals cases and decisions, which will require precise adjudication of standards like "privacy" and "fairness," as well as solutions for complex situations in which separate principles within a single document are in tension with one another.

The thirty-six documents in the *Principled Artificial Intelligence* were curated for variety, with a focus on documents that have been especially visible or influential. As noted above, a range of sectors, geographies, and approaches are represented. Given our subjective sampling method and the fact that the field of ethical and rights-respecting AI is still very much emergent, we expect that perspectives will continue to evolve beyond those reflected here. We hope that this paper and the data visualization that accompanies it can be a resource to advance the conversation on ethical and rights-respecting AI.

[2] Hannah Hilligoss, Filippo A. Raso and Vivek Krishnamurthy, 'It's not enough for AI to be "ethical"; it must also be "rights respecting"', Berkman Klein Center Collection (October 2018) https://medium.com/berkman-klein-center/its-not-enough-for-ai-to-be-ethical-it-must-also-be-rights-respecting-b87f7e215b97.

How to Use these Materials

Data Visualization
The Principled AI visualization, designed by Arushi Singh and Melissa Axelrod, is arranged like a wheel. Each document is represented by a spoke of that wheel, and labeled with the sponsoring actors, date, and place of origin. The one exception is that the OECD and G20 documents are represented together on a single spoke, since the text of the principles in these two documents is identical.[3] The spokes are sorted first alphabetically by the actor type and then by date, from earliest to most recent.

Inside the wheel are nine rings, which represent the eight themes and the extent to which each document makes reference to human rights. In the theme rings, the dot at the intersection with each spoke indicates the percentage of principles falling under the theme that the document addresses: the larger the dot, the broader the coverage. Because each theme contains different numbers of principles (ranging from three to ten), it's instructive to compare circle size within a given theme, but not between then.

In the human rights ring, a diamond indicates that the document references human rights or related international instruments, and a star indicates that the document uses international human rights law as an overall framework.

[3] Note that while the OECD and G20 principles documents share a single spoke on the data visualization, for purposes of the quantitative analysis underlying this paper, they have been counted as separate documents.

PRINCIPLED ARTIFICIAL INTELLIGENCE

A Map of Ethical and Rights-Based Approaches to Principles for AI

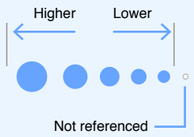
Authors: Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, Madhulika Srikumar

Designers: Arushi Singh (arushisingh.net) and Melissa Axelrod (melissaaxelrod.com)

HOW TO READ:

Date, Location
Document Title
Actor

COVERAGE OF THEMES:



References International Human Rights

Explicitly Adopts Human Rights Framework

The size of each dot represents the percentage of principles in that theme contained in the document. Since the number of principles per theme varies, it's informative to compare dot sizes within a theme but not between themes.

The principles within each theme are:

Privacy:
Privacy
Control over Use of Data
Consent
Privacy by Design
Recommendation for Data Protection Laws
Ability to Restrict Processing
Right to Rectification
Right to Erasure

Accountability:
Accountability
Recommendation for New Regulations
Impact Assessment
Evaluation and Auditing Requirement
Verifiability and Replicability
Liability and Legal Responsibility
Ability to Appeal
Environmental Responsibility
Creation of a Monitoring Body
Remedy for Automated Decision

Safety and Security:
Security
Safety and Reliability
Predictability
Security by Design

Transparency and Explainability:
Explainability
Transparency
Open Source Data and Algorithms
Notification when Interacting with an AI
Notification when AI Makes a Decision about an Individual
Regular Reporting Requirement
Right to Information
Open Procurement (for Government)

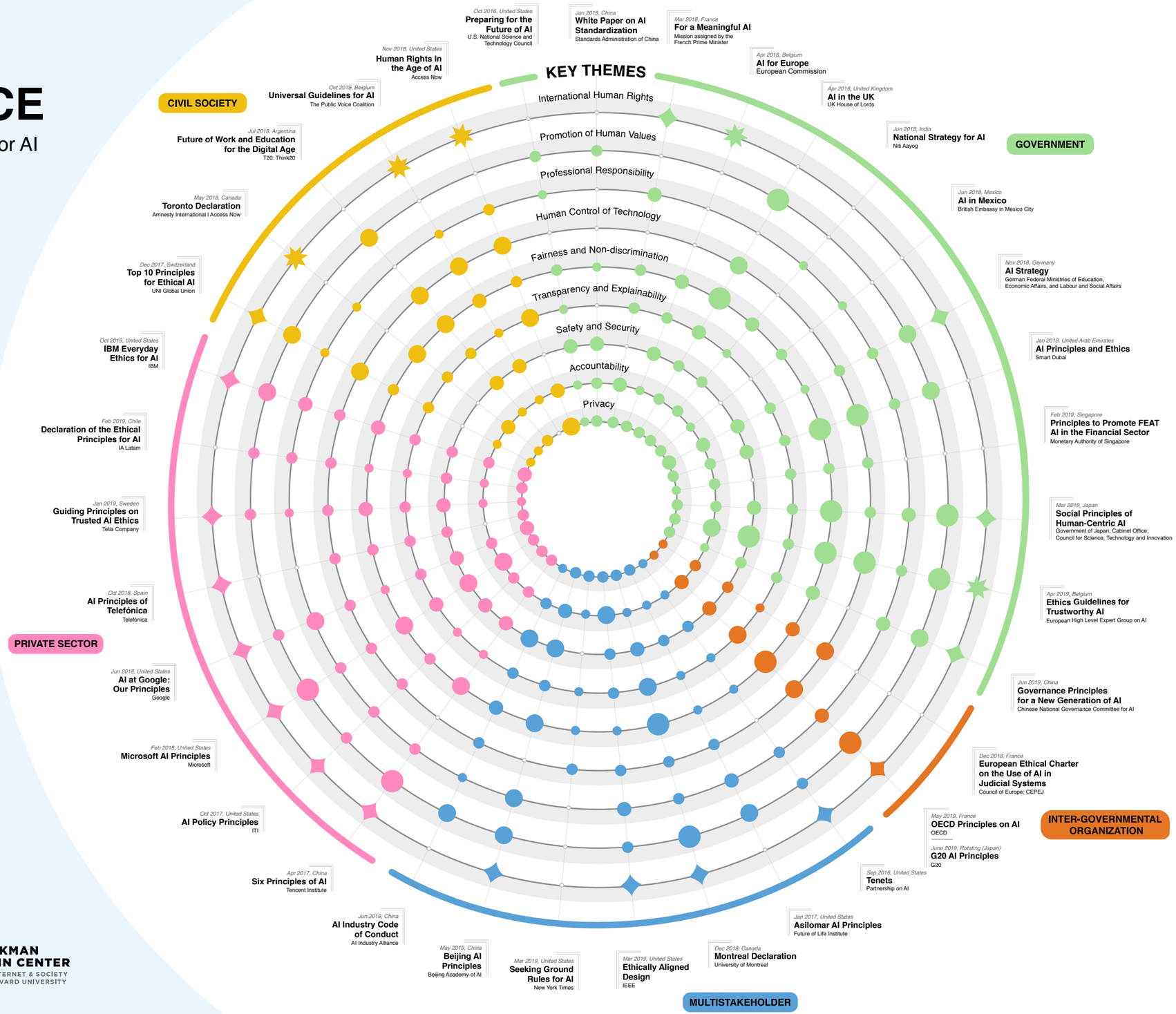
Fairness and Non-discrimination:
Non-discrimination and the Prevention of Bias
Fairness
Inclusiveness in Design
Inclusiveness in Impact
Representative and High Quality Data
Equality

Human Control of Technology:
Human Control of Technology
Human Review of Automated Decision
Ability to Opt out of Automated Decision

Professional Responsibility:
Multistakeholder Collaboration
Responsible Design
Consideration of Long Term Effects
Accuracy
Scientific Integrity

Promotion of Human Values:
Leveraged to Benefit Society
Human Values and Human Flourishing
Access to Technology

BERKMAN KLEIN CENTER
FOR INTERNET & SOCIETY
AT HARVARD UNIVERSITY



White Paper
Much as the principles documents underlying our research come from a wide variety of stakeholders in the ongoing conversation around ethical and rights-respecting AI, so too we expect a variety of readers for these materials. It is our hope that they will be useful to policymakers, academics, advocates, and technical experts. However, different groups may wish to engage with the white paper in different ways:

- Those looking for a **high-level snapshot of the current state of thinking in the governance of AI** may be best served by reviewing the data visualization (p. 8), and reading the Executive Summary (p. 4) and Human Rights section (p. 64), dipping into the discussion of themes (beginning p. 20) only where they are necessary to resolve a particular interest or question.
- Those looking to do **further research** on AI principles will likely find the discussions of the themes and principles (beginning p. 20) and Human Rights section (p. 64) most useful, and are also invited to contact the authors with requests to access the underlying data.
- Those tasked with **drafting a new set of principles** may find that the data visualization (p. 8) and discussions of the themes and principles within them (beginning p. 20) can function to offer a head start on content and approaches thereto, particularly as references to existing principles that are most likely to be useful source material.
- Those seeking closer **engagement with primary source documents** may variously find the data visualization (p. 8), timeline (p. 18), or bibliography (p. 68) to act as a helpful index.

2. Definitions and Methodology

Definition of Artificial Intelligence

The definition of artificial intelligence, or "AI", has been widely debated over the years, in part because the definition changes as technology advances.[4] In collecting our dataset, we did not exclude documents based on any particular definition of AI. Rather, we included documents that refer specifically to AI or a closely equivalent term (for example, IEEE uses "autonomous and intelligent systems").[5] In keeping with the descriptive approach we have taken in this paper, we'll share a few definitions found in our dataset. The European Commission's High-Level Expert Group on Artificial Intelligence offers a good place to start:

"Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions."[6]

Aspects of this definition are echoed in those found in other documents. For example, some documents define AI as systems that take action, with autonomy, to achieve a predefined goal, and some add that these actions are generally tasks that would otherwise require human intelligence.[7]

---

[4] This is known as the "odd paradox" – when technologies lose their classification as "AI" because more impressive technologies take their place. *See*, Pamela McCorduck, 'Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence', 2nd ed. (Natick, MA: A. K. Peters, Ltd., 2004).

[5] IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 'Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems' (2019) First Edition <https://ethicsinaction.ieee.org/>.

[6] European Commission's High-Level Expert Group on Artificial Intelligence, 'Ethics Guidelines for Trustworthy AI' (2019) p. 36 <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>.

[7] UK House of Lords, Select Committee on Artificial Intelligence, 'AI in the UK: Ready, Willing and Able?' (2018) Report of Session 2017-19 <https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf>; Mission assigned by the French Prime Minister, 'For a Meaningful Artificial Intelligence: Toward a French and European Strategy' (2018) <https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf>..

Other documents define AI by the types of tasks AI systems accomplish – like "learning, reasoning, adapting, and performing tasks in ways inspired by the human mind,"[8] or by its sub-fields like knowledge-based systems, robotics, or machine learning.[9]

Definition of Relevant Documents

While all of the documents use the term "AI" or an equivalent, not all use the term "principles," and delineating which documents on the subject of ethical or rights-respecting AI should be considered "principles" documents was a significant challenge. Our working definition was that principles are normative (in the sense that lawyers use this term) declarations about how AI generally *ought* to be developed, deployed, and governed. While the intended audience of our principles documents varies, they all endeavor to shape behavior of an audience - whether internal company principles to follow in AI development or broadly targeted principles meant to further develop societal norms about AI.

Because a number of documents employed terminology other than "principles" while otherwise conforming to this definition, we included them.[10] The concept of "ethical principles" for AI has encountered pushback both from ethicists, some of whom object to the imprecise usage of the term in this context, as well as from some human rights practitioners, who resist the recasting of fundamental human rights in this language. Rather than disaggregate AI principles from the other structures (international human rights, domestic or regional regulations, professional norms) in which they are intertwined, our research team took pains to assess principles documents in context and to flag external frameworks where relevant. In doing so, we drew inspiration from the work of Urs Gasser, Executive Director of the Berkman Klein Center for Internet & Society and Professor of Practice at Harvard Law School, whose theory on "digital constitutionalism" describes the significant role the articulation of principles by a diverse set of actors might play as part of the "proto-constitutional discourse" that leads to the crystallization of comprehensive governance norms.

Our definition of principles excluded documents that were time-bound in the sense of observations about advances made in a particular year[11] or goals to be accomplished over a particular period. It also excluded descriptive statements about AI's risks and benefits. For example, there are numerous compelling reports that assess or comment on the ethical implications of AI, some even containing recommendations for next steps, that don't advance a particular set of principles[12] and were thus excluded from this dataset. However, where a report included a recommendations section which did correspond to our definition, we included that section (but not the rest of the report) in our dataset,[13] and more generally, when only a certain page range from a broader document conformed to our definition, we limited our sample to those pages. The result of these choices is a narrower set of documents that we hope lends itself to side-by-side comparison, but notably excludes some significant literature.

We also excluded documents that were formulated solely as calls to a discrete further action, for example that that funding be committed, new agencies established, or additional research done on a particular topic, because they function more as a policy objective than a principle. By this same logic, we excluded national AI strategy documents that call for the creation of principles without advancing any.[14] However, where documents otherwise met our definition but contained individual principles such as calls for further research or regulation of AI (under the Accountability theme, see Section 3.2), we did include them. We also included the principle that those building and implementing AI should routinely consider the long-term effects of their work (under Professional Responsibility, see Section 3.7). Rather than constitute a discrete task, this call for further consideration functions as a principle in that it advocates that a process of reflection be built into the development of any AI system.

Finally, we excluded certain early instances of legislation or regulation which closely correspond to our definition of principles.[15] The process underlying the passage of governing law is markedly different than the one which resulted in other principles documents we were considering, and we were conscious of the fact that the goal of this project was to facilitate side-by-side comparison, and wanted to select documents that could fairly be evaluated that way. For the same reason, we excluded documents that looked at only a specific type of technology, such as facial recognition. We found that the content of principles documents was strongly affected by restrictions of technology type, and thus side-by-side comparison of these documents with others that focused on AI generally was unlikely to be maximally useful. On the other hand, we included principles documents that are sector-specific, focusing for example on the impacts of AI on the workforce or criminal justice, because they were typically similar in scope to the general documents.

---

[8] Information Technology Industry Council, 'AI Policy Principles' (2017) <https://www.itic.org/resources/AI-Policy-Principles-FullReport2.pdf>.

[9] German Federal Ministry of Education and Research, the Federal Ministry for Economic Affairs and Energy, and the Federal Ministry of Labour and Social Affairs, 'Artificial Intelligence Strategy' (2018) <https://www.ki-strategie-deutschland.de/home.html>; Access Now, 'Human Rights in the Age of Artificial Intelligence' (2018) <https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf>.

[10] For example, the Partnership on AI's document is the "Tenets," the Public Voice and European High Level Expert Group's documents are styled as "guidelines," the Chinese AI Industry's document is a "Code of Conduct" and the Toronto Declaration refers to "responsibilities" in Principle 8.

[11] AI Now Institute, New York University, 'AI Now Report 2018' (December 2018) https://ainowinstitute.org/AI_Now_2018_Report.pdf.

[12] AI Now Institute, New York University, 'AI Now Report 2018' (December 2018) https://ainowinstitute.org/AI_Now_2018_Report.pdf.

[13] *See generally*, Access Now (n 9).

[14] For example, in 2017 the government of Finland published *Finland's Age of Artificial Intelligence*, which was excluded from our dataset because it does not include principles for socially beneficial AI. *See*, Ministry of Economic Affairs and Employment of Finland, 'Finland's Age of Artificial Intelligence' (2017) http://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/160391/TEMrap_47_2017_verkkojulkaisu.pdf?sequence=1&isAllowed=y

[15] *See*, Treasury Board of Canada Secretariat, 'Directive on Automated Decision-Making' (Feb. 2019) https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592

Due to the flexibility of our definition, there remains a broad range among the documents we did include, from high-level and abstract statements of values, to more narrowly focused technical and policy recommendations. While we questioned whether this should cause us to narrow our focus still further, because the ultimate goal of this project is to provide a description of the current state of the field, we decided to retain the full range of principle types we observed in the dataset, and encourage others to dive deeper into particular categories according to their interests.

Document Search Methodology

The dataset of thirty-six documents on which this report and the associated data visualization are based was assembled using a purposive sampling method. Because a key aim of the project from the start was to create a data visualization that would facilitate side by side comparison of individual documents, it was important that the dataset be manageably sized, and also that it represent a diversity of viewpoints in terms of stakeholder, content, geography, date, and more. We also wanted to ensure that widely influential documents were well represented. For this reason, we determined that purposive sampling with the goal of maximum variation among influential documents in this very much emergent field was the most appropriate strategy.[16]

Our research process included a wide range of tools and search terms. To identify eligible documents, our team used a variety of search engines, citations from works in the field, and expertise and personal recommendations from others in the Berkman Klein Center community. Because the principles documents are not academic publications, we did not make extensive use of academic databases. General search terms included a combination of "AI" or "artificial intelligence" and "principles," "recommendations," "strategy," "guideline," and "declaration," amongst others. We also used knowledge from our community to generate the names of organizations – companies, governments, civil society actors, etc. – might have principles documents, and then we then searched those organizations' websites and publications.

In order to ensure that each document earned its valuable real estate in our visualization, we required that it represent the views of an organization or institution; be authored by relatively senior staff; and, in cases of multistakeholder documents, contain a breadth of involved experts. It is worth noting that some government documents are expert reports commissioned by governments rather than the work of civil servants, but all documents included in this category were officially published.

Our search methodology has some limitations. Due to the language limitations of our team, our dataset only contains documents available in English, Chinese, French, German, and Spanish. While we strove for broad geographical representation, we were unable to locate any documents from the continent of Africa, although we understand that certain African states may be currently engaged in producing AI national strategy documents which may include some form of principles. Furthermore, we recognize the possibility of network bias – because these principles documents are often shared through newsletters or mailing lists, we discovered some documents through word of mouth from those in our network. That being said, we do not purport to have a complete dataset, an admirable task which has been taken up by others.[17] Rather we have put together a selection of prominent principles documents from an array of actors.

Principle and Theme Selection Methodology

As principles documents were identified, they were reviewed in team meetings for conformity with our criteria. Those that met the criteria were assigned to an individual team member for hand coding. That team member identified the relevant pages of the document, in the case that the principles formed a sub-section of a longer document, and hand-coded all text in that section. In the initial phase, team members were actively generating the principle codes that form the basis of our database. They used the title of the principle in the document, or if no title was given or the title did not thoroughly capture the principle's content, paraphrased the content of the principle. If an identical principle had already been entered into the database, the researcher coded the new document under that principle rather than entering a duplicate.

When the team had collected and coded approximately twenty documents, we collated the list of principles, merging close equivalents, to form a final list of forty-seven principles. We then clustered the principles, identifying ones that were closely related both in terms of their dictionary meanings (e.g. fairness and non-discrimination) as well as ones that were closely linked in the principles documents themselves (e.g. transparency and explainability). We arrived at eight total themes, each with between three and ten principles under it:

- Privacy (8 principles)
- Accountability (10 principles)
- Safety and security (4 principles)
- Transparency and explainability (8 principles)
- Fairness and non-discrimination (6 principles)
- Human control of technology (3 principles)
- Professional responsibility (5 principles)
- Promotion of human values (3 principles)

[16] For background on purposive sampling, *See* Patton, M. Q., "Qualitative evaluation and research methods" (1990) (2nd ed.). Newbury Park, CA: Sage Publications.

[17] Anna Jobin, Marcello Ienca and Effy Vayena, "The global landscape of AI ethics guidelines", Nature Machine Intelligence (September 2019) https://doi.org/10.1038/s42256-019-0088-2; https://algorithmwatch.org/en/project/ai-ethics-guidelines-global-inventory/

We also collected data on references to human rights in each document, whether to human rights as a general concept or to specific legal instruments such as the UDHR or the ICCPR. While this data is structured similarly to the principles and themes, with individual references coded under the heading of International Human Rights, because the references appear in different contexts in different documents and we do not capture that in our coding, we do not regard it as a theme in the same way that the foregoing concepts are. See Section 4 for our observations of how the documents in our dataset engage with human rights.

Both the selection of principles that would be included in the dataset and the collation of those principles into themes were subjective, though strongly informed by content of the early documents in our dataset and the researchers' immersion in them. This has led to some frustrations about their content. For example, when we released the draft data visualization for feedback, we were frequently asked why sustainability and environmental responsibility did not appear more prominently. While the authors are sensitive to the significant impact AI is having, and will have, on the environment,[18] we did not find a concentration of related concepts in this area that would rise to the level of a theme, and as such have included the principle of "environmental responsibility" under the Accountability theme as well as discussion of AI's environmental impacts in the "leveraged to benefit society" principle under the Promotion of Human Values theme. It may be that as the conversation around AI principles continues to evolve, sustainability becomes a more prominent theme.

Following the establishment of the basic structure of principles and themes, we were conservative in the changes we made because work on the data visualization, which depended on their consistency, was already underway. We did refine the language of the principles in the dataset, for example from "Right to Appeal" to "Ability to Appeal," when many of the documents that referenced an appeal mechanism did not articulate it as a user's right. We also moved a small number of principles from one theme to another when further analysis of their contents demanded; the most prominent example of this is that "Predictability," which was included under the Accountability theme at the time our draft visualization was released in summer 2019, has been moved to the Safety and Security theme.

Because the production of the data visualization required us to minimize the number of these changes, and because our early document collection (on which the principles and themes were originally based) was biased toward documents from the U.S. and E.U., there are a small number of principles from documents – predominantly non-Western documents – that do not fit comfortably into our dataset. For example, the Japanese AI principles include a principle of fair competition which combines intranational competition law with a caution that "[e]ven if resources related to AI are concentrated in a specific country, we must not have a society where unfair data collection and infringement of sovereignty are performed under that country's dominant position."[19] We have coded this language within the "access to technology" principle under the Promotion of Human Values theme, but it does push at the edges of our definition of that principle, and is imperfectly captured by it. Had this document been part of our initial sample, its contents might have resulted in our adding to or changing the forty-seven principles we ultimately settled on.

We therefore want to remind our readers that this is a fundamentally partial and subjective approach. We view the principles and themes we have advanced herein as simply one heuristic through which to approach AI principles documents and understand their content. Other people could have made, and will make in future, other choices about which principles to include and how to group them.

[18] Roel Dobbe and Meredith Whittaker, 'AI and Climate Change: How they're connected, and what we can do about it', AI Now Institute (2019). Retrieved from https://medium.com/@ainowinstitute.

[19] This is Principle 4.1.5. Principle of Fair Competition in Japanese Cabinet Office, Council for Science, Technology and Innovation, 'Social Principles of Human-Centric Artificial Intelligence' (2019) <https://www8.cao.go.jp/cstp/english/humancentricai.pdf>.

PRINCIPLED ARTIFICIAL INTELLIGENCE

A Map of Ethical and Rights-Based Approaches to Principles for AI

DOCUMENT TIMELINE

Nature of Actors

- Civil Society
- Government
- Inter-governmental Organization
- Multistakeholder
- Private Sector

BERKMAN KLEIN CENTER
FOR INTERNET & SOCIETY
AT HARVARD UNIVERSITY

3. Themes among AI Principles

This section describes in detail our findings with respect to the eight themes, as well as the principles they contain:

- Privacy
- Accountability
- Safety and security
- Transparency and explainability
- Fairness and non-discrimination
- Human control of technology
- Professional responsibility
- Promotion of human values

Coverage of each theme offers a view into its core features, relevance, and connection to other themes and principles. Further, we offer a detailed look at the principles under each theme, including insights generated by comparing how the principles were variously framed by the documents in our dataset.

3.1. Privacy

Privacy – enshrined in international human rights law and strengthened by a robust web of national and regional data protection laws and jurisprudence – is significantly impacted by AI technology. Fueled by vast amounts of data, AI is used in surveillance, advertising, healthcare decision-making, and a multitude of other sensitive contexts. Privacy is not only implicated in prominent implementations of AI, but also behind the scenes, in the development and training of these systems.[20] Consequently, privacy is a prominent theme[21] across the documents in our dataset, consisting of eight principles: "consent," "control over the use of data," "ability to restrict data processing," "right to rectification," "right to erasure," "privacy by design," "recommends data protection laws," and "privacy (other/general)."

The General Data Protection Regulation of the European Union (GDPR) has been enormously influential in establishing safeguards for personal data protection in the current technological environment, and many of the documents in our dataset were clearly drafted with provisions of the GDPR in mind. We also see strong connections between principles under the Privacy theme and the themes of Fairness and Non-Discrimination, Safety and Security, and Professional Responsibility.

PRINCIPLES UNDER THIS THEME

- **25%** Consent
- **42%** Control over the Use of Data
- **11%** Ability to Restrict Processing
- **8%** Right to Rectification
- **6%** Right to Erasure
- **17%** Privacy by Design
- **17%** Recommends Data Protection Laws
- **92%** Privacy (Other/General)

Percentage reflects the number of documents in the dataset that include each principle

[20] Mission assigned by the French Prime Minister (n 8) p. 114 ("Yet it appears that current legislation, which focuses on the protection of the individual, is not consistent with the logic introduced by these systems [AI]—i.e. the analysis of a considerable quantity of information for the purpose of identifying hidden trends and behavior—and their effect on groups of individuals. To bridge this gap, we need to create collective rights concerning data.") .

[21] Privacy principles are present in 97% of documents in the dataset. All of the principles written by government, private, and multistakeholder groups reference principles under the Privacy theme. Among documents sourced from civil society, only one, the Public Voice Coalition AI guidelines, did not refer to privacy.

Consent

Broadly, "consent" principles reference the notion that a person's data should not be used without their knowledge and permission. Informed consent is a closely related but more robust principle – derived from the medical field – which requires individuals be informed of risks, benefits, and alternatives. Arguably, some formulation of "consent" is a necessary component of a full realization of other principles under the Privacy theme, including "ability to restrict processing," "right to rectification," "right to erasure," and "control over the use of data."

Documents vary with respect to the depth of their description of consent, breaking into two basic categories: documents that touch lightly on it, perhaps outlining a simple notice-and-consent regime,[22] and documents that invoke informed consent specifically or even expand upon it.[23] A few documents, such as Google's AI principles and IA Latam's principles, do not go beyond defining consent as permission, but as a general matter, informed consent or otherwise non-perfunctory processes to obtain consent feature prominently in the corpus.

The boldest departures from the standard notice-and-consent model can be found in the Chinese White Paper on AI Standardization and Indian AI strategy. The Chinese document states that "the acquisition and informed consent of personal data in the context of AI should be redefined" and, among other recommendations, states "we should begin regulating the use of AI which could possibly be used to derive information which exceeds what citizens initially consented to be disclosed."[24] The Indian national strategy cautions against unknowing consent and recommends a mass-education and awareness campaign as a necessary component of implementing a consent principle in India.[25]

Control over the Use of Data

"Control over the use of data" as a principle stands for the notion that data subjects should have some degree of influence over how and why information about them is used. Certain other principles under the privacy theme, including "consent," "ability to restrict processing," "right to rectification," and "right to erasure" can be thought of as more specific instantiations of the control principle since they are mechanisms by which a data subject might exert control. Perhaps because this principle functions as a higher-level articulation, many of the documents we coded under it are light in the way of definitions for "control."

Generally, the documents in our dataset are of the perspective that an individual's ability to determine how their data is used and for what purpose should be qualified in various ways. Microsoft commits to giving consumers "*appropriate* controls so they can choose how their data is used"[26] and IEEE notes that where minors and those with diminished capacity are concerned, recourse to guardianship arrangements may be required.[27] However, several documents do contain articulations of the control principle that are more absolute. The IBM AI principles state that "Users should *always* maintain control over what data is being used and in what context."[28] On the other hand, the German AI strategy clearly states the importance of balancing and repeatedly articulates people's control over their personal data as a qualified "right." The German document suggests the use of "pseudonymized and anonymized data" as potential tools to "help strike the right balance between protecting people's right to control their personal data and harnessing the economic potential of big-data applications."[29]

There is some differentiation between the documents on the question of where control ought to reside. Some dedicate it to individuals, which is typical of current systems for data control. On the other hand, some documents would locate control in specially dedicated tools, institutions, or systems. For example, the European Commission's High-Level Expert Group describes the creation of "data protocols" and "duly qualified personnel" who would govern access to data.[30] IEEE proposes the implementation of a technology that would allow individuals to assign "an online agent" to help make "case-by-case authorization decisions as to who can process what personal data for what purpose." This technology might even be a dynamically learning AI itself – evaluating data use requests by third parties in an "autonomous and intelligent" manner.[31] Lastly, AI in the UK advocates "data trusts" that would allow individuals to "make their views heard and shape … decisions" through some combination of consultative procedures, "personal data representatives," or other mechanisms.[32]

Ability to Restrict Processing

The "ability to restrict processing" refers to the power of data subjects to have their data restricted from use in connection with AI technology. Some documents coded for this principle articulate this power as a legally enforceable right, while others stop short of doing so. For example, the Access Now report would "give people the ability to *request* that an entity stop using or limit the use of personal information."[33] Notably, Article 18 of the GDPR has legally codified this right with respect to data processing more generally, but documents within our dataset diverge in some respects from the GDPR definition.

The extent to which data subjects should be able to restrict the processing of their data is clearly in contention. For instance, the Montreal Declaration asserts that people have a "right to digital disconnection" and imposes a positive obligation

---

[22] *See generally* German Federal Ministry of Education and Research, the Federal Ministry for Economic Affairs and Energy, and the Federal Ministry of Labour and Social Affairs (n 9).German Federal Ministry of Education and Research, the Federal Ministry for Economic Affairs and Energy, and the Federal Ministry of Labour and Social Affairs (n 10); Google, 'AI at Google: Our Principles' (2018) <https://www.blog.google/technology/ai/ai-principles/>; Smart Dubai, 'Artificial Intelligence Principles and Ethics' (2019) <https://smartdubai.ae/initiatives/ai-principles-ethics> ;IA Latam, 'Declaración de Principios Éticos Para La IA de Latinoamérica' (2019) <http://ia-latam.com/etica-ia-latam/>; Chinese National Governance Committee for the New Generation Artificial Intelligence, led by China's Ministry of Science and Technology, 'Governance Principles for a New Generation of Artificial Intelligence: Develop Responsible Artificial Intelligence' (2019) <http://www.chinadaily.com.cn/a/201906/17/WS5d07486ba3103dbf14328ab7.html>.

[23] *See generally* Standard Administration of China and Paul Triolo, 'White Paper on Artificial Intelligence Standardization' excerpts in English published by New America (January 2018) <https://www.newamerica.org/cybersecurity-initiative/digichina/blog/translation-excerpts-chinas-white-paper-artificial-intelligence-standardization/>; Beijing Academy of Artificial Intelligence, 'Beijing AI Principles' (2019) (English translation available upon request) <https://www.baai.ac.cn/blog/beijing-ai-principles?categoryId=394>; Niti Aayog, 'National Strategy for Artificial Intelligence: #AI for All (Discussion Paper)' (2018) <https://www.niti.gov.in/writereaddata/files/document_publication/NationalStrategy-for-AI-Discussion-Paper.pdf>; IBM, 'IBM Everyday Ethics for AI' (2019) <https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf>.

[24] Standard Administration of China and Triolo (n 24) Principle 3.3.3.

[25] Niti Aayog (n 24) p. 88.

[26] Microsoft, 'AI Principles' (2018) p. 68 <https://www.microsoft.com/en-us/ai/our-approach-to-ai> (emphasis added).

[27] IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (n 5) p. 23

[28] IBM (n 24) p. 44.

[29] German Federal Ministry of Education and Research, the Federal Ministry for Economic Affairs and Energy, and the Federal Ministry of Labour and Social Affairs (n 10) pp. 8, 16, 18, 28.

[30] European Commission's High-Level Expert Group on Artificial Intelligence (n 6) p. 17.

[31] IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (n 5) p. 23.

[32] UK House of Lords, Select Committee on Artificial Intelligence (n 8) p. 126.

[33] Access Now (n 10) p. 31.

on AI-driven systems to "explicitly offer the option to disconnect at regular intervals, without encouraging people to stay connected,"[34] and an earlier draft of the European High Level Expert Group guidelines placed a positive obligation on government data controllers to "systematically" offer an "express opt-out" to citizens.[35] However, the final version of the HLEG guidelines was far less expansive, narrowing the right to opt-out to "citizen scoring" technologies in "circumstances where … necessary to ensure compliance with fundamental rights."[36]

Right to Rectification

The "right to rectification" refers to the right of data subjects to amend or modify information held by a data controller if it is incorrect or incomplete. As elsewhere where the word "right" is contained in the title, we only coded documents under this principle where they explicitly articulated it as a right or obligation. High-quality data contributes to safety, fairness, and accuracy in AI systems, so this principle is closely related to the themes of Fairness and Non-Discrimination and Safety and Security. Further, the "right to rectification" is closely related to the "ability to restrict processing," insofar as they are both part of a continuum of potential responses a data subject might have in response to incorrect or incomplete information.

Rectification is not a frequently invoked principle, appearing in only three documents within our dataset. The Access Now report recommends a right to rectification closely modeled after that contained in Article 16 of the GDPR. The Singapore Monetary Authority's AI principles place a positive obligation on firms to provide data subjects with "online data management tools" that enable individuals to review, update, and edit information for accuracy.[37] Finally, the T20 report on the future of work and education addresses this principle from a sector-specific viewpoint, describing a right held by employees and job applicants to "have access to the data held on them in the workplace and/or have means to ensure that the data is accurate and can be rectified, blocked, or erased if it is inaccurate."[38]

Right to Erasure

The "right to erasure" refers to an enforceable right of data subjects to the removal of their personal data. Article 17 of the GDPR also contains a right to erasure, which allows data subjects to request the removal of personal data under a defined set of circumstances, and provides that the request should be evaluated by balancing rights and interests of the data holder, general public, or other relevant parties. The Access Now report models its recommendation off of Article 17, stating:

> [T]he Right to Erasure provides a pathway for deletion of a person's personal data held by a third party entity when it is no longer necessary, the information has been misused, or the relationship between the user and the entity is terminated.[39]

However, other documents in the dataset advance a notion of the right to erasure distinct from the GDPR. Both the Chinese AI governance principles and the Beijing AI Principles include a call for "revocation mechanisms."[40] In contrast to the Access Now articulation, the Beijing AI Principles provide for access to revocation mechanisms in "unexpected circumstances."[41] Further, the Beijing document conditions that the data and service revocation mechanism must be "reasonable" and that practices should be in place to ensure the protection of users' rights and interests. The version of the erasure principle in the T20 report on the future of work and education is even more narrowly tailored, and articulates a right to erasure for data on past, present, and potential employees held by employers if it is inaccurate or otherwise violates the right to privacy.[42]

Privacy by Design

"Privacy by design," also known as data protection by design, is an obligation on AI developers and operators to integrate considerations of data privacy into the construction of an AI system and the overall lifecycle of the data. Privacy by design is codified in Article 25 of the GDPR, which stipulates data controllers must "implement appropriate technical and organisational measures..." during the design and implementation stage of data processing "to protect the rights of data subjects."[43] Perhaps in recognition of these recent regulatory advances, IBM simply commits to adhering to national and international rights laws during the design of an AI's data access permissions.[44]

In the private sector, privacy by design is regarded as an industry best practice, and it is under these terms that Google and Telefónica consider the principle. Google's AI principles document does not use the phrase "privacy by design" but it does commit the company to incorporate Google's privacy principles into the development and use of AI technologies and to "encourage architectures with privacy safeguards."[45] Telefónica also points to its privacy policy and methodologies, stating: "In order to ensure compliance with our Privacy Policy we use a Privacy by Design methodology. When building AI systems, as with other systems, we follow Telefónica's Security by Design approach." ITI goes a step further, committing to "ethics by design," a phrase that can be best understood as the integration of principles into the design of AI systems in a manner beyond what is legally required, and connects strongly with the "responsible design" principle under the Professional Responsibility theme.

Recommends Data Protection Laws

The "recommends data protection laws" principle, simply put, is that new government regulation is a necessary component of protecting privacy in the face of AI technologies. Documents produced on behalf of the governments of France, Germany,

[34] University of Montreal, 'Montreal Declaration for a Responsible Development of Artificial Intelligence' (2018) p. 10 (*See* Principle 3.3) <https://www.montrealdeclaration-responsibleai.com/the-declaration>.

[35] Draft European Commission's High-Level Expert Group on Artificial Intelligence, 'Ethics Guidelines for Trustworthy AI' (Dec. 2018) p. 7 (*See* Principle 3.5 Citizens rights) <https://ec.europa.eu/digital-single-market/en/news/draft-ethics-guidelines-trustworthy-ai >. <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>.

[36] European Commission's High-Level Expert Group on Artificial Intelligence (n 6) p. 34.

[37] Monetary Authority of Singapore, 'Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore's Financial Sector' (2019) p. 11 <http://www.mas.gov.sg/~/media/MAS/News%20and%20Publications/Monographs%20and%20Information%20Papers/FEAT%20Principles%20Final.pdf>.

[38] Think 20, 'Future of Work and Education for the Digital Age' (2018) p. 5 <https://www.g20-insights.org/wp-content/uploads/2018/07/TF1-1-11-Policy-Briefs_T20ARG_Towards-a-G20-Framework-For-Artificial-Intelligence-in-the-Workplace.pdf>.

[39] Access Now (n 10) p. 31.

[40] Chinese National Governance Committee for the New Generation Artificial Intelligence, led by China's Ministry of Science and Technology (n 22) Principle 4.

[41] Beijing Academy of Artificial Intelligence (n 23) (*See* Principle 2.2, English translation available upon request.)

[42] Think 20 (n 39) p. 5.

[43] GDPR Art. 25 The GDPR definition and enforcement mechanism is an instructive example of privacy by design and Article 25 even specifies techniques, such as pseudonymization and data minimization, for data processors to implement.

[44] IBM (n 24) p. 44.

[45] Google (n 23) (*See* Principle 5.)

Mexico, and India each call for the development of new data privacy and data protection frameworks. These calls for regulation tend to be aspirational in their framing, with a common acknowledgement – neatly articulated in the Access Now report – that "data protection legislation can anticipate and mitigate many of the human rights risks posed by AI."[46] Other documents add that the "diverse and fast changing nature of the technology" requires a "continually updated" privacy protection regime.[47] The importance of agile regulatory frameworks is reiterated in the AI in Mexico document, which advises Mexico's National Institute for Transparency, Access to Information and Protection of Personal Data "to keep pace with innovation."[48]

The European documents that address this principle do so in the context of an already highly protective regime. The German strategy document suggests that there exists a gap in that regime, and calls for a new Workers' Data Protection Act "that would protect employees' data in the age of AI."[49] This narrow approach contrasts with the French strategy document, which critiques current legislation, and the rights framework more fundamentally, as too focused on "the protection of the individual" to adequately contend with the potential collective harms machine learning and AI systems can perpetuate. The French document calls for the creation of new "collective rights concerning data."[50] Even outside of Europe, the GDPR's influence is felt where the Indian AI strategy points towards existing practice in Europe – specifically, the GDPR and France's right to explanation for administrative algorithmic decisions – as a standard for Indian regulators to use as potential benchmarks.[51] Like the German AI strategy, the Indian AI strategy recommends establishing sector-specific regulatory frameworks to supplement a central privacy protection law.[52]

Privacy (Other/General)

Documents that were coded for the "privacy (other/general)" principle generally contain broad statements on the relevance of privacy protections to the ethical or rights-respecting development and deployment of AI. This was the single most popular principle in our dataset; nearly all of the documents in our dataset contained it.[53] Given the breadth of coverage for this principle, it's interesting to observe significant variety in the justifications for its importance. Many actors behind principles documents root the privacy principle in compliance with law, whether international human rights instruments or national or regional laws such as the GDPR, but others offer alternative rationales.

Privacy is frequently called out as the prime example of the relevance of a rights framework to AI technology. The OECD and G20 AI principles call for "respect [for] the rule of law, human rights and democratic values," including respect for privacy.[54] The Toronto Declaration, which takes human rights as an overall framework for its approach to AI governance, also highlights the importance of privacy, stating that "States must adhere to relevant national and international laws and regulations that codify and implement human rights obligations protecting against discrimination and other related rights harms, for example data protection and privacy laws."[55] Finally, in the private sector, where AI principles most commonly take the form of internal company commitments, Telia Company engages to examine the "how we manage human rights risks and opportunities, such as privacy."[56] Other private sector actors including Microsoft, Telefónica, IA Latam, and IBM, describe respect of privacy as a legal obligation and in most cases refer to privacy as a right.

Outside of compliance, we found a wealth of other grounds for the primacy of privacy. The German AI strategy describes strong privacy standards as not only necessary from a legal and ethical standpoint but as "a competitive advantage internationally."[57] Google, and ITI describe respect of user privacy as a corporate responsibility owed to users and a business imperative.[58] The U.S. Science and Technology Council report balances consumer privacy against the value of "rich sets of data."[59] Other non-legal justifications included cybersecurity benefits,[60] alignment with public opinion,[61] and the author institution's preexisting public commitment to a set of privacy principles.[62]

---

[46] Access Now (n 10) p. 30.

[47] Niti Aayog (n 24) p. 87.

[48] British Embassy in Mexico City, 'Artificial Intelligence in Mexico (La Inteligencia Artificial En México)' (2018) p. 49 <https://docs.wixstatic.com/ugd/7be025_ba24a518a53a4275af4d7ff63b4cf594.pdf>.

[49] German Federal Ministry of Education and Research, the Federal Ministry for Economic Affairs and Energy, and the Federal Ministry of Labour and Social Affairs (n 10) p. 28.

[50] Mission assigned by the French Prime Minister (n 8) p. 114.

[51] Niti Aayog (n 24) p. 87.

[52] Niti Aayog (n 24) p. 87.

[53] The three documents that did not include this principle are the Public Voice Coalition AI guidelines, the Ground Rules for AI conference paper, and the Singapore Monetary Authority's AI principles. The Public Voice Coalition AI guidelines is not coded for any principle in the Privacy theme, although in external materials such as the explanatory memorandum and references section, the organization makes it clear that privacy and data protection laws were highly influential; particularly in the framing of their "transparency" principle. *See* The Public Voice Coalition, 'Universal Guidelines for Artificial Intelligence' (2018) <https://thepublicvoice.org/ai-universal-guidelines/>.

[54] Organisation for Economic Co-operation and Development, 'Recommendation of the Council on Artificial Intelligence' (2019) p. 7 (*See* Principle 1.2) <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>; G20 Trade Ministers and Digital Economy Ministers, 'G20 Ministerial Statement on Trade and Digital Economy' (2019) p. 11 (*See* Principle 1.2) <https://www.mofa.go.jp/files/000486596.pdf>.

[55] Amnesty International, Access Now, 'Toronto Declaration: Protecting the Right to Equality and Non-Discrimination in Machine Learning Systems' (2018) p. 23 <https://www.accessnow.org/cms/assets/uploads/2018/08/The-Toronto-Declaration_ENG_08-2018.pdf>.

[56] Telia Company, 'Guiding Principles on Trusted AI Ethics' (2019) principle 3 <https://www.teliacompany.com/globalassets/telia-company/documents/about-telia-company/public-policy/2018/guiding-principles-on-trusted-ai-ethics.pdf>.

[57] German Federal Ministry of Education and Research, the Federal Ministry for Economic Affairs and Energy, and the Federal Ministry of Labour and Social Affairs (n 9) principle 16.

[58] Information Technology Industry Council (n 8) p. 1; Microsoft (n 26) p. 66.

[59] United States Executive Office of the President, National Science and Technology Council Committee on Technology, 'Preparing for the Future of Artificial Intelligence' (2016) p. 20 <https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf>.

[60] Microsoft (n 27) p. 68.

[61] IBM (n 24) p. 44.

[62] *See generally* Google (n 22); Telefónica, 'AI Principles of Telefónica' (2018) <https://www.telefonica.com/en/web/responsible-business/our-commitments/ai-principles>; Microsoft (n 26).

3.2. Accountability

On its face, the term "artificial intelligence" suggests an equivalence with human intelligence. Depending on who you ask, the age of autonomous AIs is either upon us or uncertain centuries in the future, but concerns about who will be accountable for decisions that are no longer made by humans – as well as the potentially enormous scale of this technology's impacts on the social and natural world – likely lie behind the prevalence of the Accountability theme in our dataset.[63] Almost all documents that we analyzed mention at least one Accountability principle: "recommends adoption of new regulations," "verifiability and replicability," "impact assessments," "environmental responsibility," "evaluation and auditing requirements," "creation of a monitoring body," "ability to appeal," "remedy for automated decision," "liability and legal responsibility," and "accountability per se."

PRINCIPLES UNDER THIS THEME

- **36%** Verifiability and Replicability
- **53%** Impact Assessments
- **17%** Environmental Responsibility
- **47%** Evaluation and Auditing Requirement
- **17%** Creation of a Monitoring Body
- **22%** Ability to Appeal
- **11%** Remedy for Automated Decision
- **31%** Liability and Legal Responsibility
- **53%** Recommends Adoption of New Regulations
- **69%** Accountability Per Se

Percentage reflects the number of documents in the dataset that include each principle

[63] Accountability principles are present in 97% of documents in the dataset. Only one document did not mention an accountability principle. This company, Telefónica, received the highest score in the 2019 Ranking Digital Rights report, and it will be interesting to see how its ranking is impacted when, in RDR's next report, it adds AI governance to its questionnaire. Telefónica, 'AI Principles of Telefónica' (October 2018).

The documents reflect diverse perspectives on the mechanisms through which accountability should be achieved. It's possible to map the principles within the Accountability theme across the lifecycle of an AI system, in three essential stages: design (pre-deployment), monitoring (during deployment), and redress (after harm has occurred).

Design	Monitoring	Redress
Verifiability and Replicability	Evaluation and Auditing Requirements	Remedy for Automated Decision
Impact Assessment	Creation of a Monitoring Body	Liability and Legal Responsibility
Environmental Responsibility	Ability to Appeal	Recommends Adoption of New Regulations

Of course, each principle may have applicability across multiple stages as well. For example, the "verifiability and replicability" and "environmental responsibility" principles listed under the design stage in the above table will also be relevant in the monitoring and redress phases, but for optimal implementation should be accounted for when the system is designed.

The Accountability theme shows strong connections to the themes of Safety and Security, Transparency and Explainability, and Human Control of Technology.[64] Accountability principles are frequently mentioned together with the principle of transparent and explainable AI,[65] often highlighting the need for accountability as a means to gain the public's trust[66] in AI and dissipate fears.[67]

Verifiability and Replicability

The principle of "verifiability and replicability" provides for several closely related mechanisms to ensure AI systems are functioning as they should: an AI experiment ought to "exhibit[] the same behavior when repeated under the same conditions"[68] and provide sufficient detail about its operations that it may be validated.

[64] Access Now (n 10) p. 33; Google (n 23) (*See* Principle 4.)

[65] Mission assigned by the French Prime Minister (n 8) p. 113; Amnesty International, Access Now (n 56) p. 9; UNI Global Union, 'Top 10 Principles for Ethical Artificial Intelligence' (2017) p. 6 <http://www.thefutureworldofwork.org/media/35420/uni_ethical_ai.pdf>; IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (n 5) pp. 29-30 (*See* Principle 6); Standard Administration of China and Triolo (n 24) (*See* Principle 3.3.1.); German Federal Ministry of Education and Research, the Federal Ministry for Economic Affairs and Energy, and the Federal Ministry of Labour and Social Affairs (n 10) p. 16; Japanese Cabinet Office, Council for Science, Technology and Innovation (n 20) p. 10 (*See* Principle 4.1.6.)

[66] German Federal Ministry of Education and Research, the Federal Ministry for Economic Affairs and Energy, and the Federal Ministry of Labour and Social Affairs (n 10) p. 16.

[67] *See e.g.*, IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (n 5) pp. 29-30 (*See* Principle 6): "We have learnt that users build a relationship with AI and start to trust it after just a few meaningful interactions. With trust, comes responsibility.".

[68] European Commission's High-Level Expert Group on Artificial Intelligence (n 6) p. 17.

The German AI Strategy highlights that a verifiable AI system should be able to "effectively prevent distortion, discrimination, manipulation and other forms of improper use."[69] The development of verifiable AI systems may have institutional components along with technical ones. Institutionally, auditing institutions could "verify algorithmic decision-making in order to prevent improper use, discrimination and negative impacts on society"[70] and "new standards, including standards for validation or certification agencies on how AI systems have been verified"[71] could be developed.

Impact Assessments

The "impact assessments" principle captures both specific calls for human rights impact assessments (HRIAs) as well as more general calls for the advance identification, prevention, and mitigation of negative impacts of AI technology. One way to measure negative impacts of AI systems is to evaluate its "risks and opportunities" for human rights,[72] whether through HRIAs[73] or human rights due diligence.[74] Where HRIAs are called for, documents frequently also provide structure for their design: the Access Now report, for example, outlines that the assessment should include a consultation with relevant stakeholders "particularly any affected groups, human rights organizations, and independent human rights and AI experts."[75] For other actors – often those less closely grounded in the daily management of technology's human rights harms – this principle translated to calls for the assessment of "both direct and indirect harm as well as emotional, social, environmental, or other non-financial harm."[76]

We observed that some documents use the terminology of potential *harm*[77] and others call for the identification of *risks*.[78] The emphasis, particularly among the latter category of documents, is on prevention, and impact assessments are an accountability mechanism because a sufficiently dire assessment (where risks are "too high or impossible to mitigate"[79]) should prevent an AI technology from being deployed or even developed. Some documents suggest that an AI system should only be used after evaluating its "purpose and objectives, its benefits, as well as its risks."[80] In this context, it is particularly important that the AI system can be tested in a controlled environment and scaled-up as appropriate.[81] The Smart Dubai AI principles document calls for the use of AI systems only if they are "backed by respected and evidence-based academic research, and AI developer organizations."[82]

Environmental Responsibility

The principle of "environmental responsibility" reflects the growing recognition that AI, as a part of our human future, will necessarily interact with environmental concerns, and that those who build and implement AI technology must be accountable for its ecological impacts. The documents address environmental responsibility from two different angles.

Some documents capture this principle through an insistence that the environment should be a factor that is considered within the assessment of potential harm.[83] IA Latam's principles, for example, stress that the impact of AI systems should not "represent a threat for our environment."[84] Other documents go further, moving from a prohibition on negative ramifications to prescribe that AI technologies must be designed "to protect the environment, the climate and natural resources"[85] or to "promote the sustainable development of nature and society."[86]

Evaluation and Auditing Requirement

The "evaluation and auditing requirement" principle articulates the importance of not only building technologies that are capable of being audited,[87] but also to use the learnings from evaluations to feed back into a system and to ensure that it is continually improved, "tuning AI models periodically to cater for changes to data and/or models over time."[88]

A frequent focus is on the importance of humans in the auditing exercise, either as an auditing authority[89] or as users of AI systems who are solicited for feedback.[90] The Toronto Declaration calls upon developers to submit "systems that have a significant risk of resulting in human rights abuses to *independent* third-party audits."[91] The T20 report on the future of work and education focuses instead on breadth of input, highlighting the need for training data and features to "be

[69] German Federal Ministry of Education and Research, the Federal Ministry for Economic Affairs and Energy, and the Federal Ministry of Labour and Social Affairs (n 10) p. 38.

[70] German Federal Ministry of Education and Research, the Federal Ministry for Economic Affairs and Energy, and the Federal Ministry of Labour and Social Affairs (n 10) p. 38.

[71] IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (n 5) p. 28, addressing the topic within the principle of transparency.

[72] Telia Company (n 56) (*See* Principle 3.)

[73] Access Now (n 10) p. 32; IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (n 5) p. 19; Council of Europe, European Commission For The Efficiency of Justice, 'European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment' (2018) p. 8 (*See* Principle 1) <https://rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699c>.

[74] Amnesty International, Access Now (n 56) p. 12.

[75] Access Now (n 10) p. 34.

[76] Access Now (n 10) p. 34.

[77] Access Now (n 10) p. 34; European Commission's High-Level Expert Group on Artificial Intelligence (n 6) p. 19.

[78] Niti Aayog (n 24) p. 87; Smart Dubai (n 23) p. 23 (*See* Principle 1.2.2.3.); Chinese National Governance Committee for the New Generation Artificial Intelligence, led by China's Ministry of Science and Technology (n 23) (*See* Principle 8, English translation available upon request); Japanese Cabinet Office, Council for Science, Technology and Innovation (n 20) pp. 8-9 (using the term 'harm' within the principle of privacy protection and 'risks' within the principle of ensuring security; each time elaborating on impact assessment.)

[79] Amnesty International, Access Now (n 56) p. 13 (*See* para. 48.)

[80] The Public Voice Coalition (n 54) (*See* Principle 5.)

[81] Organisation for Economic Co-operation and Development (n 54) p. 9 (*See* Principle 2.3); G20 Trade Ministers and Digital Economy Ministers (n 54) p. 13 (*See* Principle 2.3.)

[82] Smart Dubai (n 23) p. 22 (*See* Principle 1.2.2.1.)

[83] European Commission's High-Level Expert Group on Artificial Intelligence (n 6) p. 19.

[84] IA Latam (n 22) (See Principle 5, English translation available upon request.)

[85] German Federal Ministry of Education and Research, the Federal Ministry for Economic Affairs and Energy, and the Federal Ministry of Labour and Social Affairs (n 10) p. 20.

[86] Beijing Academy of Artificial Intelligence (n 42) (*See* Principle 1.1 English translation available upon request.)

[87] Amnesty International, Access Now (n 54) p. 9 (*See* para. 32, particularly within the context of government acquisitions of AI systems); Mission assigned by the French Prime Minister (n 6) p. 113 (call for the development of capacities to understand and audit AI systems).

[88] Smart Dubai (n 23) p. 23 (*See* Principles 1.2.2.4 and 1.2.2.5.)

[89] Future of Life Institute, 'Asilomar AI Principles' (2017) p. 8 <https://futureoflife.org/ai-principles/?cn-reloaded=1>.

[90] Google (n 23) (*See* Principle 4.)

[91] Amnesty International, Access Now (n 56) p. 13 (*See* para. 47) (emphasis added).

reviewed by many eyes to identify possible flaws and to counter the 'garbage in garbage out' trap."[92]

Some, but not all, documents have drafted their "evaluation and auditing" principles to contain significant teeth. Some documents recommend the implementation of mechanisms that allow an eventual termination of use. Such a termination is recommended, in particular, if the AI systems "would violate international conventions or human rights."[93] The Access Now report suggests the development of "a failsafe to terminate acquisition, deployment, or any continued use if at any point an identified human rights violation is too high or unable to be mitigated."[94]

Creation of a Monitoring Body

The principle of "creation of a monitoring body" reflects a repeated recognition that some new organization or structure may be required to create and oversee standards and best practices in the context of AI. Visions for how these bodies may be constituted and what activities they would undertake vary.

The Ethically Aligned Design document situates the need for this new body in its pursuit to ensure that AI systems do "not infringe upon human rights, freedoms, dignity, and privacy."[95] Microsoft's AI principles suggest the creation of "internal review boards" – internal, we presume, to the company, but not to the teams that are building the technology. The Toronto Declaration stresses that any monitoring body should be independent and might include "judicial authorities when necessary."[96] The German AI strategy outlines the creation of a national AI observatory, which could also be tasked to monitor that AI systems are designed socially compatible and to develop auditing standards.[97]

Ability to Appeal

The principle of an "ability to appeal" concerns the possibility that an individual who is the subject of a decision made by an AI could challenge that decision. The ability to appeal connects with the theme of Human Control of Technology, in that it's often mentioned in connection with the principle of "right to human review of an automated decision."[98] Some documents in fact collapse the two.[99] The Access Now report calls the human in the loop an element that adds a "layer of accountability."[100]

In some individual documents, this principle is parsed more neatly, as for example in the Access Now report which explains that there should be both an ability to *challenge the use* of an AI system and an ability to *appeal a decision* that has been "informed or wholly made by an AI system."[101] The ability to appeal the use of or recommendation made by an AI system could be realized in form of a judicial review.[102] Further, some documents limit the ability to appeal only to "significant automated decisions."[103]

A subset of documents recognize as part of this principle the importance of making AI subjects aware of existing procedures to vindicate their rights[104] or to broaden the accessibility of channels for the exercise of subjects' rights.[105] In order to enable AI subjects to challenge the outcome of AI systems, the OECD and G20 AI principles suggest that the outcome of the system must be "based on plain and easy-to-understand information on the factors, and the logic that served as the basis for the prediction, recommendation or decision."[106]

Remedy for Automated Decision

The principle of "remedy for automated decision" is fundamentally a recognition that as AI technology is deployed in increasingly critical contexts, its decisions will have real consequences, and that remedies should be available just as they are for the consequences of human actions. The principle of remedy is intimately connected to the ability to appeal, since where appeal allows for the rectification of the decision itself, remedy rectifies its consequences.[107]

There is a bifurcation in many of the documents that provide for remedy between the remedial mechanisms that are appropriate for state use of AI versus those that companies should implement for private use. For example, the Toronto Declaration has separate principles for company and state action, providing that companies may "for example, creat[e] clear, independent, visible processes for redress following adverse individual or societal effects, and designat[e] roles in the entity responsible for the timely remedy of such issues"[108] whereas states should provide "reparation that, where appropriate, can involve compensation, sanctions against those responsible, and guarantees of non-repetition. This may be possible using existing laws and regulations or may require developing new ones."[109] Other documents suggest further important delineations of responsibilities, including between vendors and clients.[110]

Liability and Legal Responsibility

The principle of "liability and legal responsibility" refers to the concept that it is necessary to ensure that the individuals or entities at fault for harm

[92] Think 20 (n 39) p. 6.

[93] Partnership on AI, 'Tenets' (2016) (*See* Principle 6) <https://www.partnershiponai.org/tenets/>.

[94] Access Now (n 10) p. 33.

[95] IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (n 5) p. 19 (*See* Principle 1.)

[96] Amnesty International, Access Now (n 55) p. 10.

[97] German Federal Ministry of Education and Research, the Federal Ministry for Economic Affairs and Energy, and the Federal Ministry of Labour and Social Affairs (n 9) p. 26.

[98] UNI Global Union (n 66) p. 7; Google (n 23) (*See* Principle 4.)

[99] Think 20 (n 38) p. 8.

[100] Access Now (n 9) p. 32.

[101] Access Now (n 9) p. 33.

[102] Amnesty International, Access Now (n 55) p. 14.

[103] Smart Dubai (n 23) p. 9.

[104] Smart Dubai (n 23) p. 9.

[105] Monetary Authority of Singapore (n 38) p. 11.

[106] Organisation for Economic Co-operation and Development (n 54) p. 8 (*See* Principle 1.3); G20 Trade Ministers and Digital Economy Ministers (n 54) p. 11 (*See* Principle 1.3.)

[107] Tencent Institute (n 58) (*See* Principle 4, English translation available upon request.)

[108] Amnesty International, Access Now (n 55) p. 15 (*See* Principle 53.)

[109] Amnesty International, Access Now (n 55) p. 15 (*See* Principle 56.)

[110] Access Now (n 10) p. 35 (See para. 3.)

caused by an AI system can be held accountable. While other forms of automation and algorithmic decision making have existed for some time, emerging AI technologies can place further distance between the result of an action and the actor who caused it, raising questions about who should be held liable and under what circumstances. These principles call for reliable resolutions to those questions.

Many documents point out that existing systems may be sufficient to guarantee legal responsibility for AI harms, with actors including Microsoft and the Indian AI strategy looking to tort law and specifically negligence as a sufficient solution. Others, such as the Chinese AI Industry Code of Conduct, assert that there is additional work to be done to "[c]larify the rights and obligations of parties at each stage in research and development, design, manufacturing, operation and service of AI, to be able to promptly determine the responsible parties when harm occurs."[111]

There exists some reluctance to hold developers liable for the consequences of AI's deployment. The Chinese White Paper on AI Standardization distinguishes in its principle of liability between liability at the level of development and at the level of deployment, recommending transparency as the most appropriate accountability mechanism at the development level and suggesting the establishment of a reasonable system of liability and compensation post-deployment.[112] The Montreal Declaration makes a similar distinction, stating "[w]hen damage or harm has been inflicted by an [AI system that]… is proven to be reliable and to have been used as intended, it is not reasonable to place blame on the people involved in its development or use."[113]

Recommends Adoption of New Regulations

The "recommends adoption of new regulations" principle reflects a position that AI technology represents a significant enough departure from the status quo that new regulatory regimes are required to ensure it is built and implemented in an ethical and rights-respecting manner. Some documents that contain this principle refer to existing regulations,[114] but there is a general consensus that it is necessary to reflect on the adequacy of those frameworks.[115] Documents that contain this principle frequently express an urgent need for clarity about parties' respective responsibilities.[116] A few documents address the fact that "one regulatory approach will not fit all AI applications"[117] and emphasize the need to adopt context specific regulations, for example, regarding the use of AI for surveillance and similar activities that are likely to interfere with human rights.[118]

Among statements of this principle, we see a variety of justifications for future regulation, some of which are recognizable from other themes in our data: the regulation should ensure that the development and use of AI is safe and beneficial to society;[119] implement oversight mechanisms "in contexts that present risk of discriminatory or other rights-harming outcomes;"[120] and identify the right balance between innovation and privacy rights.[121]

There is also a common emphasis on the need for careful balancing in crafting regulation. The trade industry group ITI cautions that new regulations might "inadvertently or unnecessarily impede the responsible development and use of AI."[122] On the other hand, the OECD AI principles and G20 AI principles state that appropriate policy and regulatory frameworks can "encourage innovation and competition for trustworthy AI."[123] Many documents recognize that new laws and regulations are appropriate if lawmakers use them alongside self-regulation and existing policy tools. The AI for Europe document states that "self-regulation can provide a first set of benchmarks" but that the European Commission should "monitor developments and, if necessary, review existing legal frameworks."[124] The Standards Administration of China suggested that new regulations might be based on "universal regulatory principles"[125] that would be formulated at an international level.

Accountability Per Se

Like many of our themes, the Accountability theme contains an "accountability" principle, but in this specific case, only to those documents that explicitly use the word "accountability" or "accountable" (25 of the 36 documents) were coded under this principle. Because principles documents are frequently challenged as toothless or unenforceable, we were interested to see how documents grappled with this term specifically. In this context, documents converge on a call for developing "accountability frameworks"[126] that define the responsibility of different entities "at each stage in research and development, design, manufacturing, operation and service."[127]

Notably, a few documents emphasize that the responsibility and accountability of AI systems cannot lie with the technology itself, but should be "apportioned between those who design, develop and deploy [it]."[128] Some documents propose specific entities that should be held accountable if harm occurs, including the government,[129]

---

[111] Artificial Intelligence Industry Alliance, 'Artificial Intelligence Industry Code of Conduct (Consultation Version)' (2019) (*See* Principle 8, English translation available upon request) <https://www.secrss.com/articles/11099>.

[112] Standard Administration of China (n 23) (*See* Principle 3.3.2.)

[113] University of Montreal (n 35) p. 16 (*See* Principle 9.5.)

[114] Mission assigned by the French Prime Minister (n 8) p. 114 (referring to the French Data Protection Act of 1978 and the GDPR (2018.)

[115] European Commission, 'Artificial Intelligence for Europe: Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee, and the Committee of the Regions' COM (2018) p. 16 <https://ec.europa.eu/digital-single-market/en/news/communication-artificial-intelligence-europe> (Stressing, in particular, the need to reflect on "the suitability of some established rules on safety and civil law questions on liability.")

[116] UK House of Lords, Select Committee on Artificial Intelligence (n 8) p. 135 (*See* para. 56.)

[117] Information Technology Industry Council (n 8) p. 4.

[118] Access Now (n 9) p. 32.

[119] Beijing Academy of Artificial Intelligence (n 23) (*See* Preamble, English translation available upon request); Tencent Institute (n 58) (*See* Principle 18, English translation available upon request.)

[120] Amnesty International, Access Now (n 55) p. 11.

[121] British Embassy in Mexico City (n 49) p. 49.

[122] Information Technology Industry Council (n 8) p. 4.

[123] Organisation for Economic Co-operation and Development (n 54) p. 9 (*See* Principle 2.3); G20 Trade Ministers and Digital Economy Ministers (n 54) p. 13 (*See* Principle 2.3.)

[124] European Commission (n 115) p. 16.

[125] Standard Administration of China (n 23) (See Principle 3.3.1.)

[126] Information Technology Industry Council (n 8) p. 4.

[127] Artificial Intelligence Industry Alliance (n 111) (*See* Article 8, English translation available upon request.)

[128] Smart Dubai (n 22) p. 7.

[129] Access Now (n 9) p. 33.

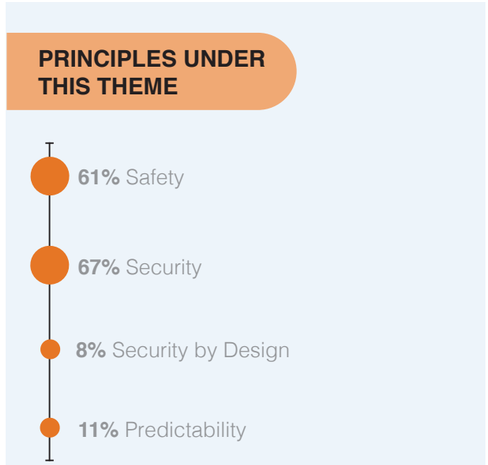
companies and their business partners,[130] researchers, developers and users.[131] The OECD AI principles and G20 AI principles suggest that accountability should adapt to the context in which the technology is used.[132]

3.3. Safety and Security

Given early examples of AI systems' missteps[133] and the scale of harm they may cause, concerns about the safety and security of AI systems were unsurprisingly a significant theme among principles in the documents we coded.[134] There appears to be a broad consensus across different actor types on the centrality of Safety and Security, with about three-quarters of the documents addressing principles within this theme. There are four principles under it: "safety," "security," "security by design," and "predictability."

It is worth distinguishing, up front, the related concepts of safety and security. The principle of safety generally refers to proper internal functioning of an AI system and the avoidance of unintended harms. By contrast, security addresses external threats to an AI system. However, documents in our dataset often mention the two principles together, and indeed they are closely intertwined. This observation becomes particularly evident when documents use the related term "reliability":[135] a system that is reliable is safe, in that it performs as intended, and also secure, in that it is not vulnerable to being compromised by unauthorized third parties.

There are connections between this theme and the Accountability, Professional Responsibility, and Human Control of Technology themes. In many ways, principles under these other themes can be seen, at least partially, as implementation mechanisms for the goals articulated under Safety and Security.



[130] Telia Company (n 56) (*See* Principle 5.)

[131] University of Montreal (n 34) p. 14 (*See* Principle 7.)

[132] Organisation for Economic Co-operation and Development (n 54) p. 8 (*See* Principle 1.5); G20 Trade Ministers and Digital Economy Ministers (n 54) p. 12 (*See* Principle 1.5.)

[133] *See* National Transportation Safety Board Office of Public Affairs, "'Inadequate Safety Culture' Contributed to Uber Automated Test Vehicle Crash - NTSB Calls for Federal Review Process for Automated Vehicle Testing on Public Roads," (Nov. 19, 2019), <https://www.ntsb.gov/news/press-releases/Pages/NR20191119c.aspx> (describing the results of the National Transportation Safety Board's investigation in the fatal collision between an automated test vehicle operated by Uber and a pedestrian in Tempe, Arizona. Stating: "Contributing to the crash was Uber ATG's inadequate safety risk assessment procedures, ineffective oversight of the vehicle operators and a lack of adequate mechanisms for addressing operators' automation complacency – all consequences of the division's inadequate safety culture."); *see also* Cade Metz and Scott Blumenthal, "How A.I. Could be Weaponized to Spread Disinformation," *The New York Times*, (June 7, 2019), <https://www.nytimes.com/interactive/2019/06/07/technology/ai-text-disinformation.html> (discussing the disinformation threat AI driven technologies that can create "false images and sounds that are indistinguishable from the real thing" and automated text-generation systems might pose to the online information ecosystem.)

[134] Safety and Security principles are present in 81% of documents in the dataset.

[135] Microsoft (n 27) p. 61; Partnership on AI (n 94) (*See* Principle 6); Beijing Academy of Artificial Intelligence (n 24) (*See* Principle 1.4, English translation available upon request); Japanese Cabinet Office, Council for Science, Technology and Innovation (n 20) p. 10 (*See* Principle 4.1.7.); European Commission's High-Level Expert Group on Artificial Intelligence (n 6) p. 17; Think 20 (n 39) p. 7; University of Montreal (n 35) p. 8 (*See* Principle 8.3); Chinese National Governance Committee for the New Generation Artificial Intelligence, led by China's Ministry of Science and Technology (n 23) (*See* Principle 5, English translation available upon request.)

Accountability measures are key guarantors of AI safety, including verifiability[136] and the need to monitor the operation of AI systems after their deployment.[137] Individuals and organizations behind AI technology have a key role in ensuring it is designed and used in ways that are safe and secure. Safety is thus frequently mentioned in connection with the need to ensure controllability by humans.[138]

Safety

The principle of "safety" requires that an AI system be reliable and that "the system will do what it is supposed to do without harming living beings or [its] environment."[139] Articulations of this principle focus both on safety measures to be taken both before AI systems are deployed[140] and after, "throughout their operational lifetime."[141] Safety measures during development require that AI systems are "built and tested to prevent possible misuse."[142] Building systems safely means avoiding "risks of harm"[143] by assessing safety risks[144] including potential human rights violations.[145] Testing procedures should not only apply to likely scenarios, but also establish that a system "responds safely to unanticipated situations and does not evolve in unexpected ways."[146]

Testing and monitoring of AI systems should continue after deployment according to a few articulations of the "safety" principle. This is particularly relevant where the document focuses on machine learning technology, which is likely to evolve following implementation as it continues to receive input of new information. Developers of AI systems cannot always "accurately predict the risks"[147] associated with such systems ex ante. There are also safety risks associated with AI systems being implemented in ways that their creators did not anticipate, but one document suggests that designing AI that could be called safe might require the technology makes "relatively safe decisions" "even when faced with different environments in the decision-making process."[148]

Finally, two documents coded for the "safety" principle specifically call for the development of safety regulations to govern AI. One call relates specifically to the regulation of autonomous vehicles[149] and the other is more general, calling for "high standards in terms of safety and product liability"[150] within the EU. Other documents call for public awareness campaigns to promote safety.[151] For example, IEEE's Ethically Aligned Design suggests that "in the same way police officers have given public safety lectures in schools for years; in the near future they could provide workshops on safe [AI systems]."[152]

Security

The principle of "security" concerns an AI system's ability to resist external threats. Much of the language around security in our dataset is high level, but in broad terms, the documents coded here call for three specific needs to protect against security threats: the need to test the resilience of AI systems;[153] to share information on vulnerabilities[154] and cyberattacks;[155] and to protect privacy[156] and "the integrity and confidentiality of personal data."[157] With regard to the latter need, the ITI AI Policy Principles suggest that the security of data could be achieved through anonymization, de-identification, or aggregation, and they call on governments to "avoid requiring companies to transfer or provide access to technology, source code, algorithms, or encryption keys as conditions for doing business."[158] The Chinese White Paper on AI Standardization suggests that the implementation of security assurance requirements could be facilitated through a clear distribution of liability and fault between developers, product manufacturers, service providers and end users.[159]

A number of documents, concentrated in the private sector, emphasize the "integral"[160] role of security in fostering trust in AI systems.[161] The ITI AI Policy Principles state that AI technology's success depends on users' "trust that their personal and sensitive data is protected and handled appropriately."[162]

Security by Design

The "security by design" principle, as its name suggests, is related to the development of secure AI systems. The European High Level Expert Group guidelines observes that these "values-

[136] Future of Life Institute (n 90) (*See* Principle 2); Smart Dubai (n 23) p. 9.

[137] Google (n 22) (*See* Principle 3); Chinese National Governance Committee for the New Generation Artificial Intelligence, led by China's Ministry of Science and Technology (n 22) (*See* Principle 5, English translation available upon request.)

[138] Information Technology Industry Council (n 9) p. 3; Smart Dubai (n 23) p. 9. Think 20 (n 39) p. 7.

[139] European Commission's High-Level Expert Group on Artificial Intelligence (n 6) p. 17.

[140] Telia Company (n 57) (*See* Principle 6); Think 20 (n 39) p. 7; Google (n 23) (*See* Principle 3.)

[141] Future of Life Institute (n 90) (*See* Principle 2); Smart Dubai (n 23) p. 9.

[142] Telia Company (n 56) (*See* Principle 6.)

[143] Google (n 22) (*See* Principle 3.)

[144] European Commission's High-Level Expert Group on Artificial Intelligence (n 6) pp. 16-17; Organisation for Economic Co-operation and Development (n 55) p. 8 (*See* Principle 1.4); G20 Trade Ministers and Digital Economy Ministers (n 55) pp. 11-12 (*See* Principle 1.4); Smart Dubai (n 23) p. 9; Think 20 (n 39) p. 7; Japanese Cabinet Office, Council for Science, Technology and Innovation (n 20) p. 10 (*See* Principle 4.1.7.); The Public Voice Coalition (n 54) (*See* Principle 8); Beijing Academy of Artificial Intelligence (n 24) (*See* Principle 1.4, English translation available upon request.)

[145] Partnership on AI (n 94) p. 6.

[146] Think 20 (n 39) p. 7; Japanese Cabinet Office, Council for Science, Technology and Innovation (n 20) p. 9 (stating, "it is not always possible for AI to respond appropriately to rare events or deliberate attacks" and consequently arguing that society should be empowered to balance risks and benefits.)

[147] Standard Administration of China (n 23) (*See* Principle 3.3.1.)

[148] Standard Administration of China (n 23) (*See* Principle 3.3.1.)

[149] United States Executive Office of the President, National Science and Technology Council Committee on Technology (n 59) p. 17.

[150] European Commission (n 115) p. 15.

[151] Tencent Institute (n 58) (*See* Principle 16, English translation available upon request); Japanese Cabinet Office, Council for Science, Technology and Innovation (n 19) p. 9 (*See* Principle 4, stating "Society should always be aware of the balance between the benefits and risks.")

[152] IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (n 5) p. 31 (*See* Principle 5.)

[153] Google (n 22) (*See* Principle 3.)

[154] University of Montreal (n 34) p. 15 (*See* Principle 8.5.)

[155] Information Technology Industry Council (n 8) p. 4.

[156] Microsoft (n 27) p. 66; Smart Dubai (n 23) p. 9; Think 20 (n 39) p. 20; Information Technology Industry Council (n 9) p. 4; European Commission (n 116) p. 15.

[157] University of Montreal (n 34) p. 15 (*See* Principle 8.4.)

[158] Information Technology Industry Council (n 8) p. 4.

[159] Standard Administration of China (n 23) (*See* Principle 3.3.1.)

[160] Information Technology Industry Council (n 8) p. 4

[161] See IA Latam (n 23) (*See* Principle 10, English translation available upon request); Telefónica (n 63) (See Principle 4); The Public Voice Coalition (n 54) (*See* Principle 9); Telia Company (n 57) (*See* Principle 6); Google (n 23) (*See* Principle 3.)

[162] Information Technology Industry Council (n 8) p. 4.

by-design" principles may provide a link between abstract principles and specific implementation decisions.[163]

A few documents argue that existing and widely adopted security standards should apply for the development of AI systems. The German AI Strategy suggests that security standards for critical IT infrastructure should be used[164] and the Microsoft AI Principles mention that principles from other engineering disciplines of robust and fail-safe design can be valuable.[165] Similarly, the European High Level Expert Group guidelines argue for AI systems to be built with a "fallback plan" where, in the event of a problem, a system would switch its protocol "from statistical to rule-based" decision-making or require the intervention of a human before continuing.[166]

Predictability

The principle of "predictability" is concisely defined in the European High Level Expert Group guidelines, which state that for a system to be predictable, the outcome of the planning process must be consistent with the input.[167] Predictability is generally presented as a key mechanism to ensure that AI systems have not been compromised by external actors. As the German AI strategy puts it, "transparent, predictable and verifiable" AI systems may "effectively prevent distortion, discrimination, manipulation and other forms of improper use."[168] As in the "security" principle, there is an observable connection between predictable AI systems and public trust, with the Beijing AI Principles observing that improving predictability, alongside other "ethical design approaches" should help "to make the system trustworthy."[169]

[163] European Commission's High-Level Expert Group on Artificial Intelligence (n 6) p. 21.

[164] German Federal Ministry of Education and Research, the Federal Ministry for Economic Affairs and Energy, and the Federal Ministry of Labour and Social Affairs (n 10) p. 37.

[165] Microsoft (n 27) p. 64.

[166] European Commission's High-Level Expert Group on Artificial Intelligence (n 6) p. 17 (*See* Principle 1.2 Technical robustness and safety.)

[167] European Commission's High-Level Expert Group on Artificial Intelligence (n 6) p. 22.

[168] German Federal Ministry of Education and Research, the Federal Ministry for Economic Affairs and Energy, and the Federal Ministry of Labour and Social Affairs (n 10) p. 38.

[169] Beijing Academy of Artificial Intelligence (n 24) (*See* Principle 1.5, English translation available upon request.)

3.4. Transparency and Explainability

Perhaps the greatest challenge that AI poses from a governance perspective is the complexity and opacity of the technology. Not only can it be difficult to understand from a technical perspective, but early experience has already proven that it's not always clear when an AI system has been implemented in a given context, and for what task. The eight principles within the theme of Transparency and Explainability are a response to these challenges: "transparency," "explainability," "open source data and algorithms," "open government procurement," "right to information," "notification when interacting with an AI," "notification when AI makes a decision about an individual," and "regular reporting." The principles of transparency and explainability are some of the most frequently occurring individual principles in our dataset, each mentioned in approximately three-quarters of the documents.[170]

It is interesting to note a bifurcation among the principles under this theme, where some, including "explainability" and the ability to be notified when you are interacting with an AI or subject to an automated decision, are responses to entirely new governance challenges posed by the specific capabilities of current and emerging AI technologies. The rest of the principles in this theme, such as "open source data and algorithms" and "regular reporting" are well-established pillars of technology governance, now applied specifically to AI systems.

PRINCIPLES UNDER THIS THEME

- **72%** Transparency
- **78%** Explainability
- **28%** Open Source Data and Algorithms
- **3%** Open Government Procurement
- **11%** Right to Information
- **19%** Notification When AI Makes a Decision about an Individual
- **25%** Notification when Interacting with AI
- **17%** Regular Reporting

Percentage reflects the number of documents in the dataset that include each principle

[170] Transparency and Explainability principles are present in 94% of documents in the dataset. Only two documents do not include any principles under this theme. These are two government actors, the Standards Administrations of China and the report prepared by the British Embassy in Mexico City. Jeffrey Ding and Paul Triolo, "White Paper on Artificial Intelligence Standardization (Available Excerpts in English)," *New America*, January 2018, https://www.newamerica.org/cybersecurity-initiative/digichina/blog/translation-excerpts-chinas-white-paper-artificial-intelligence-standardization/; and "Artificial Intelligence in Mexico (La Inteligencia Artificial En México)" (Mexico City: British Embassy in Mexico City, June 2018), https://docs.wixstatic.com/ugd/7be025_ba24a518a53a4275af4d7ff63b4cf594.pdf.

Transparency and Explainability is connected to numerous other themes, most especially Accountability,[171] because principles within it may function as a "prerequisite for ascertaining that [such other] principles are observed."[172] It is also connected to the principle of predictability within the Safety and Security theme and to the Fairness and Non-discrimination theme.[173] The German government notes that individuals can only determine if an automated decision is biased or discriminatory if they can "examine the basis – the criteria, objectives, logic – upon which the decision was made."[174] Transparency and Explainability is a foundation for the realization of other many other principles.

Transparency

The principle of "transparency" is the assertion that AI systems should be designed and implemented in such a way that oversight of their operations are possible. The documents in the dataset vary in their suggestions about how transparency might be applied across institutions and technical systems throughout the AI lifecycle. The European High Level Expert Group guidelines note that transparency around "the data, the system, and the business models" all matter.[175] Some documents emphasize the importance of technical transparency, such as providing the relevant authorities with access to source code.[176]

Transparency throughout an AI system's life cycle means openness throughout the design, development, and deployment processes. While most documents treat transparency as binary — that is, an AI system is either transparent or it is not — several articulate the transparency principle as one that entities will strive for, with increased disclosure over time.[177] Some raise concerns about the implications of an over-broad transparency regime, which could give rise to conflicts with privacy-related principles.[178] IEEE's Ethically Aligned Design recommends the development of "new standards that describe measurable, testable levels of transparency, so that systems can be objectively assessed and levels of compliance determined."[179] Where sufficient transparency cannot be achieved, the Toronto Declaration calls upon states to "refrain from using these systems at all in high-risk contexts."[180]

Explainability

"Explainability" is defined in various ways, but is at its core about the translation of technical concepts and decision outputs into intelligible,[181] comprehensible formats suitable for evaluation. The T20 report on the future of work and education, for example, highlights the importance of "clear, complete and testable explanations of what the system is doing and why."[182] Put another way, a satisfactory explanation "should take the same form as the justification we would demand of a human making the same kind of decision."[183]

Many of the documents note that explainability is particularly important for systems that might "cause harm,"[184] have "a significant effect on individuals,"[185] or impact "a person's life, quality of life, or reputation."[186] The AI in the UK document suggests that if an AI system has a "substantial impact on an individual's life" and cannot provide "full and satisfactory explanation" for its decisions, then the system should not be deployed.[187]

The principle of explainability is closely related to the Accountability theme as well as the principle of "right to human review of automated decision" under the Human Control of Technology theme.[188] The Toronto Declaration mentions explainability as a necessary requirement to "effectively scrutinize" the impact of AI systems on "affected individuals and groups," to establish responsibilities, and to hold actors to account.[189] The European Commission's policy statement also connects explainability to the principle of nondiscrimination, as the development of understandable AI is crucial for minimizing "the risk of bias or error."[190] The need for explainability will become increasingly important as the capabilities and impact of AI systems compound.[191]

Open Source Data and Algorithms

The principle of "open source data and algorithms" is, as noted in the introduction to this theme, a familiar concept in technology governance, and it operates similarly in the context of AI as in other computer systems. The majority of documents that address it emphasize the value of the development of common algorithms[192] and open research and collaboration to support the advancement of the technology.[193] The Montreal Declaration describes this as a "socially equitable objective"[194] and the Beijing AI Principles note that open source solutions may be useful "to avoid data/platform monopolies, to share the benefits of AI development to the greatest extent, and

---

[171] *See*, e.g., Mission assigned by the French Prime Minister (n 7) p. 38.

[172] UNI Global Union (n 66) p. 7 (*See* Principle 1.)

[173] *See*, e.g., Council of Europe: European Commission For The Efficiency of Justice (CEPEJ), "European Ethical Charter on the Use of AI in Judicial Systems," p. 11 (*See* Principle 4); T20: Think20, "Future of Work and Education for the Digital Age", p. 7.

[174] German Federal Ministry of Education and Research, the Federal Ministry for Economic German Federal Ministry of Education and Research, the Federal Ministry for Economic Affairs and Energy, and the Federal Ministry of Labour and Social Affairs (n 10) p. 38.

[175] European Commission's High-Level Expert Group on Artificial Intelligence (n 6) p. 18.

[176] University of Montreal (n 34) (*See* Principle 5.3, stating: "[C]ode for algorithms, whether public or private, must always be accessible to the relevant public authorities and stakeholders for verification and control purposes.")

[177] Chinese National Governance Committee for the New Generation Artificial Intelligence, led by China's Ministry of Science and Technology (n 22) (*See* Principle 5, English translation available upon request); Telia Company (n 56) (*See* Principle 7); Artificial Intelligence Industry Alliance (n 111) (*See* Principle 6, English translation available upon request.)

[178] *See*, e.g., Monetary Authority of Singapore (n 37) p. 12 (*See* Principle 8.1, stating: "excessive transparency could create confusion or unintended opportunities for individuals to exploit or manipulate.")

[179] IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (n 5) p. 28.

[180] Amnesty International, Access Now (n 55) p. 9.

[181] We have coded "intelligibility," which is less common but does appear in at least three documents, as equivalent to explainability.

[182] Think 20 (n 39) p. 7.

[183] University of Montreal (n 34) p. 12 (*See* Principle 5.2.)

[184] Future of Life Institute (n 89) (*See* Principle 7); *See also*, University of Montreal (n 34) p. 12 (*See* Principle 5.2.)

[185] Smart Dubai (n 22) p. 8.

[186] University of Montreal (n 34) p. 12 (*See* Principle 5.2.)

[187] UK House of Lords, Select Committee on Artificial Intelligence (n 7) p. 40.

[188] Future of Life Institute (n 89) (*See* Principle 8.)

[189] Amnesty International, Access Now (n 55) p. 9.

[190] European Commission (n 115) p. 15.

[191] IBM (n 24) p. 28.

[192] University of Montreal (n 34) p. 13 (*See* Principle 6.7.)

[193] IA Latam (n 22) (*See* Principle 11, English translation available upon request.)

[194] University of Montreal (n 34) principle 6.7.

to promote equal development opportunities for different regions and industries."[195] Further, numerous documents also call for public and private investment in open datasets.[196]

The T20 report on the future of work and education focuses on the balance between transparency and the potential negative effect of open source policies on algorithmic innovation. One solution, they posit, is "algorithmic verifiability", which would "require companies to disclose information allowing the effect of their algorithms to be independently assessed, but not the actual code driving the algorithm."[197] Recognizing that data or algorithm disclosure is not sufficient to achieve transparency or explainability, the IEEE stresses the importance of disclosing the underlying algorithm to validation or certification agencies that can effectively serve as auditing and accountability bodies.[198]

Open Government Procurement

"Open government procurement," the requirement that governments be transparent about their use of AI systems, was only present in one document in our dataset. The Access Now report recommends that: "When a government body seeks to acquire an AI system or components thereof, procurement should be done openly and transparently according to open procurement standards. This includes publication of the purpose of the system, goals, parameters, and other information to facilitate public understanding. Procurement should include a period for public comment, and states should reach out to potentially affected groups where relevant to ensure an opportunity to input."[199]

It is notable that the Access Now report is one of the few documents in our dataset that specifically adopts a human rights framework. This principle accounts for the special duty of governments under Principle 5 of the UN Guiding Principles on Business and Human Rights to protect against human rights abuses when they contract with private businesses.

Right to Information

The "right to information" concerns the entitlement of individuals to know about various aspects of the use of, and their interaction with, AI systems. This might include "information about the personal data used in the decision-making process,"[200] "access to the factors, the logic, and techniques that produced the outcome" of an AI system,[201] and generally "how automated and machine learning decision-making processes are reached."[202]

As elsewhere where the word "right" is contained in the title of the principle, we only coded documents where they were explicitly articulated as a right or obligation. The OECD and G20 AI principles, for instance, do not call for an explicit "right to information" for users, and thus were not coded here, even though they recommend that those adversely affected by an AI system should be able to challenge it based on "easy-to-understand information on the factors, and the logic that served as the basis for the prediction, recommendation or decision."[203] One document specifically articulates the right to information as extending beyond a right to technical matter and data to the "obligation [that it] should be drawn up in plain language and be made easily accessible."[204]

Notification when AI Makes a Decision about an Individual

The definition of the principle of "notification when an AI system makes a decision about an individual" is facially fairly clear: where an AI has been employed, the person to whom it was subject should know. The AI in UK document stresses the importance of this principle to allow individuals to "experience the advantages of AI, as well as to opt out of using such products should they have concerns."[205] If people don't know when they are subject to automated decisions, they won't have the autonomy to decide whether or not they consent, or the information to reach their own conclusions about the overall value that AI provides.

In this respect, the notification principle connects to the themes of Human Control of Technology and Accountability. For example, the European Commission not only suggests that individuals should be able to opt out,[206] but also that they should be "informed on how to reach a human and how to ensure that a system's decisions can be checked or corrected,"[207] which is an important component of accountability. Access Now emphasizes the special importance of this principle when an AI system "makes a decision that impacts an individual's rights."[208]

Notification when Interacting with an AI

The principle of "notification when interacting with an AI system," a recognition of AI's increasing ability to pass the Turing test at least in limited applications, stands for the notion that humans should always be made aware when they are engaging with technology rather than directly with another person. Examples of when this principle is relevant include chatbot interactions,[209] facial recognition systems, credit scoring systems, and generally "where machine learning systems are used in the public sphere."[210]

Like "notification when an AI system makes a decision about an individual," this principle is a precondition to the actualization of other principles, including in the Accountability and Human Control of Technology themes. However, this principle is broader than the preceding one because it requires notification even in passive uses of AI systems. In the deployment of facial recognition systems, for example, the "decision"

[195] Beijing Academy of Artificial Intelligence (n 23) (*See* Principle 1.7, English translation available upon request.)

[196] Organisation for Economic Co-operation and Development (n 54) p. 8 (*See* Principle 2.1); G20 Trade Ministers and Digital Economy Ministers (n 54) p. 13 (*See* Principle 2.1.)

[197] Think 20 (n 38) p. 7.

[198] IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (n 5) p. 28 (*See* Principle 5.)

[199] Access Now (n 9) p. 32.

[200] German Federal Ministry of Education and Research, the Federal Ministry for Economic Affairs and Energy, and the Federal Ministry of Labour and Social Affairs (n 9) p. 38.

[201] The Public Voice Coalition (n 53) (*See* Principle 1.)

[202] Amnesty International, Access Now (n 55) p. 9.

[203] Organisation for Economic Co-operation and Development (n 54) p. 8 (*See* Principle 1.3); G20 Trade Ministers and Digital Economy Ministers (n 54) p. 11 (*See* Principle 1.3.)

[204] German Federal Ministry of Education and Research, the Federal Ministry for Economic Affairs and Energy, and the Federal Ministry of Labour and Social Affairs (n 10) p. 38.

[205] UK House of Lords, Select Committee on Artificial Intelligence (n 7) p. 27.

[206] European Commission's High-Level Expert Group on Artificial Intelligence (n 6) p.7.

[207] European Commission (n 115) p. 17.

[208] Access Now (n 9) p. 33.

[209] University of Montreal (n 34) p. 12 (*See* Principle 5.9.)

[210] Amnesty International, Access Now (n 55) p. 9.

principle might be interpreted to only require disclosure if an action is taken (e.g. an arrest), whereas the "interaction" principle might require notices that the facial recognition system is in use to be posted in public spaces, much like CCTV signs. Among other glosses on this principle, the European Commission notes that "consideration should be given to when users should be informed on how to reach a human"[211] and the OECD and G20 AI principles call out that that a system of notifications of AI interactions may be especially important "in the workplace."[212]

Regular Reporting

"Regular reporting" as a principle stands for the notion that organizations that implement AI systems should systematically disclose important information about their use. This might include "how outputs are reached and what actions are taken to minimize rights-harming impacts,"[213] "discovery of … operating errors, unexpected or undesirable effects, security breaches, and data leaks,"[214] or the "evaluation of the effectiveness"[215] of AI systems. The regular reporting principle can be interpreted as another implementation mechanism for transparency and explainability, and the OECD and G20 AI principles further call for governments to step in and develop internationally comparable metrics to measure AI research, development, and deployment and to gather the necessary evidence to support these claims.[216]

[211] European Commission (n 116) p. 17

[212] Organisation for Economic Co-operation and Development (n 54) p. 8 (*See* Principal 1.3); G20 Trade Ministers and Digital Economy Ministers (n 54) p. 11 (*See* Principal 1.3.)

[213] Access Now (n 9) p.33.

[214] University of Montreal (n 34) p. 12 (*See* Principle 5.4.)

[215] Amnesty International, Access Now (n 55) p. 49.

[216] Organisation for Economic Co-operation and Development (n 54) p. 9 (*See* Principle 2.5); G20 Trade Ministers and Digital Economy Ministers (n 54) p. 14 (*See* Principle 2.5.)

3.5. Fairness and Non-discrimination

Algorithmic bias – the systemic under- or over-prediction of probabilities for a specific population – creeps into AI systems in a myriad of ways. A system might be trained on unrepresentative, flawed, or biased data.[217] Alternatively, the predicted outcome may be an imperfect proxy for the true outcome of interest[218] or the outcome of interest may be influenced by earlier decisions that are themselves biased. As AI systems increasingly inform or dictate decisions, particularly in sensitive contexts where bias long predates their introduction such as lending, healthcare, and criminal justice, ensuring fairness and non-discrimination is imperative. Consequently, the Fairness and Non-discrimination theme is the most highly represented theme in our dataset, with every document referencing at least one of its six principles: "non-discrimination and the prevention of bias," "representative and high-quality data," "fairness," "equality," "inclusiveness in impact," and "inclusiveness in design."[219]

Within this theme, many documents point to biased data – and the biased algorithms it generates – as the source of discrimination and unfairness in AI, but a few also recognize the role of human systems and institutions in perpetuating or preventing discriminatory or otherwise harmful impacts. Examples of language that focuses on the technical side of bias include the Ground Rules for AI conference paper ("[c]ompanies should strive to avoid bias in A.I. by drawing on diverse data sets")[220] and the Chinese White Paper on AI Standardization ("we should also

PRINCIPLES UNDER THIS THEME

- **89%** Non-discrimination and the Prevention of Bias
- **36%** Representative and High Quality Data
- **56%** Fairness
- **25%** Equality
- **42%** Inclusiveness in Impact
- **47%** Inclusiveness in Design

Percentage reflects the number of documents in the dataset that include each principle

[217] *E.g.*, Jeffrey Dastin, "Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women," Reuters, (Oct. 9, 2018), https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

[218] A bail decision algorithm, for example, may predict for "failure to appear" instead of flight risk to inform decisions about pretrial release. This conflates flight with other less severe causes of nonappearance (i.e. an individual may miss a court date due to inability to access transportation, childcare, or sickness) that may warrant a less punitive, lower-cost intervention than detention.

[219] Fairness and Non-discrimination principles are present in 100% of documents in the dataset.

[220] New York Times' New Work Summit, 'Seeking Ground Rules for AI' (March 2019) principle 5 <https://www.nytimes.com/2019/03/01/business/ethical-ai-recommendations.html>.

be wary of AI systems making ethically biased decisions").[221] While this concern is warranted, it points toward a narrow solution, the use of unbiased datasets, which relies on the assumption that such datasets exist. Moreover, it reflects a potentially technochauvinistic orientation – the idea that technological solutions are appropriate and adequate fixes to the deeply human problem of bias and discrimination.[222] The Toronto Declaration takes a wider view on many places bias permeates the design and deployment of AI systems:

All actors, public and private, must prevent and mitigate against discrimination risks in the design, development and application of machine learning technologies. They must also ensure that there are mechanisms allowing for access to effective remedy in place before deployment and throughout a system's lifecycle.[223]

Within the Fairness and Non-discrimination theme, we see significant connections to the Promotion of Human Values theme, with principles such as "fairness" and "equality" sometimes appearing alongside other values in lists coded under the "Human Values and Human Flourishing" principle.[224] There are also connections to the Human Control of Technology, and Accountability themes, principles under which can act as implementation mechanisms for some of the higher-level goals set by Fairness and Non-discrimination principles.

Non-discrimination and the Prevention of Bias
The "non-discrimination and the prevention of bias" principle articulates that bias in AI – in the training data, technical design choices, or the technology's deployment – should be mitigated to prevent discriminatory impacts. This principle was one of the most commonly included ones in our dataset[225] and, along with others like "fairness" and "equality" frequently operates as a high-level objective for which other principles under this theme (such as "representative and high-quality data" and "inclusiveness in design") function as implementation mechanisms.[226]

Deeper engagement with the principle of "non-discrimination and the prevention of bias" included warnings that AI is not only replicating existing patterns of bias, but also has the potential to significantly scale discrimination and to discriminate in unforeseen ways.[227] Other documents recognized that AI's great capacity for classification and differentiation could and should be proactively used to identify and address discriminatory practices in current systems.[228] The German Government commits to assessing how its current legal protections against discrimination cover – or fail to cover – AI bias, and to adapt accordingly.[229]

Representative and High Quality Data
The principle of "representative and high quality data," driven by what is colloquially referred to as the "garbage in, garbage out" problem, is defined as the use of appropriate inputs to an AI system, which relates accurately to the population of interest. The use of a dataset that is not representative leads to skewed representation of a group in the dataset compared to the actual composition of the target population, introduces bias, and reduces the accuracy of the system's eventual decisions. It is important that the data be high quality and apposite to the context in which the AI system will be deployed, because a representative dataset may nonetheless be informed by historical bias.[230] Some quality measures for data include accuracy, consistency, and validity. As the definition suggests, the documents in our dataset often directly connected this principle to the goal of mitigating the discriminatory impacts of AI.

The Montreal Declaration and the European Charter on AI in judicial systems call for representative and high quality data but state that even using the gold standard in data could be detrimental if the data are used for "deterministic analyses."[231] The Montreal Declaration's articulation of this principle warns against using data "to lock individuals into a user profile, fix their personal identity, or confine them to a filtering bubble, which would restrict and confine their possibilities for personal development."[232] Some documents, including the European Charter on AI in judicial systems, explicitly call for special protections for marginalized groups and for particularly sensitive data, defined as "alleged racial or ethnic origin, socio-economic background, political opinions, religious or philosophical beliefs, trade union membership, genetic data, biometric data, health-related data or data concerning sexual life or sexual orientation."[233]

Fairness
The "fairness" principle was defined as equitable and impartial treatment of data subjects by AI systems. We used this definition, drawn from common usage, over a technical one because articulations of fairness in the documents coded under this principle are not especially technical or overly specific in spite of the rich vein of academic research by AI and machine learning academics around competing mathematical formalizations of fairness.[234] However, Microsoft adds to its principle "AI systems should treat all people fairly" the further elaboration that "industry and academia should continue the promising work underway to develop analytical techniques to detect and address potential unfairness, like methods that systematically assess the data used to train AI systems for appropriate representativeness and document information about its origins and characteristics."[235]

---

[221] Standard Administration of China (n 23) (*See* Principle 3.3.2.)

[222] M. Broussard coined the term "technochauvinism" in her recent book *Artificial Unintelligence*.

[223] Amnesty International, Access Now (n 55) p.6 (*See* Principle 17.)

[224] Organisation for Economic Co-operation and Development (n 54) p. 7 (*See* Principle 1.2.); G20 Trade Ministers and Digital Economy Ministers (n 54) p. 11 (*See* Principle 1.2.)

[225] Only four documents in the dataset did not cite this principle: Asilomar AI Principles, PAI Tenets, U.S. Science and Technology Council report, and Ethically Aligned Design from the IEEE.

[226] European Commission (n 115) p. 13.

[227] Monetary Authority of Singapore (n 37) p. 6 (stating: "While the use of AIDA [Artificial Intelligence and Data Analytics] could enable analysis based on segmentation and clustering of data, this also means that differentiation between groups could take place at a greater scale and faster speed. The use of AIDA may also create the ability to identify or analyse new types of differentiation that could not previously be done. This could perpetuate cases of unjustified differentiation at a systemic level if not properly managed."); *See also*, Mission assigned by the French Prime Minister (n 7) pp. 121-122.

[228] Council of Europe, European Commission for the Efficiency of Justice (n 73) pp.9-10 (stating: "However, the use of machine learning and multidisciplinary scientific analyses to combat such discrimination should be encouraged.")

[229] German Federal Ministry of Education and Research, the Federal Ministry for Economic Affairs and Energy, and the Federal Ministry of Labour and Social Affairs (n 9) p.37.

[230] For example, a lending algorithm trained on a dataset of previously successful applicants will be "representative" of the historical applicant pool but will also replicate any past biases that informed who received a loan.

[231] Council of Europe, European Commission for the Efficiency of Justice (n 73) p. 9.

[232] University of Montreal (n 34) p.14 (*See* Principle 7.4.)

[233] European Commission's High-Level Expert Group on Artificial Intelligence (n 6).

[234] Arvind Narayanan, "Translation tutorial: 21 fairness definitions and their politics," tutorial presented at the Conference on Fairness, Accountability, and Transparency, (Feb. 23, 2018), <https://www.youtube.com/embed/jIXIuYdnyyk>

[235] Microsoft (n 27) p. 58.

There was general consensus in the documents about the importance of fairness with regard to marginalized populations. For example, the Japanese AI principles include the imperative that "all people are treated fairly without unjustified discrimination on the grounds of diverse backgrounds such as race, gender, nationality, age, political beliefs, religion, and so on."[236] Similarly, the Chinese AI Industry Code of Conduct states that "[t]he development of artificial intelligence should ensure fairness and justice, avoid bias or discrimination against specific groups or individuals, and avoid placing disadvantaged people at a more unfavorable position."[237] The European High Level Expert Group guidelines term this the "substantive dimension" of fairness, and also point to a "procedural dimension of fairness [which] entails the ability to contest and seek effective redress against decisions made by AI systems and by the humans operating them," which we coded under the "ability to appeal" principle in the Accountability theme.

Equality

The principle of "equality" stands for the idea that people, whether similarly situated or not, deserve the same opportunities and protections with the rise of AI technologies. "Equality" is similar to "fairness" but goes farther, because of fairness's focus on similar outcomes for similar inputs. As the European High Level Expert Group guidelines puts it:

> "Equality of human beings goes beyond non-discrimination, which tolerates the drawing of distinctions between dissimilar situations based on objective justifications. In an AI context, equality entails that the same rules should apply for everyone to access to information, data, knowledge, markets and a fair distribution of the value added being generated by technologies."[238]

There are essentially three different ways that equality is represented in the documents in our dataset: in terms of human rights, access to technology, and guarantees of equal opportunity through technology. In the human rights framing, the Toronto Declaration notes that AI will pose "new challenges to equality" and that "[s]tates have a duty to take proactive measures to eliminate discrimination."[239] In the access to technology framing, documents emphasize that all people deserve access to the benefits of AI technology, and that systems should be designed to facilitate that broad access.[240]

Documents that take on what we have termed the guarantees of equal opportunity framing go a bit farther in their vision for how AI systems may or should implement equality. The Montreal Declaration asserts that AI systems "must help eliminate relationships of domination between groups and people based on differences of power, wealth, or knowledge" and "must produce social and economic benefits for all by reducing social inequalities and vulnerabilities."[241] This framing makes clear the relationship between the "equality" principle and the principles of "non-discrimination and the prevention of bias" and "inclusiveness in impact."

[236] Japanese Cabinet Office, Council for Science, Technology and Innovation (n 20) p. 10.

[237] Artificial Intelligence Industry Alliance (n 111) (*See* Principle 3, English translation available upon request.)

[238] European Commission's High-Level Expert Group on Artificial Intelligence (n 6) p. 7.

[239] Amnesty International, Access Now (n 55) pp. 5, 10.

[240] European Commission's High-Level Expert Group on Artificial Intelligence (n 6) p. 18.

[241] University of Montreal (n 34) p. 13 (*See* Principles 6.2 and 6.3.)

Inclusiveness in Impact

"Inclusiveness in impact" as a principle calls for a just distribution of AI's benefits, particularly to populations that have historically been excluded. There was remarkable consensus in the language that documents employed to reflect this principle, including concepts like "shared benefits" and "empowerment":

Document	Language of principle
Asilomar AI Principles	Shared Benefit: AI technologies should benefit and empower as many people as possible.[242]
Microsoft's AI principles	Inclusiveness – AI systems should empower everyone and engage people. If we are to ensure that AI technologies benefit and empower everyone, they must incorporate and address a broad range of human needs and experiences. Inclusive design practices will help system developers understand and address potential barriers in a product or environment that could unintentionally exclude people. This means that AI systems should be designed to understand the context, needs and expectations of the people who use them.[243]
Partnership on AI Tenets	We will seek to ensure that AI technologies benefit and empower as many people as possible[244]
Smart Dubai AI principles	We will share the benefits of AI throughout society: AI should improve society, and society should be consulted in a representative fashion to inform the development of AI[245]
T20 report on the future of work and education	Benefits should be shared: AI should benefit as many people as possible. Access to AI technologies should be open to all countries. The wealth created by AI should benefit workers and society as a whole as well as the innovators.[246]
UNI Global Union's AI principles	Share the Benefits of AI Systems: AI technologies should benefit and empower as many people as possible. The economic prosperity created by AI should be distributed broadly and equally, to benefit all of humanity.[247]

The European High Level Expert Group guidelines add some detail around what "benefits" might be shared: "AI systems can contribute to wellbeing by seeking achievement of a fair, inclusive and peaceful society, by helping to increase citizen's mental autonomy, with equal distribution of economic, social and political opportunity."[248] There is a clear connection to the principles we have catalogued under the Promotion of Human Values theme, especially the principle of "leveraged to benefit society."

[242] Future of Life Institute (n 89) (*See* Principle 14.)

[243] Microsoft (n 26) p. 69.

[244] Partnership on AI (n 93) (Principle 1.)

[245] Smart Dubai (n 22) p. 11.

[246] Think 20 (n 38) p. 7

[247] UNI Global Union (n 65) p. 8 (*See* Principle 6.)

[248] European Commission's High-Level Expert Group on Artificial Intelligence (n 6) p. 9.

Inclusiveness in Design
The "inclusiveness in design" principle stands for the idea that ethical and rights-respecting AI requires more diverse participation in the development process for AI systems. This principle is expressed in two different ways. The first and more common interpretation calls for diverse AI design teams. For example, the AI for Europe document from the European Commission affirms that "More women and people of diverse backgrounds, including people with disabilities, need to be involved in the development of AI, starting from inclusive AI education and training, in order to ensure that AI is non-discriminatory and inclusive."[249] The European High Level Expert Group guidelines add that "Ideally, teams are not only diverse in terms of gender, culture, age, but also in terms of professional backgrounds and skill sets."[250]

The second interpretation holds that a broad cross-section of society should have the opportunity to weigh in on what we use AI for and in what contexts; specifically, that there should be "a genuinely diverse and inclusive social forum for discussion, to enable us to democratically determine which forms of AI are appropriate for our society."[251] The Toronto Declaration emphasizes the importance of including end users in decisions about the design and implementation of AI in order to "ensure that systems are created and used in ways that respect rights – particularly the rights of marginalised groups who are vulnerable to discrimination."[252] This interpretation is similar to the Multistakeholder Collaboration principle in our Professional Responsibility category, but it differs in that it emphasizes bringing into conversation all of society – specifically those most impacted by AI – and not just a range of professionals in, for example, industry, government, civil society organizations, and academia.

[249] European Commission (n 116) p. 13.

[250] European Commission's High-Level Expert Group on Artificial Intelligence (n 6) p. 23.

[251] Mission assigned by the French Prime Minister (n 7) p. 114.

[252] Amnesty International, Access Now (n 55) p. 6 (*See* Principle 18.)

3.6. Human Control of Technology

From prominent Silicon Valley magnates' concerns about the Singularity to popular science fiction dystopias, our society, governments, and companies alike are grappling with a potential shift in the locus of control from humans to AI systems. Thus, it is not surprising that Human Control of Technology is a strong theme among the documents in our dataset,[253] with significant representation for the three principles that fall under it: "human review of automated decision," "ability to opt out of automated decision," and "human control of technology (other/general)."

There are connections between the principles in the Human Control of Technology theme and a number of other themes, because human involvement is often presented as a mechanism to accomplish those ends. Human control can facilitate objectives within the themes of Safety and Security, Transparency and Explainability, Fairness and Non-discrimination, and the Promotion of Human Values. For example, the OECD and G20 AI principles refer to human control as a "safeguard"[254] and UNI Global Union claims that transparency in both decisions and outcomes requires "the right to appeal decisions made by AI/algorithms, and having it reviewed by a human being."[255]

PRINCIPLES UNDER THIS THEME

- **33%** Human Review of Automated Decision
- **8%** Ability to Opt out of Automated Decisions
- **64%** Human Control of Technology (Other/General)

Percentage reflects the number of documents in the dataset that are included each principle

Human Review of Automated Decision
The principle of "human review of automated decision" stands for the idea that where AI systems are implemented, people who are subject to their decisions should be able to request and receive human review of those decisions. In contrast to other principles under this theme, the "human review of automated decision" principle is always ex post in is implementation, providing the opportunity to remedy an objectionable result. Although the documents in our dataset are situated in a variety of contexts, there is remarkable commonality between them in the articulation of this principle. The underlying rationale, when explicit, is that "Humans interacting with AI systems must be able to keep full and effective self-determination over themselves."[256]

[253] Human Control of Technology principles are present in 69% of documents in the dataset, with the Human Control of Technology (Other/General) principle most strongly represented.

[254] Organisation for Economic Co-operation and Development (n 54) p. 7 (*See* Principle 1.2); G20 Trade Ministers and Digital Economy Ministers (n 54) p. 11 (*See* Principle 1.2.)

[255] UNI Global Union (n 65) p. 7 (*See* Principle 1.)

[256] European Commission's High-Level Expert Group on Artificial Intelligence (n 6) p. 12.

The most salient differences among the documents are in the breadth of circumstances in which they suggest that human review is appropriate, and the strength of the recommendation. Many of the documents apply the principle of human review in all situations in which an AI system is used, but a handful constrain its application to situations in which the decision is "significant."[257] Further, the principles generally present human review as desirable, but two documents, the Access Now report and the Public Voice Coalition AI guidelines, articulate it as a right of data subjects. The European Charter on AI in judicial systems also contains a strong version of the human review principle, specifying that if review is requested, the case should be heard by a competent court.[258]

Ability to Opt out of Automated Decision

The "ability to opt out of automated decision" principle is defined, as its title suggests, as affording individuals the opportunity and choice not to be subject to AI systems where they are implemented. The AI in the UK document explains its relevance by saying:

> "It is important that members of the public are aware of how and when artificial intelligence is being used to make decisions about them, and what implications this will have for them personally. This clarity, and greater digital understanding, will help the public experience the advantages of AI, as well as to opt out of using such products should they have concerns."[259]

Of course, individuals interact with AI systems in numerous ways: their information may be used as training data; they may be indirectly impacted by systemic deployments of AI, and they may be personally subject to automated decisions. Perhaps because these principles are articulated with relative brevity, or perhaps because of the significant challenges in implementation, only three documents contained this principle: AI in the UK, the European High Level Expert Group guidelines, and the Smart Dubai AI principles. All documents articulated this principle as a natural corollary of the right to notification when interacting with an AI system. The latter two documents disagree about the extent of the principle's implementation, with Smart Dubai saying that entities should "consider" providing the ability to opt out "where appropriate"[260] and the European document standing for a "meaningful opportunity for human choice."[261]

Human Control of Technology (Other/General)

The "human control of technology (other/general)" principle requires that AI systems are designed and implemented with the capacity for people to intervene in their actions. This was the most commonly referenced principle[262] under the theme of Human Control of Technology, and most of the documents that included it framed it broadly, as in our definition. The Asilomar AI principles' version is illustrative: "Humans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives."[263] Where the documents included a theoretical grounding for this principle, it was typically the preservation of human autonomy. For example, the Montreal Declaration states that AI systems should be built and used "respecting people's autonomy, and with the goal of increasing people's control over their lives and their surroundings."[264]

Numerous documents emphasize the importance not only of human-chosen objectives, which were included in the Asilomar principle, but the Promotion of Human Values and human quality of life.[265]Telefónica's AI Principles require that their uses of AI "be driven by value-based considerations"[266] and IA Latam's principles state that the use of AI should not only be under human control but be for the common good.[267] Others focus on the stemming the capacity of AI systems to be used to manipulate[268] or mislead people.[269]

A number of private sector principles stand out for their more granular versions of this principle, which demonstrate some connection with the theme of Professional Responsibility, because they are addressed quite directly to the developers and users of AI tools. Microsoft's AI principles include multiple steps to ensure human control, including "[e]valuation of when and how an AI system should seek human input during critical situations, and how a system controlled by AI should transfer control to a human in a manner that is meaningful and intelligible."[270] The IBM AI principles remind developers that they must identify and design for other users. The policy notes that they "may not have control over how data or a tool will be used by user, client, other external source."[271] Telia's AI principles state that the company "monitor[s] AI solutions so that we are continuously ready to intervene."[272]

Finally, emphasizing the role of people in the process in a different way, UNI Global Union asserts that AI systems must maintain "the legal status of tools, and legal persons [must] retain control over, and responsibility for, these machines at all times."[273] The Public Voice Coalition's principle of human control extends perhaps the farthest, explicitly stating that an institution has an obligation to terminate an AI system if they are no longer able to control it.[274]

[257] Smart Dubai (n 23) p. 9.

[258] Council of Europe, European Commission for the Efficiency of Justice (n 73) p. 12.

[259] UK House of Lords, Select Committee on Artificial Intelligence (n 8) p. 27.

[260] Smart Dubai (n 23) p. 26.

[261] European Commission's High-Level Expert Group on Artificial Intelligence (n 6) p. 12.

[262] 64% of documents included the "human control of technology (other/general)" principle.

[263] Future of Life Institute (n 89) (*See* Principle 16.)

[264] University of Montreal (n 34) p. 9 (*See* Principle 2.)

[265] Japanese Cabinet Office, Council for Science, Technology and Innovation (n 20) p. 4.

[266] Telefónica (n 62) (*See* Principle 3.)

[267] IA Latam (n 22) (*See* Principle 1, English translation available upon request.)

[268] European Commission's High-Level Expert Group on Artificial Intelligence (n 6) p. 12.

[269] University of Montreal (n 34) p. 9 (*See* Principle 2.)

[270] Microsoft (n 27) p. 65.

[271] IBM (n 24) p. 18.

[272] Telia Company (n 56) p. 3 (*See* Principle 4.)

[273] UNI Global Union (n 65) p. 8 (*See* Principle 4.)

[274] The Public Voice Coalition (n 53) (*See* Principle 12.)

3.7. Professional Responsibility

The theme of Professional Responsibility brings together principles that are targeted at individuals and teams who are responsible for designing, developing, or deploying AI-based products or systems. These principles reflect an understanding that the behavior of such professionals, perhaps independent of the organizations, systems, and policies that they operate within, may have a direct influence on the ethics and human rights impacts of AI. The theme of Professional Responsibility was widely represented in our dataset[275] and consists of five principles: "accuracy," "responsible design," "consideration of long-term effects," "multistakeholder collaboration," and "scientific integrity."

There are significant connections between the Professional Responsibility theme and the Accountability theme, particularly with regard to the principle of "accuracy." Articulations of the principle of "responsible design" often connect with the theme of Promotion of Human Values, and sometimes suggest Human Control of Technology as an aspect of this objective.

PRINCIPLES UNDER THIS THEME

- **19%** Accuracy
- **44%** Responsible Design
- **33%** Consideration of Long Term Effects
- **64%** Multi-stakeholder Collaboration
- **6%** Scientific Integrity

Percentage reflects the number of documents in the dataset that include each principle

Accuracy

The principle of "accuracy" is usefully defined by the European High Level Expert Group guidelines, which describe it as pertaining "to an AI's confidence and ability to correctly classify information into the correct categories, or its ability to make correct predictions, recommendations, or decisions based on data or models."[276] There is a split among the documents, with some understanding "accuracy" as a goal and others as an ongoing process.

The Google AI principles are focused narrowly on the goal of preventing the use of AI in the creation and dissemination of false information, making "accurate information readily available"[277] and the Montreal Declaration similarly avers that AI "should be designed with a view to containing [the] dissemination" of "untrustworthy information."[278] By contrast, the European High Level Expert Group guidelines are emblematic of the process-based approach, recommending that developers establish an internal definition of "accuracy" for the use case; develop a method of measurement; verify the harms caused by inaccurate predictions and measure the frequency of such predictions; and finally institute a "series of steps to increase the system's accuracy."[279] In cases when inaccurate predictions cannot be avoided, these guidelines suggest that systems indicate the likelihood of such errors.[280] Relying on a similar understanding of accuracy, the IEEE recommends operators measure the effectiveness of AI systems through methods that are "valid and accurate, as well as meaningful and actionable."[281]

The principle of accuracy is frequently referred to alongside the similar principle of "verifiability and replicability" under the Accountability theme. The Public Voice Coalition, for instance, recommends that institutions must ensure the "accuracy, reliability, and validity of decisions."[282] The two can be distinguished as "accuracy" is targeted at developers and users, promoting careful attention to detail on their part. By contrast, the principle of replicability focuses on the technology, asking whether an AI system delivers consistent results under the same conditions, facilitating post-hoc evaluation by scientists and policymakers.

Responsible Design

The principle of "responsible design" stands for the notion that individuals must be conscientious and thoughtful when engaged in the design of AI systems. Indeed, even as the phrasing of this principle might differ from document to document, there is a strong consensus that professionals are in a unique position to exert influence on the future of AI. The French AI strategy emphasizes the crucial role that researchers, engineers and developers play as "architects of our digital society."[283] This document notes that professionals play an especially important part in emerging technologies since laws and norms cannot keep pace with code and cannot solve for every negative effect that the underlying technology may bring about.[284]

The Partnership on AI Tenets prompt research and engineering communities to "remain socially responsible, and engage directly with the potential influences of AI technologies on wider society."[285] This entails, to some degree, an obligation to become informed about society, which other documents address directly. The IBM AI principles require designers and developers not only to encode values that are sensitive to different contexts but also to engage in collaboration to better recognize existing values.[286] The Tencent and Microsoft AI principles capture this idea by calling for developers to ensure that design

[275] Professional Responsibility principles are present in 78% of documents in the dataset.

[276] European Commission's High-Level Expert Group on Artificial Intelligence (n 6) p. 17.

[277] Google (n 22) (*See* Principle 1.)

[278] University of Montreal (n 34) p.9 (*See* Principle 2.5.)

[279] European Commission's High-Level Expert Group on Artificial Intelligence (n 6) p. 17.

[280] European Commission's High-Level Expert Group on Artificial Intelligence (n 6) p. 17.

[281] IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (n 5) p. 25 (*See* Principle 4.)

[282] The Public Voice Coalition (n 53) (*See* Principle 6.)

[283] Mission assigned by the French Prime Minister (n 8) p. 114.

[284] Mission assigned by the French Prime Minister (n 8) p. 114.

[285] Partnership on AI (n 93) (*See* Principle 6.)

[286] IBM (n 24) p. 22.

is "aligned with human norms in reality"[287] and to involve domain experts in the design and deployment of AI systems.[288] We note a rare interaction among the documents when the Indian AI strategy recommends that evolving best practices such as the recommendations by the Global Initiative on Ethics of Autonomous and Intelligent Systems by IEEE be incorporated in the design of AI systems.[289]

Consideration of Long Term Effects

The principle of "consideration of long term effects" is characterized by deliberate attention to the likely impacts, particularly distant future impacts, of an AI technology during the design and implementation process. The documents that address this principle largely view the potential long-term effects of AI in a pluralistic manner. For instance, the German AI strategy highlights that AI is a global development and policymakers will need to "think and act globally" while considering its impact during the development stage[290]; and the Asilomar principles recognize that highly-developed AI must be for the benefit of all of humanity and not any one sub-group.[291] The Montreal Declaration recommends that professionals must anticipate the increasing risk of AI being misused in the future and incorporate mechanisms to mitigate that risk.[292]

Some of the documents base their articulations of this principle on the premise that AI capabilities in the future may be vastly advanced compared to the technology we know today. The Beijing AI principles recommend that research on potential risks arising out of augmented intelligence, artificial general intelligence and superintelligence be encouraged.[293] These documents take the position that possibility of catastrophic or existential risks arising out of AI systems in the future cannot not be ruled out and professionals must work towards avoiding or mitigating such impacts.[294]

Multistakeholder Collaboration

"Multistakeholder collaboration" is defined as encouraging or requiring that designers and users of AI systems consult relevant stakeholder groups while developing and managing the use of AI applications. This was the most commonly included of the principles under Professional Responsibility.[295] Broadly, the documents reflect either a tool-specific or a general policy vision for multistakeholderism.

The IBM AI principles are emblematic of a tool-specific vision, specifying that developers should try to consult with policymakers and academics as they build AI systems to bring in different perspectives.[296] Additionally, the principles recommend that a feedback loop or open dialogue be established with users allowing them to highlight biases or other on-ground challenges that the system might bring about once deployed.[297] The Toronto Declaration calls for meaningful consultation with users and especially marginalized groups during the design and application of machine learning systems.[298] Access Now also suggests that human rights organizations and independent human rights and AI experts be included during such consultations.[299]

Documents that espouse a general policy function for multistakeholderism call for collaboration across the globe, rather than around any particular tool. Participants may be drawn from universities, research institutions, industry, policymaking, and the public at large to examine AI developments across sectors and use cases. The Japanese and Chinese AI strategies, for instance, push for international cooperation on AI research and use, to build a "non-regulatory, non-binding" framework.[300] This interpretation of multistakeholderism is focused on the utility of building a normative consensus on the governance of AI technologies. This vision is also seen as a policy vehicle through which governments can educate and train their populations to ensure an easy transition and safety as labor markets continue to modernize.[301]

Scientific Integrity

The principle of "scientific integrity" means that those who build and implement AI systems should be guided by established professional values and practices. Interestingly, both documents that include this relatively little-mentioned principle are organizations driven at least in significant part by engineers and technical experts. Google's AI principles recognize scientific method and excellence as the bedrock for technological innovation, including AI. The company makes a commitment to honor "open inquiry, intellectual rigor, integrity, and collaboration" in its endeavors.[302] The IEEE acknowledges the idea of scientific rigor in its call for creators of AI systems to define metrics, make them accessible, and measure systems.[303]

[287] Tencent Institute (n 58) (*See* Principle 12, English translation available upon request.)

[288] Microsoft (n 27) p. 65.

[289] Niti Aayog (n 24) p. 87.

[290] German Federal Ministry of Education and Research, the Federal Ministry for Economic Affairs and Energy, and the Federal Ministry of Labour and Social Affairs (n 10) p. 40.

[291] Future of Life Institute (n 89) (*See* Principle 23.)

[292] University of Montreal (n 34) p. 15 (See Principle 8.)

[293] Beijing Academy of Artificial Intelligence (n 23) (*See* Principle 3.5, English translation available upon request.)

[294] Beijing Academy of Artificial Intelligence (n 23) (*See* Principle 3.5, English translation available upon request.)

[295] 64% of documents included it in one form of another.

[296] IBM (n 24) p. 24.

[297] IBM (n 24) p. 36.

[298] Amnesty International, Access Now (n 56) p. 6.

[299] Access Now (n 10) p. 34.

[300] Japanese Cabinet Office, Council for Science, Technology and Innovation (n 20) p. 6.

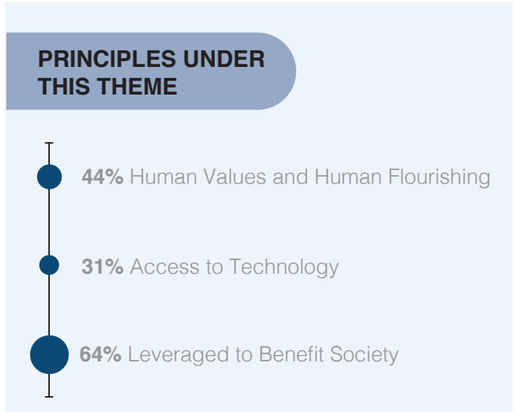
[301] *See* e.g., Organisation for Economic Co-operation and Development (n 54) p. 9 (*See* Principle 2.4.); G20 Trade Ministers and Digital Economy Ministers (n 54) p. 13 (*See* Principle 2.4.)

[302] Google (n 22) (*See* Principle 6.)

[303] IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (n 5) p. 25.

3.8. Promotion of Human Values

With the potential of AI to act as a force multiplier for any system in which it is employed, the Promotion of Human Values is a key element of ethical and rights-respecting AI.[304] The principles under this theme recognize that the ends to which AI is devoted, and the means by which it is implemented, should correspond with and be strongly influenced by social norms. As AI's use becomes more prevalent and the power of the technology increases, particularly if we begin to approach artificial general intelligence, the imposition of human priorities and judgment on AI is especially crucial. The Promotion of Human Values category consists of three principles: "human values and human flourishing," "access to technology," and "leveraged to benefit society."



While principles under this theme were coded distinctly from explicit references to human rights and international instruments of human rights law, there is a strong and clear connection. References to human values and human rights were often adjacent to one another, and where the documents provided more specific articulations of human values, they were are largely congruous with existing guarantees found in international human rights law. Moreover, principles that refer to human values often include explicit references to fundamental human rights or international human rights, or mention concepts from human rights frameworks and jurisprudence such as human dignity or autonomy. The OECD and G20 AI principles also add "internationally recognized labor rights" to this list.[305]

There is also an overlap between articulations of the Promotion of Human Values and social, economic, or environmental concepts that are outside the boundaries of political and civil rights,[306] including among documents coded under the principle of AI "leveraged to benefit society." Principle 3, "Make AI Serve People and Planet," from the UNI Global Union's AI principles, is emblematic, calling for: "throughout their entire operational process, AI systems [to] remain compatible and increase the principles of human dignity, integrity, freedom, privacy and cultural and gender diversity, as well as … fundamental human rights. In addition, AI systems must protect and even improve our planet's ecosystems and biodiversity."[307]

Human Values and Human Flourishing

The principle of "human values and human flourishing" is defined as the development and use of AI with reference to prevailing social norms, core cultural beliefs, and humanity's best interests. As the Chinese AI Governance Principles put it, this principle means that AI should "serve the progress of human civilization."[308] This is the broadest of the three principles in the Promotion of Human Values theme and is mentioned in 44 percent of documents. Most documents do not delve especially deeply into what they intend by "human values" beyond references to concepts like self-determination,[309] but the Montreal Declaration contains a somewhat idiosyncratic list, calling for AI systems that "permit the growth of the well-being of all sentient beings" by, inter alia, "help[ing] individuals improve their living conditions, their health, and their working conditions, … allow[ing] people to exercise their mental and physical capacities [and]… not contribut[ing] to increasing stress, anxiety, or a sense of being harassed by one's digital environment."[310]

Many of the documents that refer to the theme of "human values and human flourishing" are concerned with how the societal impacts of AI can be managed through AI system design. Tencent's AI principles state that "The R&D of artificial intelligence should respect human dignity and protect human rights and freedoms."[311] The Smart Dubai AI principles says we should "give AI systems human values and make them beneficial to society,"[312] suggesting that it is possible to build AI systems that have human values embedded in their code.[313] However, most, if not all, of these documents also acknowledge that human values will also need to be promoted in the implementation of AI systems and "throughout the AI system lifecycle."[314]

Access to Technology

The "access to technology" principle represents statements that the broad availability of AI technology, and the benefits thereof, is a vital element of ethical and rights-respecting AI. Given the significant transformational potential of AI, documents that include this principle worry that AI might contribute to the growth of inequality. The ITI AI Policy Principles, a private sector document, focus on the economic aspect, stating that "if the value [created by AI] favors only certain incumbent entities, there is a risk of exacerbating existing wage, income, and wealth gaps."[315] At least one civil society document shares this concern: the T20 report on the future of work and education

[304] Promotion of Human Values principles are present in 69% of documents in the dataset.

[305] Organisation for Economic Co-operation and Development (n 54) p. 7 (*See* Principle 1.2.); G20 Trade Ministers and Digital Economy Ministers (n 54) p. 11 (*See* Principle 1.2.)

[306] *See generally*, Future of Life Institute (n 89); European Commission (n 115).

[307] UNI Global Union (n 66) p. 7 (*See* Principle 3.)

[308] Chinese National Governance Committee for the New Generation Artificial Intelligence, led by China's Ministry of Science and Technology (n 22) (*See* Principle 1, English translation available upon request.)

[309] Think 20 (n 38) p. 7.

[310] University of Montreal (n 34) p. 8 (Principle 1.)

[311] Tencent Institute (n 58) (*See* Principle 1, English translation available upon request.)

[312] Smart Dubai (n 22) p. 10.

[313] One document from our draft dataset that is no longer included in the final version, SAGE's The Ethics of Code: Developing AI for Business with Five Core Principles has a similar Principle as found in the Smart Dubai document, stating in Principle 3: "…Reinforcement learning measures should be built not just based on what AI or robots do to achieve an outcome, but also on how AI and robots align with human values to accomplish that particular result."

[314] Organisation for Economic Co-operation and Development (n 54) p. 7 (See Principle 1.2.); G20 Trade Ministers and Digital Economy Ministers (n 54) p. 11 (*See* Principle 1.2.)

[315] Information Technology Industry Council (n 9) p. 5 (*See* "Democratizing Access and Creating Equality of Opportunity.")

avers that "The wealth created by AI should benefit workers and society as a whole as well as the innovators."[316] The Japanese AI principles, while acknowledging the economic dimension of this issue (observing that "AI should not generate a situation where wealth and social influence are unfairly biased towards certain stakeholders"[317]), emphasize the sociopolitical dimensions of inequality, including the potential that AI may unfairly benefit certain states or regions as well as contribute to "a digital divide with so-called 'information poor' or 'technology poor' people left behind."[318]

Some versions of the "access to technology" principle are premised on the notion that broad access to AI technology itself, as well as the education necessary to use and understand it, is the priority. The Chinese AI governance principles provide that "Stakeholders of AI systems should be able to receive education and training to help them adapt to the impact of AI development in psychological, emotional and technical aspects."[319] The ITI AI Policy Principles focus on educating and training people who have traditionally been marginalized by or excluded from technological innovation, calling for the "diversification and broadening of access to the resources necessary for AI development and use, such as computing resources, education, and training."[320] Two documents, Microsoft's AI Principles and the European High Level Expert Group guidelines, go beyond this to reflect a vision for "[a]ccessibility to this technology for persons with disabilities,"[321] noting that in some cases "AI-enabled services… are already empowering those with hearing, visual and other impairments."[322]

Leveraged to Benefit Society

The principle that AI be "leveraged to benefit society" stands for the notion that AI systems should be employed in service of public-spirited goals. The documents vary in the specificity with which they articulate goals. Where they are specific in the goals they list, they may include social, political, and economic factors. Examples of beneficial ends in the European High Level Expert Group guidelines include: "Respect for human dignity... Freedom of the individual... Respect for democracy, justice and the rule of law... Equality, non-discrimination and solidarity - including the rights of persons at risk of exclusion... Citizens' rights… including the right to vote, the right to good administration or access to public documents, and the right to petition the administration."[323] The High Level Expert Group and the German AI strategy were the two documents to explicitly include the environment and sustainable development as factors in their determination of AI that is "leveraged to benefit society."[324]

There is a notable trend among the documents that include this principle to designate it as a *precondition* for AI development and use. IEEE's Ethically Aligned Design document uses strong language to assert that it is not enough for AI systems to be profitable, safe, and legal; they must also include human well-being as a "primary success criterion for development."[325] Google's AI principles contain a similar concept, stating that the company "will proceed [with the development of AI technology] where we believe that the overall likely benefits substantially exceed the foreseeable risks and downsides" after taking "into account a broad range of social and economic factors."[326]

[316] Think 20 (n 38) p. 7.

[317] Japanese Cabinet Office, Council for Science, Technology and Innovation (n 20) p. 9.

[318] Japanese Cabinet Office, Council for Science, Technology and Innovation (n 20) p. 7.

[319] Chinese National Governance Committee for the New Generation Artificial Intelligence, led by China's Ministry of Science and Technology (n 22) (*See* Principle 2.3., English translation available upon request.)

[320] Information Technology Industry Council (n 9) p. 5 (*See* "Democratizing Access and Creating Equality of Opportunity.")

[321] European Commission's High-Level Expert Group on Artificial Intelligence (n 6) p. 18.

[322] Microsoft (n 27) p. 70.

[323] European Commission's High-Level Expert Group on Artificial Intelligence (n 6) p. 11 (*See* Principle 2.1.)

[324] European Commission's High-Level Expert Group on Artificial Intelligence (n 5) p. 32 (*See* "Example of Trustworthy AI"); German Federal Ministry of Education and Research, the Federal Ministry for Economic Affairs and Energy, and the Federal Ministry of Labour and Social Affairs (n 9) p. 9.

[325] IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (n 5) pp. 21-22 (*See* Principle 2.)

[326] Google (n 23) (*See* Principle 1.)

4. International Human Rights

In recent years, the human rights community has become more engaged with digital rights, and with the impacts of AI technology in particular. Even outside of human rights specialists, there has been an increasing appreciation for the relevance of international human rights law and standards to the governance of artificial intelligence.[327] To an area of technology governance that is slippery and fast-moving, human rights law offers an appealingly well-established core set of concepts, against which emerging technologies can be judged. To the broad guarantees of human rights law, principles documents offer a tailored vision of the specific – and in some cases potentially novel – concerns that AI raises.

Accordingly, when coding the principles documents in our dataset, we also made observations on each document's references to human rights, whether as a general concept or specific human-rights related documents such as the Universal Declaration of Human Rights, International Covenant on Civil and Political Rights, the United Nations Guiding Principles on Business & Human Rights and the United Nations Sustainable Development Goals. Twenty-three of the documents in our dataset (64%) made a reference of this kind. We also noted when documents stated explicitly that they had employed a human rights framework, and five of the thirty-six documents (14%) did so.

Given the increasing visibility of AI in the human rights community and the apparent increasing interest in human rights among those invested in AI governance, we had expected that the data might reveal a trend toward increasing emphasis on human rights in AI principles documents. However, our dataset was small enough, and the timespan sufficiently compressed, that no such trend is apparent.

As illustrated in the table below, private sector and civil society documents were most likely to reference human rights. At the outset of our research, we had expected that principles documents from the private sector would be less likely to refer to human rights and government documents more likely. Among the principles documents we looked at – admittedly not designed to be a complete or even representative sample – we were wrong. The actor type with the single greatest proportion of human rights references were the documents from the private sector; only one omitted a reference to human rights. By contrast, less than half of documents authored by or on behalf of government actors did contain some reference to human rights.[328]

Nature of actor	Number of documents	Number with any reference to human rights	
Civil society	5	4	80%
Government	13	6	46%
Intergovernmental organization	3	2	67%
Multistakeholder initiative	7	4	57%
Private sector	8	7	88%
Total	**36**	**23**	**64%**

There are multiple possible explanations for this. It may be that the agencies or individuals in government who have been tasked with drafting and contributing to principles documents were not selected for their expertise with human rights law, or it may be that national laws, such as the GDPR, are perceived as more relevant.

The documents also exhibit significant variation in the degree to which they are permeated by human rights law, with some using it as the framework of the whole document (denoted by a star in the data visualization), and others merely mentioning it in passing (denoted by a diamond). Using a human rights framework means that the document uses human rights as a basis for further ethical principle for the development and use of AI systems. Only five documents use a human rights framework. Three are civil society documents and two are government documents from the EU: Access Now report, AI for Europe, European High Level Expert Group guidelines, Public Voice Coalition AI guidelines, and Toronto Declaration.

[327] Filippo A. Raso, Hannah Hilligoss, and Vivek Krishnamurthy, 'Artificial Intelligence & Human Rights: Opportunities & Risks', Berkman Klein Center (September 25, 2018) https://cyber.harvard.edu/publication/2018/artificial-intelligence-human-rights.

[328] The government documents were from Europe (France, Germany, European Commission (both documents)), China and Japan.

5. Conclusion

The eight themes that surfaced in this research – Privacy, Accountability, Safety and Security, Transparency and Explainability, Fairness and Non-discrimination, Human Control of Technology, Professional Responsibility, and Promotion of Human Values – offer at least some view into the foundational requirements for AI that is ethical and respectful of human rights. However, there's a wide and thorny gap between the articulation of these high-level concepts and their actual achievement in the real world. While it is the intent of this white paper and the accompanying data visualization to provide a high-level overview, there remains more work to be done, and we close with some reflections on productive possible avenues.

In the first place, our discussion of the forty-seven principles we catalogued should make clear that while there are certainly points of convergence, by no means is there unanimity. The landscape of AI ethics is burgeoning, and if calls for increased access to technology (see Section 3.8) and multistakeholder participation (see Section 3.7) are heeded, it's likely to become yet more diverse. It would be compelling to have closer studies of the variation within the themes we uncovered, including additional mapping projects that might illustrate narrower or different versions of the themes with regard to particular geographies or stakeholder groups. It would also be interesting to look at principles geared toward specific applications of AI, such as facial recognition or autonomous vehicles.

Within topics like "fairness," the varying definitions and visions represented by the principles documents in our dataset layer on top of an existing academic literature,[329] but also on existing domestic and international legal regimes which have long interpreted these and similar concepts. Litigation over the harmful consequences of AI technology is still nascent, with just a handful of cases having been brought. Similarly, only a few jurisdictions have adopted regulations concerning AI, although certainly many of the documents in our dataset anticipate, and even explicitly call for (see Sections 3.1 and 3.2), such actions. Tracking how principles documents engage with and influence how liability for AI-related damages is apportioned by courts, legislatures, and administrative bodies, will be important.

There will be a rich vein for further scholarship on ethical and rights-respecting AI for some time, as the norms we attempt to trace remain actively in development. What constitutes "AI for good" is being negotiated both through top-down efforts such as dialogues at the intergovernmental level, as well as bottom-up, among people most impacted by the deployment of AI technology, and the organizations who represent their interests. That there are core themes to these conversations even now is due to the hard work of the many individuals and organizations who are participating in them, and we are proud to play our part.

[329] Arvind Narayanan, "Translation tutorial: 21 fairness definitions and their politics," tutorial presented at the Conference on Fairness, Accountability, and Transparency, (Feb 23 2018), available at: https://www.youtube.com/embed/jlXIuYdnyyk

6. Bibliography

Access Now, 'Human Rights in the Age of Artificial Intelligence' (2018) <https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf>
Amnesty International, Access Now, 'Toronto Declaration: Protecting the Right to Equality and Non-Discrimination in Machine Learning Systems' (2018) <https://www.accessnow.org/cms/assets/uploads/2018/08/The-Toronto-Declaration_ENG_08-2018.pdf>

Artificial Intelligence Industry Alliance, 'Artificial Intelligence Industry Code of Conduct (Consultation Version)' (2019) <https://www.secrss.com/articles/11099>

Beijing Academy of Artificial Intelligence, 'Beijing AI Principles' (2019) <https://www.baai.ac.cn/blog/beijing-ai-principles?categoryId=394>

British Embassy in Mexico City, 'Artificial Intelligence in Mexico (La Inteligencia Artificial En México)' (2018) <https://docs.wixstatic.com/ugd/7be025_ba24a518a53a4275af4d7ff63b4cf594.pdf>

Chinese National Governance Committee for the New Generation Artificial Intelligence, led by China's Ministry of Science and Technology, 'Governance Principles for a New Generation of Artificial Intelligence: Develop Responsible Artificial Intelligence' (2019) <http://www.chinadaily.com.cn/a/201906/17/WS5d07486ba3103dbf14328ab7.html>

Council of Europe, European Commission for the Efficiency of Justice, 'European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment' (2018) <https://rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699c>

European Commission, 'Artificial Intelligence for Europe: Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee, and the Committee of the Regions' COM (2018) 237 <https://ec.europa.eu/digital-single-market/en/news/communication-artificial-intelligence-europe>

European Commission's High-Level Expert Group on Artificial Intelligence, 'Ethics Guidelines for Trustworthy AI' (2018) <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>

Future of Life Institute, 'Asilomar AI Principles' (2017) <https://futureoflife.org/ai-principles/?cn-reloaded=1>

G20 Trade Ministers and Digital Economy Ministers, 'G20 Ministerial Statement on Trade and Digital Economy' (2019) <https://www.mofa.go.jp/files/000486596.pdf>

German Federal Ministry of Education and Research, the Federal Ministry for Economic Affairs and Energy, and the Federal Ministry of Labour and Social Affairs, 'Artificial Intelligence Strategy' (2018) <https://www.ki-strategie-deutschland.de/home.html>

Google, 'AI at Google: Our Principles' (2018) <https://www.blog.google/technology/ai/ai-principles/>

IA Latam, 'Declaración de Principios Éticos Para La IA de Latinoamérica' (2019) <http://ia-latam.com/etica-ia-latam/>

IBM, 'IBM Everyday Ethics for AI' (2019) <https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf>

IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 'Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems' (2019) First Edition <https://ethicsinaction.ieee.org/>

Information Technology Industry Council, 'AI Policy Principles' (2017) <https://www.itic.org/resources/AI-Policy-Principles-FullReport2.pdf>

Japanese Cabinet Office, Council for Science, Technology and Innovation, 'Social Principles of Human-Centric Artificial Intelligence' (2019) <https://www8.cao.go.jp/cstp/english/humancentricai.pdf>
Microsoft, 'AI Principles' (2018) <https://www.microsoft.com/en-us/ai/our-approach-to-ai>

Mission assigned by the French Prime Minister, 'For a Meaningful Artificial Intelligence: Toward a French and European Strategy' (2018) <https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf>
Monetary Authority of Singapore, 'Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore's Financial Sector' (2019) <http://www.mas.gov.sg/~/media/MAS/News%20and%20Publications/Monographs%20and%20Information%20Papers/FEAT%20Principles%20Final.pdf>

New York Times' New Work Summit, 'Seeking Ground Rules for AI' (March 2019) <https://www.nytimes.com/2019/03/01/business/ethical-ai-recommendations.html>

Niti Aayog, 'National Strategy for Artificial Intelligence: #AI for All (Discussion Paper)' (2018) <https://www.niti.gov.in/writereaddata/files/document_publication/NationalStrategy-for-AI-Discussion-Paper.pdf>

Organisation for Economic Co-operation and Development, 'Recommendation of the Council on Artificial Intelligence' (2019) <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>

Partnership on AI, 'Tenets' (2016) <https://www.partnershiponai.org/tenets/>

Smart Dubai, 'Artificial Intelligence Principles and Ethics' (2019) <https://smartdubai.ae/initiatives/ai-principles-ethics>

Standard Administration of China, 'White Paper on Artificial Intelligence Standardization' *excerpts in English published by New America* (January 2018) <https://www.newamerica.org/cybersecurity-initiative/digichina/blog/translation-excerpts-chinas-white-paper-artificial-intelligence-standardization/>

Telefónica, 'AI Principles of Telefónica' (2018) <https://www.telefonica.com/en/web/responsible-business/our-commitments/ai-principles>

Telia Company, 'Guiding Principles on Trusted AI Ethics' (2019) <https://www.teliacompany.com/globalassets/telia-company/documents/about-telia-company/public-policy/2018/guiding-principles-on-trusted-ai-ethics.pdf>

Tencent Institute, 'Six Principles of AI' (2017) <http://www.kejilie.com/iyiou/article/ZRZFn2.html>

The Public Voice Coalition, 'Universal Guidelines for Artificial Intelligence' (2018) <https://thepublicvoice.org/ai-universal-guidelines/>

Think 20, 'Future of Work and Education for the Digital Age' (2018) <https://www.g20-insights.org/wp-content/uploads/2018/07/TF1-1-11-Policy-Briefs_T20ARG_Towards-a-G20-Framework-For-Artificial-Intelligence-in-the-Workplace.pdf>

UK House of Lords, Select Committee on Artificial Intelligence, 'AI in the UK: Ready, Willing and Able?' (2018) Report of Session 2017-19 <https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf>

UNI Global Union, 'Top 10 Principles for Ethical Artificial Intelligence' (2017) <http://www.thefutureworldofwork.org/media/35420/uni_ethical_ai.pdf>

United States Executive Office of the President, National Science and Technology Council Committee on Technology, 'Preparing for the Future of Artificial Intelligence' (2016) <https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf>

University of Montreal, 'Montreal Declaration for a Responsible Development of Artificial Intelligence' (2018) <https://www.montrealdeclaration-responsibleai.com/the-declaration>

cyber.harvard.edu

OECD publishing

SCOPING THE OECD AI PRINCIPLES

DELIBERATIONS OF THE EXPERT GROUP ON ARTIFICIAL INTELLIGENCE AT THE OECD (AIGO)

OECD DIGITAL ECONOMY PAPERS
November 2019 **No. 291**

OECD
BETTER POLICIES FOR BETTER LIVES

Foreword

This document presents the work conducted by the Expert Group on Artificial Intelligence at the OECD (AIGO) to scope principles to foster trust in and adoption of artificial intelligence (AI), as requested by the Committee on Digital Economy Policy (CDEP). The work was developed over four in-person meetings and several teleconference calls in-between those meetings. The group concluded its discussion and agreed on this draft at its fourth and last meeting in Dubai, UAE, on 8-9 February.

This paper was approved and declassified by the CDEP on 1 July 2019 and prepared for publication by the OECD Secretariat. The description of what is an AI system and the AI system lifecycle informed the CDEP's discussion of a draft Recommendation of the Council on Artificial Intelligence on 14-15 March 2019. The OECD Council adopted this Recommendation at its 1397th Session on 22 May 2019.

This document was a contribution to IOR 1.3.1.1.3 Artificial Intelligence of the 2019-2020 Programme of Work of the CDEP. For more information, please visit www.oecd.ai.

Note to Delegations:

This document is also available on O.N.E under the reference code:

DSTI/CDEP(2019)1/FINAL

Table of contents

Foreword 2

Background 4

What is an 'AI system' 5

A Practical Reference Framework for the AI System Lifecycle 12

Annex A. Scoping principles to foster trust in and adoption of AI 18

Introduction 19
Common understanding of technical terms for the purposes of these principles 20
Principles for responsible stewardship of trustworthy AI 20
National policies for trustworthy AI 22

Annex B. List of AIGO members 25

Figures

Figure 1. A high-level conceptual view of an AI system 6
Figure 2. Detailed conceptual view of an AI System 6
Figure 3. Linking the AI System to the General Principles 9
Figure 4. Areas of the AI system in which biases can appear 10
Figure 5. AI system lifecycle 13
Figure 6. Stakeholders view of AI principles, in the framework of the AI lifecycle 14
Figure 7. AI risk-based management approach 16
Figure 8. A view of fairness considerations by AI actors within the AI system lifecycle 17

Background

In the context of its work on Artificial Intelligence (AI), the Committee on Digital Economy Policy (CDEP) agreed, at its meeting on 16-18 May 2018, to form an expert group on AI to scope principles to foster trust in and adoption of AI in society, in view of developing a Council Recommendation in the course of 2019 [DSTI/CDEP/M(2018)1, Item 10].

The group, AIGO, comprised over 50 experts from different disciplines and different sectors (government, industry, civil society, academia and the technical community; see also Annex B: List of AIGO members). AIGO held four meetings: two meetings in Paris in September and November 2018, one at MIT in January 2019, and a last meeting in Dubai, early February 2019.

Chaired by the CDEP Chair, Mr Wonki Min[1], the group's objective was to scope principles with the following characteristics: specific to AI, facilitating innovation and trust in AI, implementable, flexible to stand the test of time, and conducive to increased co-operation. At each meeting, the group discussed proposals for the principles, revised by the Secretariat based on oral input from the previous discussion and on written input. The group also formed two subgroups, to discuss and clarify particular technical aspects, namely, articulating a common understanding of "AI systems" (Chapter 1) and of the AI system lifecycle (Chapter 2). The work benefited from the diligence, engagement and substantive contributions of its members, as well as from their multi-stakeholder and multidisciplinary backgrounds.

At its meeting in Dubai on 8-9 February 2019, the group agreed on its final proposal to the Committee, which included five value-based principles that AI should promote, four recommendations for national AI policies, and a principle on international cooperation for trustworthy AI (Annex A). These principles aim to apply globally to all stakeholders and throughout the entire AI life cycle.

The group's proposal was submitted to the Committee to inform its discussion of a draft Recommendation of the Council on Artificial Intelligence on 14-15 March 2019. It was subsequently adopted by the OECD Council at Ministerial level on 22 May 2019.

What is an 'AI system'

In November 2018, AIGO set up a subgroup to develop a description of an AI system, in view of delineating the scope of applicability of the OECD Principles. This chapter details the high-level description of an AI system provided in the Principles. The description aims to be understandable, technically accurate, technology-neutral, and applicable to short and long-term time horizons. It is broad enough to encompass many of the definitions of AI commonly used by the scientific, business and policy communities.

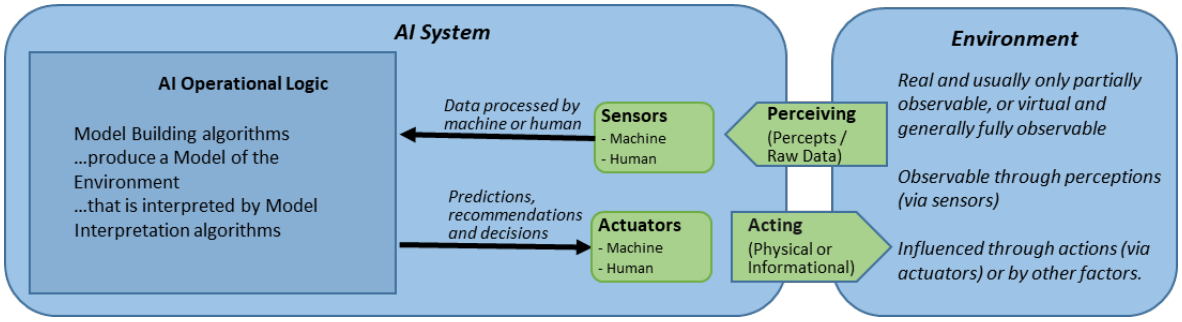
Twenty-one AI experts participated in the work of the subgroup, which was co-moderated by Mr. Marko Grobelnik from Slovenia and by Mr. Javier Juarez Mojica from Mexico. Mr. Marko Grobelnik authored the present document with input from the subgroup that met regularly from mid-December 2018 to mid-February 2019 and from the Secretariat.

Conceptual view of an AI system

The present description of what is an AI system is based on the conceptual view of AI detailed in "Artificial Intelligence: A Modern Approach" (Russel, S. & Norvig, P., 2009[1]). This view is consistent with a widely-used definition of AI as "the study of the computations that make it possible to perceive, reason, and act" (Winston, 1992[2]) and with similar general definitions (Gringsjord, S. & Govindarajulu, N.S., 2018[3]).

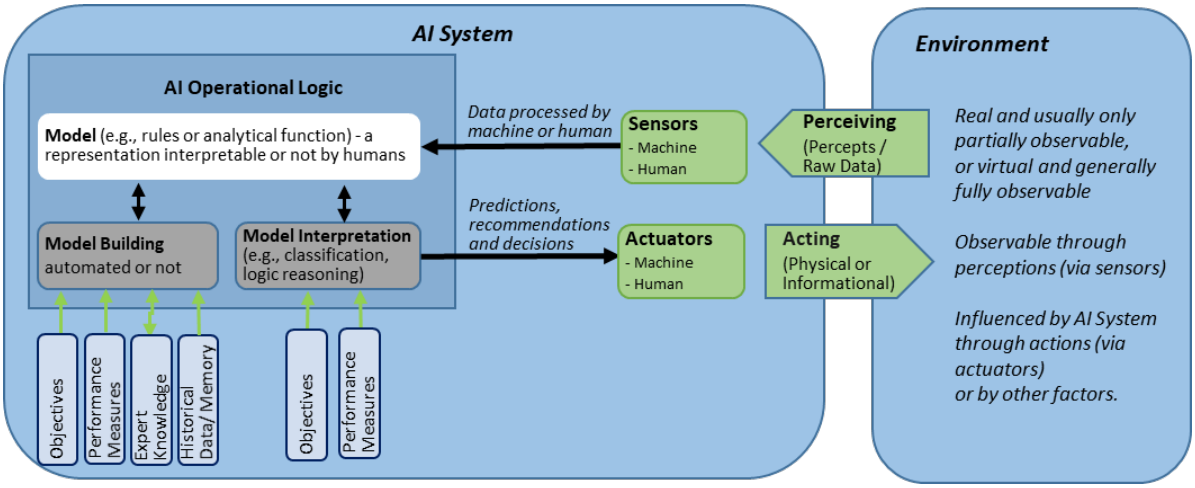
A conceptual view of AI is first presented as the high-level structure of a generic AI system (also referred to as 'Intelligent agent') (Figure 1). An AI system consists of three main elements: Sensors, Operational Logic and Actuators. Sensors collect raw data from the Environment, while Actuators take actions to change the state of the Environment. The key power of an AI system resides in its Operational Logic, which, for a given set of objectives and based on input data from Sensors, provides output for the Actuators – as recommendations, predictions or decisions – that are capable of influencing the state of the Environment.

Figure 1. A high-level conceptual view of an AI system



A more detailed structure captures the main elements that are relevant to the policy dimensions of AI systems (Figure 2). To cover different types of AI systems and different scenarios, the diagram separates the Model Building process (such as machine learning), from the Model (a data object constructed by the Model Building process), and the Model Interpretation process, which uses the Model to make predictions, recommendations and decisions, for the Actuators to influence the Environment.

Figure 2. Detailed conceptual view of an AI System



Environment

An environment in relation to an AI system is a space observable through perceptions (via Sensors) and influenced through actions (via Actuators). Sensors and Actuators are either machines or humans. Environments are either real (*e.g.* physical, social, mental) and usually only partially observable, or virtual (*e.g.* board games) and generally fully observable.

AI system

An AI system is a machine-based system that is capable of influencing the Environment by making recommendations, predictions or decisions for a given set of Objectives. It does so by utilising machine and/or human-based inputs/data to: *i)* perceive real and/or virtual environments; *ii)* abstract such perceptions into models manually or automatically; and *iii)* use Model Interpretations to formulate options for outcomes.

Credit scoring as an illustration of an AI system

A credit-scoring system illustrates a machine-based system that influences its environment (whether people are granted a loan), by making recommendations (a credit score) for a given set of objectives (credit-worthiness). It does so by utilising both machine-based inputs (historical data on people's profiles and on whether they repaid loans) and human-based inputs (a set of rules) to: *i)* perceive real environments (whether people are repaying loans on an ongoing basis); *ii)* abstract such perceptions into models automatically (a credit-scoring algorithm could for example use a statistical model) and *iii)* use model interpretations (the credit-scoring algorithm) to formulate a recommendation (a credit score) of options for outcomes (providing or denying a loan).

"Visually impaired assistant" as an illustration of an AI system

An assistant for visually impaired people illustrates a machine-based system influences its environment by making recommendations (causing a visually impaired person to avoid an obstacle or cross the street) for a given set of objectives (travel from one place to another). It does so utilising machine and/or human-based inputs (large tagged image databases of objects, written words, and even human faces) to: *i)* perceive images of the environment (a camera captures an image of what is in front of a person and sends it to an application), *ii)* abstract such perceptions into models automatically (object recognition algorithms that can recognise a traffic light, a car or an obstacle on the sidewalk) and *iii)* use model interpretation to formulate a recommendation of options for outcomes (providing an audio description of the objects detected in the environment) so the person can decide how to act and thereby influence the environment.

Model

A Model is an actionable representation of all or part of the external environment of an AI system that describes the environment's structure and/or dynamics. The model represents the core of an AI system. A model can be based on data and/or expert knowledge, by humans and/or by automated tools like machine learning algorithms. Model Interpretation is the process of deriving an outcome from a model.

Model Building

A model can be built or adjusted based on data processed either manually by humans or using automated tools like machine learning algorithms, or both. Model Building often uses Historical Data/Memory to aggregate data automatically into the Model, but can also use Expert Knowledge. Objectives (*e.g.* the output variables) and Performance Measures (*e.g.* accuracy, resources for training, representativeness of the dataset) guide the building process.

Model Interpretation

Model Interpretation is the process by which humans and/or automated tools derive an outcome from the model, in the form of recommendations, predictions or decisions. Objectives and Performance Measures guide the execution. In some cases (*e.g.*, deterministic rules), a model can offers a single recommendation, while in other cases (*e.g.*, probabilistic models), a model can offer a variety of recommendations associated with different levels of, for instance, performance measures like level of confidence, robustness or risk. In some cases, during the interpretation process, it is possible to explain why specific recommendations are made, while in other cases, explanation is almost impossible.

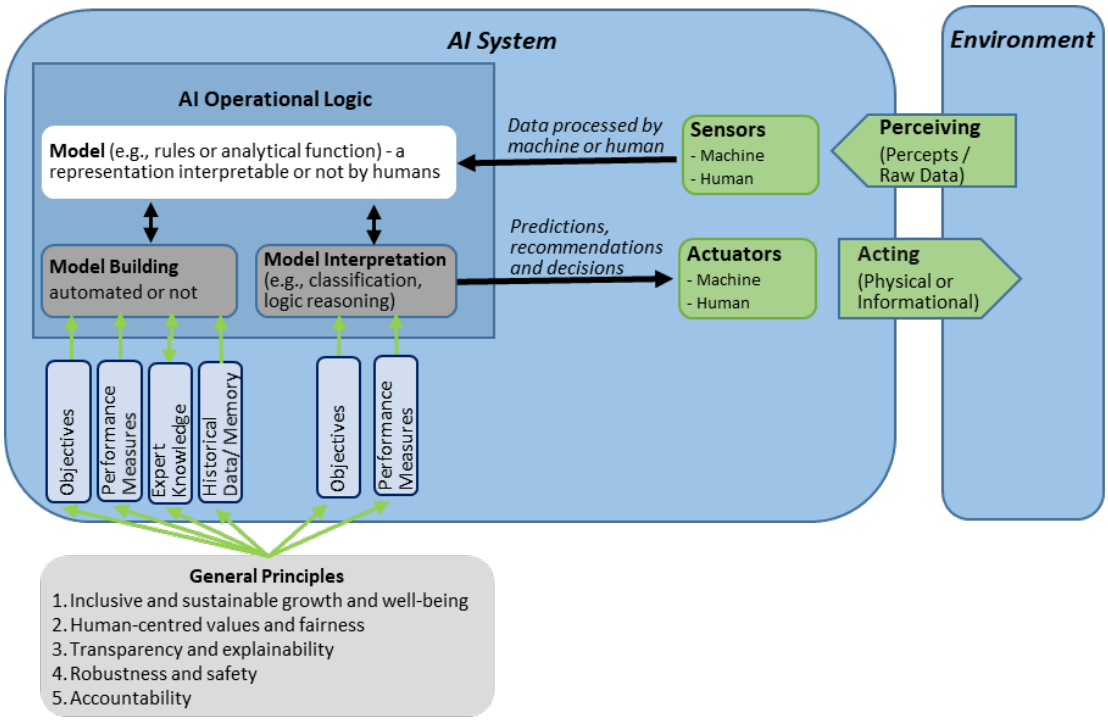
Linking AI Systems to the Principles

The above detailed AI System schema can be linked to the Principles (Figure 3).

Inclusive and sustainable growth and well-being

AI systems can detect patterns in large volumes of data from sensors and can model complex and inter-dependent environments. In turn, AI systems can positively influence the Environment by providing much more accurate and less expensive predictions, recommendations or decisions that generate productivity gains and can help address complex challenges in areas such as science, health and security.

Figure 3. Linking the AI System to the General Principles



Human values and fairness

A model is typically built to achieve specific objectives that may or may not reflect human values, from cancer detection to autonomous weapons. In addition, specific AI systems can be built to achieve a specific set of objectives but later on interpreted with different objectives, as in the case of transfer learning for example.

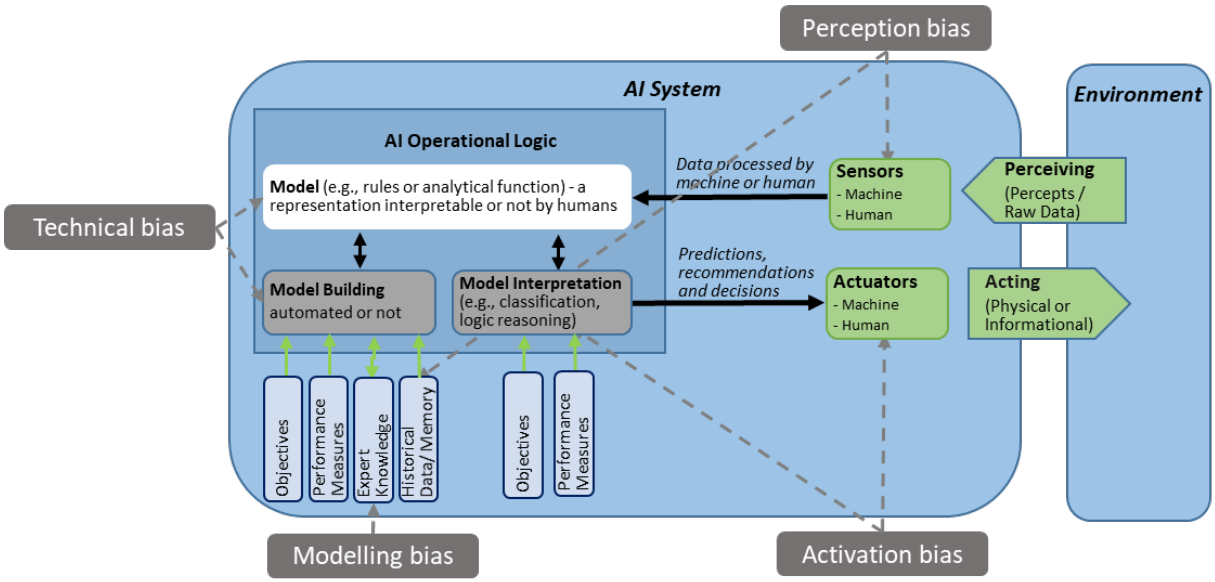
Figure 4 illustrates some of the areas of the AI System in which different types of biases – in particular, perception bias, technical bias, modelling bias and activation bias – are most pronounced. Bias can occur in each of the three main elements of the AI system:

- *Sensors,* notably via Perception bias, whereby the data collected over-represents (or under-represents) one population. Perception bias makes the AI system operate better (or worse) for that population at the expense of others.
- *Operational logic,* notably via Technical bias that arises from constraints or considerations within the technology itself, whether they are known or not. This can include the tools and algorithms an AI system uses. For example, a selected algorithm may work better or worse with a different sets of variables/features. If used in an AI system with different variables or features, its accuracy will be lower, which may introduce bias that is very hard to detect. An accident in 2016 involving a Tesla Model S and a tractor trailer provides an example of technical bias, where the Autopilot's computer vision-based vehicle detection system did not notice the white side of the tractor trailer against a brightly lit sky and did not brake.
- *Expert Knowledge,* notably via Modelling bias, whereby a human manually designing a model (or part of a model) does not take into account some aspects of the environment in building the model, consciously or unconsciously. For example, in an AI system devoted to judiciary decision management, a model can estimate the probability that a person reoffends in future. If the model implemented by a human expert does not take into account

the person's age or gender, for instance because the expert only worked with male or young offenders in the past, the model will include this modelling bias.

- *Actuators*, notably via Activation bias, which relates to how the outputs of the AI system are used in the Environment. For example, actuators such as bots generating twitter posts or news articles can have embedded bias related to the narratives generated by templates.

Figure 4. Areas of the AI system in which biases can appear



Transparency

A Model itself can be interpretable by people (for example in the case of a decision tree) or non-interpretable by people (for example, in the case of deep learning, often referred to as a "black box"). The Model Interpretation process can similarly be more or less understandable. In some cases, during the interpretation process, it is possible to explain why specific recommendations are made, while in other cases (often known as "black box models"), explanation is almost impossible and other types of accountability and transparency measures are called for.

Transparency of an AI system typically focuses on allowing people to understand how an AI system is developed, trained, and deployed; which variables are used, and which variables impact a specific prediction, recommendation or decision.

Robustness and safety

The robustness and safety of AI Systems hinges on Performance Measures that assess how well a system performs compared to specific indicators, for example indicators of accuracy, efficiency, fairness and safety. Performance Measures provide guarantees regarding how a model is built and how it is interpreted. Safety of AI Systems also pertains to Actuators, where most risks of physical and virtual harm reside.

Accountability

Accountability focuses on allocating responsibility to the appropriate organisations or individuals. The accountability of AI systems also relates largely to Performance Measures, which must respect the state of the art.

A Practical Reference Framework for the AI System Lifecycle

In November 2018, the AI Group of experts at the OECD (AIGO) established a subgroup to complement the Principles by detailing the AI system lifecycle. This chapter develops a practical reference framework in which to contextualise and consider ways to implement the Principles in the AI systems lifecycle. After providing an overview of the main phases of the AI system lifecycle, the AI lifecycle actors and the broader set of "stakeholders" affected by AI systems, this annex provides a framework for understanding the risk management approach to AI systems encouraged in the Principles.

Nineteen AI experts participated in the work of the subgroup, which was moderated by Jim Kurose from the U.S. NSF and Nozha Boujemaa from INRIA, and met regularly from mid-December 2018 to mid-February 2019.

The AI System Lifecycle

An AI system incorporates many of the phases involved in traditional software development lifecycles and system development lifecycles more generally but contains specific features.

The AI system lifecycle typically involves the following four phases: i) 'design, data and models'; which is a context-dependent sequence encompassing planning and design, data collection and processing, as well as model building; ii) 'verification and validation'; iii) 'deployment'; and iv) 'operation and monitoring' (Figure 5).

These phases can be described as follows:

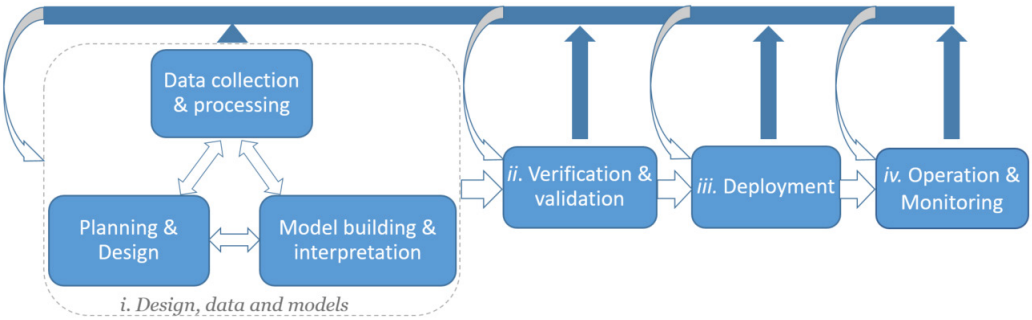
i. **Design, data and modelling** includes several activities, whose order may vary for different AI systems:
   - **Planning and design** of the AI system involves articulating the system's concept and objectives, underlying assumptions, context and requirements, and potentially building a prototype.
   - **Data collection and processing** includes gathering and cleaning data, performing checks for completeness and quality, and documenting the characteristics of the dataset. Dataset characteristics include information on how a dataset was created, its composition, its intended uses, and how it was maintained over time.
   - **Model building and interpretation** involves the creation or selection models/algorithms, their calibration and/or training and interpretation.

ii. **Verification and validation** involves executing and tuning models, with tests to assess performance across various dimensions and considerations.

iii. **Deployment** into live production involves piloting, checking compatibility with legacy systems, ensuring regulatory compliance, managing organisational change, and evaluating user experience.

iv. **Operation and monitoring** of an AI system involves operating the AI system and continuously assessing its recommendations and impacts (both intended and unintended) in light of objectives and ethical considerations. In this phase, problems are identified and adjustments made by reverting to other phases or, if necessary, deciding to retire an AI system from production.

Figure 5. AI system lifecycle



A feature that distinguishes the lifecycle of many AI systems from that of more general system development is the centrality of data and of models that rely on data for their training and evaluation. A characteristic of some AI systems based on machine learning is the capacity to iterate and evolve over time.

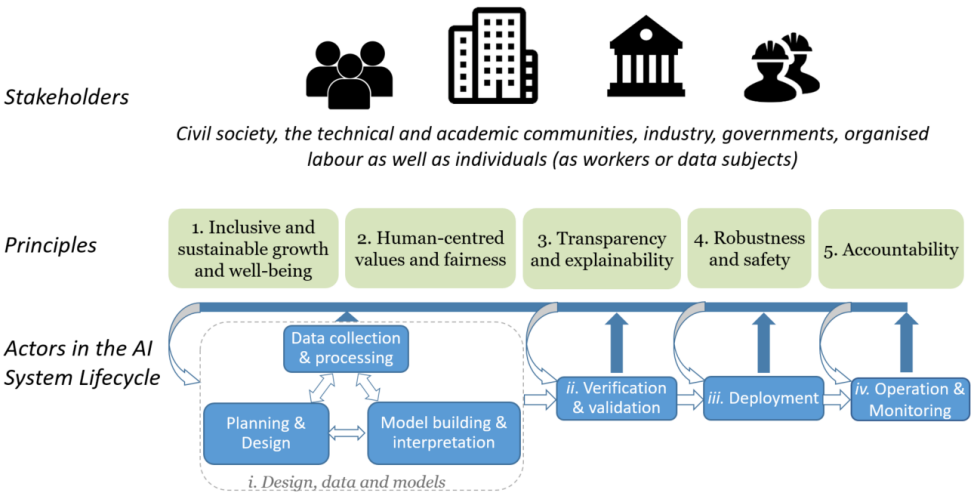
Stakeholders, AI actors and risk management for AI systems

Stakeholders

Stakeholders encompass all public and private sector organisations and individuals involved in, or affected by, AI systems, directly or indirectly. They include, inter alia, civil society, the technical and academic communities, industry, governments, labour representatives and trade unions as well as individuals as workers or data subjects. AI actors are a subset of stakeholders.

Different stakeholders will naturally view each AI principle through a different lens, with different considerations, priorities and questions (Figure 6). These questions and considerations may also differ depending on the phase of the AI system lifecycle.

Figure 6. Stakeholders view of AI principles, in the framework of the AI lifecycle



AI actors

AI actors are those who play an active role in the AI system lifecycle. Public or private sector organisations or individuals that acquire AI systems to deploy or operate them are also considered to be AI actors. AI actors include, inter alia, technology developers, systems integrators, and service and data providers.

The expertise needed at each lifecycle phase varies and may include, inter alia, data science, domain knowledge, modelling, data and model engineering, and governance oversight.

i. **Design, data and modelling:**
   - **Planning and design:** currently involves expertise such as data scientists, domain experts, and governance experts.
   - **Data collection and processing**: currently involves expertise such as data scientists, domain experts, data engineers, data providers.
   - **Model building and interpretation:** currently involves expertise such as modellers, model engineers, data scientists, domain experts.

ii. **Verification and validation:** currently involves expertise such as data scientists, data/model/systems engineers, governance experts.

iii. **Deployment:** currently involves expertise such as system integrators, developers, systems/software engineers and testers.

iv. **Operation and monitoring:** currently involves expertise such as governance experts, domain experts, and systems/software engineers.

A risk management approach for AI systems

Organisations use risk management to identify, assess, prioritise and treat potential risks that can adversely affect the behaviour of systems. Such an approach can also be used to identify risks for different stakeholders and determine how to address these risks throughout the AI system lifecycle.

AI actors implement a risk management approach in conjunction with the AI system lifecycle, both assessing and mitigating risks of the AI system as a whole as well as in each lifecycle phase. As shown in Figure 7, risk management consists of the following steps, whose relevance varies depending on the phase of the AI system lifecycle:

a) *Objectives:* define objectives, functions or properties of the AI system, in context. These functions and properties may change depending on the phase of the AI lifecycle.

b) *Stakeholders and actors:* identify stakeholders and actors involved, i.e., those directly or indirectly affected by the system's functions or properties in each lifecycle phase.

c) *Risk assessment:* assess the potential effects, both benefits and risks, for stakeholders and actors. These will vary, depending on the stakeholders and actors affected, as well as the phase in the AI system lifecycle. In all cases, potential risks to the Principles can be considered.

d) *Risk mitigation:* identify risk mitigation strategies that are appropriate to, and commensurate with, the risk. These should consider factors such as the organisation's goals and objectives, the stakeholders and actors involved, the likelihood of risks materialising and potential benefits.

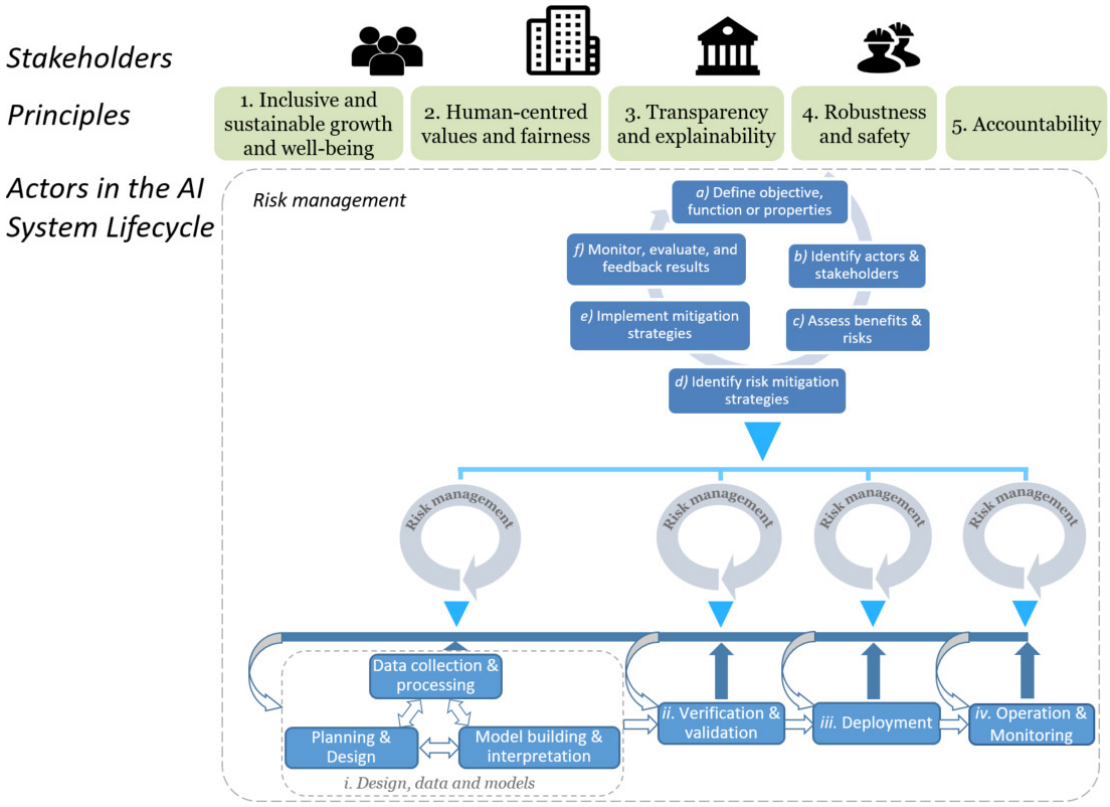
e) *Implementation:* implement risk mitigation strategies.

f) *Monitoring, evaluation and feedback:* monitor, evaluate and feedback results of the implementation.

The use of such an AI risk management system and the documentation of the decisions made at each lifecycle phase can help improve an AI system's transparency and an organisations' accountability for the system.

Figure 7. AI risk-based management approach

To be implemented at each phase of the AI system lifecycle



Illustrating the value of the AI system lifecycle practical reference framework

Fairness considerations provide an example of the value of using such a practical reference framework throughout the AI system lifecycle, to allow stakeholders to engage in more specific discussions and actions in relation to this principle. Different types of biases and other factors that affect fairness may appear in different phases of the AI system lifecycle (Figure 8), including:

i. *Design, data and modelling* phase*:*
   - *Planning and design: non-inclusive design,* whereby an AI system cannot be equally accessed and used by as many people as possible, regardless of age, gender and disability.
   - *Data collection and processing:* data can inaccurately represent the real world or reflect socially-derived artefacts that disadvantage particular groups.
     - *Reporting bias,* whereby people tend to under-report all the information available.
     - *Selection Bias,* whereby the data selected over-represents (or under-represents) one population, making the AI system operate better (or worse) for that population at the expense of others. This can be due, for example, to issues of under-coverage or non-response of some population members or due to the way that sampling is conducted.
     - *Out-group homogeneity bias,* whereby people tend to see those outside their own group as more similar to one another than those in their own groups (e.g. similar attitudes, values, personality traits, and other characteristics).

- *Model engineering, calibration and interpretation:* can also involve biases, notably
  - *Experimenter bias,* whereby the model is subconsciously influenced by the modeller's predisposed notions or beliefs.

ii. *Verification and validation* phase*:*
- *Confirmation bias:* the tendency to search for, interpret, favour, and recall information in a way that confirms one's pre-existing beliefs or hypotheses.

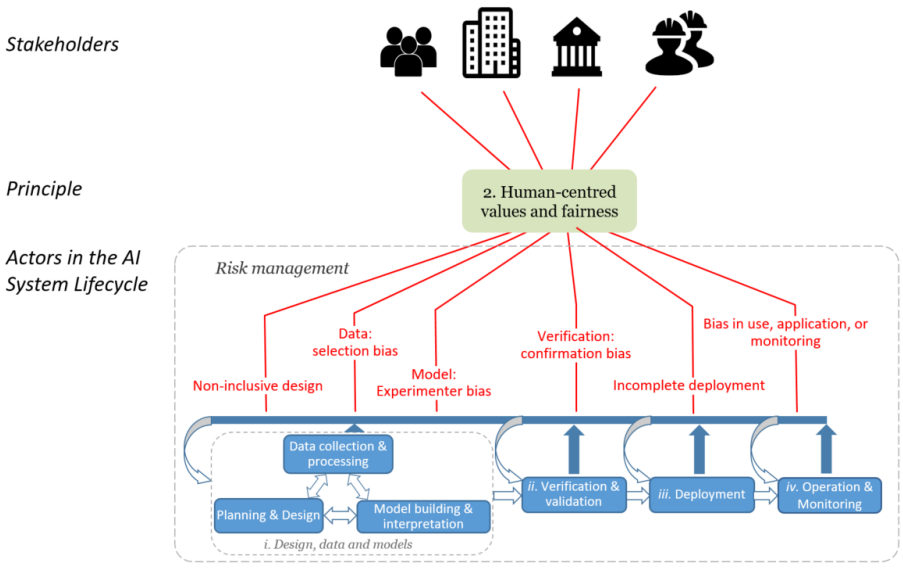
iii. *Deployment* phase*:*
- *Incomplete deployment,* whereby groups of stakeholders may be excluded from using or realising the benefits of a deployed AI system.

iv. *Operation and monitoring* phase*:*
- *Inadequate monitoring,* whereby data may not reflect the breadth of users or uses of a deployed system.

Relevant AI actors can implement risk management strategies to avoid or mitigate these and other biases throughout the AI lifecycle. For example, asking the following questions could help manage and mitigate the risk of *selection bias*:

a) *Defining the objective*: in view of the AI system's objectives, how would selection bias in the data affect its functioning and the Principles?
b) *Stakeholders and actors*: which stakeholders would be affected by selection bias and which AI actors could mitigate this risk?
c) *Risk assessment*: what risks would selection bias create and how likely are they to materialise? What would be the consequences? What level of selection bias would be acceptable in view of potential benefits of the AI system?
d) *Risk mitigation*: what risk controls could be set up during the development phase to prevent selection bias? How can AI actors ensure that this risk remains at an acceptable level?
e) *Implementation*: who should implement the selected risk controls to prevent or mitigate selection bias, when and how?
f) *Monitoring, evaluation and feedback*: how is performance measured, monitored and reviewed, and by who? Who documents and shares information on the risk management of selection bias? With whom?

Figure 8. A view of fairness considerations by AI actors within the AI system lifecycle



Annex A. Scoping principles to foster trust in and adoption of AI

This Annex presents the scoping principles to foster trust in and adoption of artificial intelligence (AI), developed over four meetings by the Expert Group on Artificial Intelligence at the OECD (AIGO). The group concluded its discussion and agreed on this draft at its fourth and last meeting in Dubai, United Arab Emirates, on 8-9 February. This proposal informed the Committee's discussion of a draft Recommendation of the Council on Artificial Intelligence on 14-15 March 2019 that was subsequently adopted by the OECD Council at Ministerial level on 22 May 2019. The present document was declassified by the Committee on 1 July 2019.

Introduction

(References)

Reference to existing OECD instruments (e.g. CDEP + CCP, including privacy and security; MNE Guidelines) and the UN SDGs; UDHR;

Reference to existing national legal, regulatory and policy frameworks applicable to AI, including those related to consumer and personal data protection, intellectual property rights and competition, while noting that such frameworks may need to be adapted;

(Transformative effect of recent developments in AI)

Notably due to recent developments, AI has pervasive, far-reaching and global implications that are transforming societies, economic sectors and the world of work, and are likely to increasingly do so in the future;

(Benefits and challenges)

AI has the potential to improve the welfare of people, to contribute to a positive sustainable global economic activity, to increase innovation and productivity, and to help respond to key global challenges, such as climate change, health crises, resource scarcity and discrimination;

At the same time, these transformations may have disparate effects within, and between, societies and economies, notably economic shifts, transitions in the labour market, deepening inequalities, such as gender, income and skills gaps, and detrimental implications on democracy, freedom, fairness, autonomy and individual control, and data privacy and security;

(Need for a global policy framework and practical guidance on AI)

Trust is a key enabler of digital transformation and, while further AI applications and their implications may be hard to foresee, trustworthiness of AI systems is a key factor for diffusion of AI and for capturing the full potential of the technology.

Given the rapid development and implementation of AI, there is a pressing need for a predictable, stable yet adaptive policy environment that promotes a human-centric approach to AI and practical guidance for trustworthy AI, and that applies to all relevant stakeholders according to their responsibility, in a context-sensitive manner.

This policy framework aims to achieve this objective, to empower individuals, public entities, businesses and workers to engage and thereby to create incentives to turn trustworthy AI into a collaborative and competitive parameter in the global marketplace. Striking an appropriate and fair balance between the opportunities offered and the challenges raised by AI applications is essential to steering AI innovation toward inclusive and sustainable growth and well-being, reduction of inequalities between countries and people, and respect of human rights and democratic values.

[**Recognition** that such policy framework should be developed, implemented, monitored and reviewed through continuous international co-operation and multi-stakeholder and interdisciplinary dialogue, that would also guarantee diversity of thought and consideration of national and regional frameworks.]

[**Indication** that:

- the framework below should be regarded as a baseline which can be supplemented by further work from all stakeholders at the OECD and in other fora.

- all principles are inter-related and should be considered as a whole.]

Common understanding of technical terms for the purposes of these principles

AI system

An AI system is a machine-based system that is capable of influencing the environment by making recommendations, predictions or decisions for a given set of objectives.

It does so by utilising machine and/or human-based inputs to: *i)* perceive real and/or virtual environments; *ii)* abstract such perceptions into models manually or automatically; and *iii)* use model interpretations to formulate options for outcomes.

Model

A model is an actionable representation of all or part of the external environment of an AI system that describes the environment's structure and/or dynamics. The model represents the core of an AI system. A model can be based on data and/or expert knowledge, by humans and/or by automated tools like machine learning algorithms. Model interpretation is the process by which humans and/or automated tools derive an outcome from the model, in the form of recommendations, predictions or decisions.

AI lifecycle

AI system lifecycle phases involve: i) 'design, data and models'; which is a context-dependent sequence encompassing planning and design, data collection and processing, as well as model building; ii) 'verification and validation'; iii) 'deployment'; and iv) 'operation and monitoring'.

AI knowledge

AI knowledge refers to the resources and skills, such as data, algorithms, models, research, know-how, training programmes, governance, processes and best practices, required to understand and participate in the AI system lifecycle.

AI actors

AI actors are those who play an active role in the AI system lifecycle. Public or private sector organisations or individuals that acquire AI systems to deploy, operate and/or use them are also considered to be AI actors.

Stakeholders

Stakeholders encompass all public and private sector organisations and individuals involved in, or affected by, AI systems, directly or indirectly. They include, inter alia, civil society, the technical and academic communities, industry, governments, labour representatives and trade unions as well as individuals as workers or data subjects. AI actors are a subset of stakeholders.

Principles for responsible stewardship of trustworthy AI

1.1. Inclusive and sustainable growth and well-being

All stakeholders should engage in responsible stewardship of trustworthy AI to achieve fair and beneficial outcomes for all people and the planet, such as empowering people and enhancing their capabilities and creativity, advancing inclusion of underrepresented populations and reducing economic and social inequalities, within and across countries, and overall invigorating sustainable economic growth and well-being.

Governments should in particular consider:

- Initiating a meaningful and iterative dialogue inclusive of all stakeholders to enhance understanding of AI, to debate AI-related opportunities and challenges for the economy, the society and the world of work, and to inform policy makers.

- Encouraging AI actors to ensure multidisciplinary collaboration and diversity of views throughout the AI lifecycle to maximise benefits and minimise the potential for harm.

- Supporting AI actors in the implementation of this principle, including through promotion of responsible AI in education and research, exchange of knowledge and best practices, guidance for responsible business conduct and incentives to turn responsible AI into a competitive advantage.

1.2. Human-centred values and fairness

AI actors should set up effective mechanisms to demonstrate respect of human rights and democratic values, including freedom, dignity, autonomy, privacy, non-discrimination, fairness and social justice, and diversity as well as core labour rights, throughout the AI lifecycle.

Governments should in particular consider:

- Encouraging AI actors to assess that AI systems respect human-centred values and fairness on an ongoing basis, and to implement safeguards by design and other measures and processes, including capacity for human final determination, that are appropriate to the context and benefit from multidisciplinary and multi-stakeholder collaboration.

- Promoting codes of ethical conduct, quality standards and quality labels, that help align AI systems with human-centred values and fairness throughout their lifecycle, and help assess AI systems' levels of compliance with these values.

- Ensuring that AI systems allow for individuals' determination over their digital identity and personal data.

1.3. Transparency and explainability

All stakeholders should promote a culture of transparency and responsible disclosure regarding AI systems. In this regard, AI actors should provide, appropriate to the context and state of art, meaningful information to all stakeholders in order to foster understanding of AI systems, to raise their awareness of their interactions with AI systems, including in the workplace, and to enable those adversely affected by an AI system to challenge its recommendations.

Governments should in particular consider:

- Promoting initiatives from AI actors to help make AI systems understandable, including through AI systems that can communicate meaningful information appropriate to the context during their operation to foster understanding of their recommendations.

- Ensuring meaningful disclosure of when and for which purpose stakeholders are interacting with an AI system and who operates it, especially when the system is unbeknownst to the stakeholders.

- Ensuring that natural and legal persons adversely affected by an AI system can obtain, appropriate to the context and state of art, information on the factors and the logic that serve as the basis for its recommendations, without having to comprehend the technology.

1.4. Robustness and safety

AI systems should be robust, in the sense that they should be able to withstand or overcome adverse conditions, and safe, in the sense that they should not pose unreasonable safety risk in normal or foreseeable use or misuse throughout their entire lifecycle.

To this end, AI actors should ensure traceability of the datasets, processes and decisions made during the lifecycle of AI systems to enable understanding of their outcomes and inquiry, where appropriate.

AI actors should also implement or reinforce their risk management approach, on a continuous basis throughout the AI lifecycle, to mitigate risks, including to digital security, as appropriate to the context.

Governments should in particular consider:

- Encouraging AI actors to assess the implications of their contribution to an AI system's lifecycle, in a manner proportionate to their role.

- Calling on AI actors to document the process and decisions made during the AI system's lifecycle, especially for systems with potentially significant consequences on people's lives, to support understanding of AI systems' outcomes and enable accountability.

- Encouraging AI actors to consult stakeholders during the AI lifecycle, including in relation to risk management processes, thus promoting stakeholder participation in all stages of AI systems' lifecycle.

1.5. Accountability

AI actors should be accountable for the proper functioning of AI systems and for the respect of the above principles, based on their individual role, the context, and state of art.

National policies for trustworthy AI

Governments should develop policies, in co-operation with all stakeholders, to promote trustworthy AI systems and achieve fair and beneficial outcomes for people and the planet, consistent with the principles above.

2.1. Investing in responsible AI research and development

Governments should consider and encourage long-term investments in inter-disciplinary basic research and development to spur innovation in trustworthy AI that would focus on challenging technical issues as well as on AI-related social implications and policy issues.

Governments should in particular consider:

- Developing high-level frameworks to coordinate whole-of-government investments, especially in promising areas underserved by market-driven investments.

- Prioritising inter-disciplinary research and development to address the ethical, legal, and social implications of AI, crosscutting issues such as bias, privacy, transparency, accountability and the safety of AI, and difficult technical challenges such as explainability.

- Building open data sets that are representative and preserve privacy in order to provide an un-biased environment for research and development and to encourage innovation and competition, and opening up existing ones, accordingly.

- Using public procurement, promoting joint public and private procurement, and establishing flexible joint venture funding systems to spur market investment in responsible research and development, to encourage broad-based evolution of the market for AI-based solutions, and to foster diffusion of AI systems that benefit society across regions, firms and demographic groups.

2.2. Fostering an enabling digital ecosystem for AI

Governments should foster an enabling ecosystem, including digital technologies and infrastructure, competitive markets as well as mechanisms for sharing AI knowledge to support the development of trustworthy AI systems.

Governments should in particular consider:

- Investing in, and providing incentives to the private sector to invest in, AI enabling infrastructure and technologies such as high-speed broadband, computing power and data storage, as well as fostering entrepreneurship for trustworthy AI systems.

- Encouraging the sharing of AI knowledge through mechanisms such as open AI platforms and data sharing frameworks while respecting privacy, intellectual property and other rights.

2.3 Providing an agile [and controlled] policy environment for AI

Governments should provide an enabling policy environment to support the agile, safe and transparent transition from research and development to deployment and operation of trustworthy AI systems. To this effect, governments should review existing laws, regulations, policy frameworks and assessment mechanisms as they apply to AI and adapt them, or develop new ones as appropriate.

Governments should further encourage that AI actors comply with the applicable national frameworks and global standards.

Governments should in particular consider:

- Using experimentation, including regulatory sandboxes, innovation centres and policy labs, to provide a controlled environment in which AI systems can be tested.

- Encouraging stakeholders to develop or adapt, through an open and transparent process, codes of conduct, voluntary standards and best practices to guide AI actors throughout the AI lifecycle, including for monitoring, reporting, assessing and addressing harmful effects or misuse of AI systems.

- Establishing and encouraging public and private sector oversight mechanisms of AI systems, as appropriate, such as compliance reviews, audits, conformity assessments and certification schemes, while considering the specific needs of and constraints faced by SMEs.

- Establishing mechanisms for continuous monitoring, reporting, assessing and addressing the implications of AI systems that may pose significant risks or target vulnerable groups.

2.4. Building human capacity and preparing for job transformation

Governments should work closely with social partners, industry, academia, and civil society to prepare for the transition in the world of work and empower people with the competences and skills necessary to use, interact and work with AI.

They should ensure that AI deployment in society goes hand in hand with equipping workers fully for a fair transition and new opportunities in the labour markets. They should do so with a view to fostering

entrepreneurship, creating quality jobs, making human work safer, more productive and more rewarding, and ensuring that no one is left behind.

Governments should in particular consider:

- Developing a policy framework conducive to the creation of new employment opportunities.

- Encouraging research on occupational and organisational changes to anticipate future skills needs and improve safety.

- Promoting a broad, flexible and equal opportunity range of life-long education, technological literacy, skills and capacity-building measures to allow people and workers to successfully engage with AI systems across the breadth of applications.

- Developing schemes, including through social dialogue, for fair transition to support people whose current jobs may be significantly transformed by AI, with a focus on training, career guidance and social safeguard systems.

- Encouraging education institutions and employers to provide interdisciplinary education and training needed for trustworthy AI, from STEM to ethics, including through apprenticeships and reskilling programmes to train AI specialists, researchers, innovators, operators and workers.

2.5 International cooperation for trustworthy AI

Governments should actively cooperate at international level, among themselves and with stakeholders in all countries, to invigorate inclusive and sustainable economic growth and well-being through trustworthy AI in all world regions, and to address global challenges.

They should work together transparently in all relevant global and regional fora to advance the adoption and implementation of these principles and progress on trustworthy AI.

Governments should in particular consider:

- Supporting international and cross-sectoral collaboration concerning these principles, including through open, global multi-stakeholder dialogues that can enable long-term expertise for trustworthy AI.

- Promoting cross-border collaboration for responsible AI innovation through sharing of AI knowledge, and maintaining [free] [transborder] flows of data with trust that safeguard security, privacy, human rights and democratic values.

- Encouraging the development of globally accepted practical technical standards, terminology, taxonomy, and measurement methodologies and indicators to guide international co-operation on trustworthy AI.

- Building AI capacity to bridge digital divides and to share the benefits of trustworthy AI among all countries.

[Provision on measurement to be added: Governments should encourage the development of internationally comparable metrics based on common measurement methodologies, standards and best practices to measure global activity related to AI research, development and deployment, and to gather the necessary evidence base to assess progress in the implementation of these principles.]

Annex B. List of AIGO members

This Annex lists the members and expert contributors to the work of the Expert Group on Artificial Intelligence at the OECD (AIGO).

The following experts contributed to the work of the AIGO as members (Table A B.1). Their contributions are greatly acknowledged.

Table A B.1. AIGO members

Name	Title	Organisation / Country	Group / Delegation
Mr. Wonki Min	[AIGO Chair] Vice-Minister and Chair of the OECD Committee on Digital Economy Policy	Ministry of Science and ICT, Korea	Korea
Mr. Tim Bradley	Minister-Counsellor	Department of Industry, Innovation and Science.	Australia
Mr. Alex Cooke	Counsellor, Department of Industry, Innovation and Science	Australian Embassy to Belgium, Luxembourg and Mission to the European Union and NATO	Australia
Ms. Elissa Strome	Executive Director of the Pan-Canadian AI Strategy	Canadian Institute for Advanced Research (CIFAR)	Canada
Mr. Lars Rugholm Nielsen	Head of Section	Danish Business Authority	Denmark
Mr. Antti Eskola	Commercial Counsellor for Innovation and Enterprise Financing Department	Ministry of Economic Affairs and Employment	Finland
Ms. Christel Fiorina	Head of Audiovisual and Multimedia office	Directorate General for Enterprise, French ministry of economy and finance	France
Mr. Bertrand Pailhes	National Coordinator for the French AI Strategy	State Digital Service, Prime Minister's Service	France
Mr. Michael Schönstein	Head of Strategic Foresight & Analysis	Policy Lab “Digital Work & Society”, Federal Ministry for Labour and Social Affairs	Germany
Mr. Nils Börnsen	Policy adviser responsible for AI policy at BMWI	Federal Ministry for Economic Affairs and Energy	Germany
Mr. András Hlács	Counsellor	Permanent Delegation of Hungary to OECD	Hungary
Mr. Osamu Sudoh	Professor, Graduate School of Interdisciplinary Information Studies	University of Tokyo	Japan
Mr. Susumu Hirano	Dean and Professor	Chuo University Graduate School of Policy Studies	Japan
Mr. Chungwon LEE	Director, Multilateral cooperation division	Ministry of Science and ICT, Korea	Korea
Mr. Seongtak Oh	Executive Director, Department of Bigdata	National Information Society Agency, Korea	Korea
Mr. Javier Juárez Mojica	[Co-moderator, 'What is AI' AIGO subgroup] IFT Commissioner	Federal Telecommunications Institute	Mexico
Mr. Wim Rullens	Senior Policy Coordinator	Ministry of Economic Affairs and Climate	Netherlands
Ms. Olivia Erdelyi	Lecturer	Canterbury University	New Zealand
Mr. Robert Kroplewski	Representative	Minister for Digitalisation of the Information Society in Poland	Poland
Mr. Andrey Ignatyev	Deputy Head of OECD Unit	Ministry of Economic Development	Russian Federation
Mr. Konstantin Vishnevskiy	Head of Department for Digital Economy Studies ISSEK HSE	Institute for Statistical Studies and Economics of Knowledge	Russian Federation
Mr. Yeong Zee Kin	Assistant Chief Executive (Data Innovation and Protection Group)	Infocomm Media Development Authority (IM;DA), Government of Singapore	Singapore
Mr. Michal Cíž	AI Policy Expert, EU Digital Single Market	Deputy Prime Minister’s Office for Investments and Informatization	Slovak Republic
Mr. Marko Grobelnik	[Co-moderator, 'What is AI' AIGO subgroup] Researcher in AI	Jozef Stefan Institute - Artificial Intelligence Lab	Slovenia

Ms. Helena Hånell McKelvey	Head of Section, Division for Digital Development	Ministry of Enterprise and Innovation	Sweden
Ms. Livia Walpen	Advisor, International Relations	Swiss Federal Office of Communications	Switzerland
Ms. Ezgi Bener	Expert on Scientific Programmes	The Scientific and Technological Research Council of Turkey (TUBITAK)	Turkey
Mr. Cyrus Hodes	Advisor to the UAE Minister for AI	UAE Ministry for AI	United Arab Emirates
Mr. Edward Teather	Senior Policy Adviser	Office for Artificial Intelligence	United Kingdom
Mr. Adam Murray	International Affairs Officer, Office of International Communications and Information Policy	U.S. Department of State	United States
Ms. Fiona Alexander	NTIA Associate Administrator	U.S. Department of Commerce	United States
Mr. Jim Kurose	[Co-moderator, 'AI system lifecycle' AIGO subgroup] Assistant Director for Computer and Information Science and Engineering, Assistant Director for AI at the Office of Science and Technology Policy	U.S. National Science Foundation	United States
Mr. Matt Chessen	A/Deputy Science and Technology Adviser to the Secretary of State	U.S. Department of State	United States
Ms. Irina Orssich	Political Analyst	European Commission	European Commission
Mr. Jean-Yves Roger	Policy Officer	European Commission	European Commission
Mr. Barry O'Brien	Government and Regulatory Affairs Executive	IBM (Ireland)	BIAC
Ms. Carolyn Nguyen	Director, Technology Policy Group	Microsoft	BIAC
Mr. Ludovic Peran	Public Policy & Gov't Relations	Google	BIAC
Mr. Noberto Andrade	Privacy and Public Policy Manager	Facebook	BIAC
Mr. Marc Rotenberg	Executive Director	Electronic Privacy Information Center (EPIC)	CSISAC
Mr. Suso Baleato	Secretary	CSISAC	CSISAC
Mr. Konstantinos Karachalios	Managing Director	IEEE	ITAC
Ms. Anna Byhovskaya	Senior Policy Advisor	TUAC - Trade Union Advisory Committee to the OECD	TUAC
Ms. Christina J. Colclough	Director Platform & Agency Workers, Digitalisation and Trade	Uni Global Union (UNI)	TUAC
Mr. Nicolas Miailhe	Co-Founder of AI Initiative	AI Initiative (civil society)	Invited expert
Ms. Verity Harding	Co-Lead	DeepMind Ethics & Society	Invited expert
Mr Jason Stanley	Design Research Practice Lead	ElementAI	Invited expert
Mr. Urs Gasser	Director, Technology Policy Group	Harvard Berkman Klein Center	Invited expert
Mr. Ryan Budish	Senior Researcher	Harvard Berkman Klein Center	Invited expert
Ms. Nozha Boujemaa	[Co-moderator, 'AI system lifecycle' AIGO subgroup] Director of Research	INRIA	Invited expert
Mr. Michel Morvan	President / Executive Chairman	IRT SystemX / Cosmo Tech	Invited expert
Mr. Taylor Reynolds	Director, Technology Policy	MIT	Invited expert
Mr. Danny Weitzner	Principal Research Scientist	MIT	Invited expert
Mr. Jonathan	PhD Candidate	MIT	Invited expert

Frankle			
Mr. Jack Clark	Policy Director	OpenAI	Invited expert
Mr Dudu Mimran	CTO	Telekom Innovation Laboratories Israel	Invited expert
Mr. Moez Chakchouk	Assistant Director-General for Communication and Information	UNESCO	Invited expert
Ms. Pam Dixon	Founder/ executive director	World Privacy Forum	Invited expert

The work of AIGO benefited from the contributions and input of other experts (Table A B.2). We gratefully acknowledge their contributions.

Table A B.2. Other contributors to AIGO

Name	Title	Organisation / Country	Group / Delegation
Ms. Karen McCabe	Senior Director, Technology Policy and International Affairs	IEEE	ITAC
Mr. Kentaro Kotsuki	Director of the Policy Research Department Institute for Information and Communications Policy (IICP)	Ministry of Internal Affairs and Communications	Japan
Mr. Tomáš Jucha	Director of Department of Innovative Technologies and International Cooperation	Deputy Prime Minister's Office for Investments and Informatization of the Slovak Republic	Slovak Republic
Mr. Timotej Šooš	Key Horizontal Projects Coordinator	Ministry of Foreign Affairs of Slovenia	Slovenia
Mr. Daniel Egloff	Professor	University of Lausanne	Switzerland
Mr. Philippe Labouchère	Project Leader for Innovation & Entrepreneurship	Swissnex Boston	Switzerland
Mr. Kelly Ross	Deputy Policy Director	American Federation of Labor and Congress	TUAC
Mr. Doug Franz		IEEE	Invited expert
Ms. Eva Thelisson	Co-Founder & CEO	AI Transparency Institute	Invited expert

In addition, we thank the following experts for their contributions to the work of the AIGO subgroups:

Name	Title	Organisation / Country
Mr. Wael Diab	Chair SC 42 (Artificial Intelligence)	ISO
Mr. James Hodson	Member of the Board of Directors and CEO	AI for Good foundation
Mr. Ali G Hessami	Chair and Tech Editor, IEEE P7000 Tech-Ethics Standard	IEEE
Mr. Abe Hsuan	IT & IP lawyer	
Mr. Grigory Marshalko	Expert of the Technical committee for standardization "Cryptography and security mechanisms", "IT security techniques", and "AI"	ISO
Mr. John Shawe Taylor	Head of Computer Science department at UCL and UNESCO AI Chair	UCL (University College London)
Ms. Ingrid Volkmer	Professor and Head, Media and Communications Program	University of Melbourne
Mr. Michael Witbrock	Head, AI Foundations Lab - Reasoning	IBM Research AI

The support of MIT Internet Policy Research Initiative and of the UAE Ministry for AI, which each hosted an AIGO meeting, is also gratefully acknowledged.

[1] The meeting at MIT in January 2019 was chaired by Ms. Fiona Alexander.