4. CONSIDÉRATIONS DE POLITIQUE PUBLIQUE

Le présent chapitre examine les considérations de politique publique à prendre en compte pour la mise en place de systèmes d'IA dignes de confiance et centrés sur l'humain. Il y est question des enjeux liés à l'éthique et à l'équité ; du respect des valeurs humaines et démocratiques, notamment celui de la vie privée ; et des risques de transposition des biais existant dans le monde analogique vers le monde numérique, à l'instar des préjugés sexistes et racistes. La nécessité de faire évoluer les systèmes d'IA pour les rendre plus fiables, sûrs, sécurisés et transparents et les doter de mécanismes permettant de déterminer clairement les responsabilités quant aux résultats obtenus y est soulignée.

La promotion de systèmes d'IA dignes de confiance passe, notamment, par des politiques ayant pour effet de favoriser l'investissement dans la recherche et le développement responsables ; de créer un écosystème numérique où la protection de la vie privée n'est pas menacée par un élargissement de l'accès aux données ; de permettre aux petites et moyennes entreprises de prospérer ; d'encourager la concurrence sans porter atteinte à la propriété intellectuelle ; et d'aider les travailleurs à passer d'un emploi à l'autre au gré de l'évolution du monde du travail.

Une IA centrée sur l'humain

L'IA joue un rôle de plus en plus prépondérant. Plus cette technologie se diffuse, plus les répercussions que peuvent avoir, sur la vie des individus, les prévisions qu'elle établit, les recommandations qu'elle formule ou les décisions qu'elle prend deviennent importantes. La communauté technique, les entreprises et les responsables politiques cherchent activement la meilleure façon d'obtenir une IA centrée sur l'humain et digne de confiance, qui présente un maximum d'avantages pour un minimum de risques et recueille l'adhésion de la société.

Encadré 4.1. Les systèmes d'IA fonctionnant comme des « boîtes noires » posent de nouveaux défis par rapport aux progrès technologiques précédents

Les réseaux neuronaux sont souvent qualifiés de « boîtes noires ». S'il est effectivement possible d'observer le comportement de ces systèmes, il existe une différence considérable entre les réseaux neuronaux et les technologies précédentes en matière de possibilité d'observation, d'où l'expression « boîtes noires ». Les réseaux neuronaux fonctionnent par itération des données sur lesquelles ils sont entraînés. Ils établissent des corrélations probabilistes complexes à plusieurs variables qui deviennent constitutives du modèle qu'ils construisent. Toutefois, ils n'indiquent pas comment les données peuvent être liées entre elles (Weinberger, 2018[1]). Les données sont bien trop complexes pour être appréhendées par le cerveau humain. Les caractéristiques qui distinguent l'IA des avancées technologiques antérieures et nuisent à la transparence et à la détermination des responsabilités sont entre autres les suivantes :

- **La possibilité d'exploration** : Les algorithmes basés sur des règles peuvent être lus et vérifiés règle par règle, ce qui permet de trouver relativement facilement certains types d'erreurs. En revanche, certaines catégories de systèmes d'apprentissage automatique, notamment les systèmes neuronaux, consistent uniquement en des relations mathématiques abstraites entre différents facteurs. Ces systèmes peuvent s'avérer extrêmement complexes et difficiles à comprendre, mêmes pour ceux qui les programment et les entraînent (OCDE, 2016).
- **Le caractère évolutif** : Certains systèmes d'apprentissage automatique fonctionnent en boucle et évoluent avec le temps, ils peuvent même modifier leur comportement de manière imprévue.
- **Le faible niveau de reproductibilité** : Il est possible que le système d'apprentissage automatique n'établisse une prévision spécifique ou ne prenne une décision particulière que dans certaines conditions et avec certaines données, lesquelles ne sont pas nécessairement reproductibles.
- **Davantage de tensions dans la protection des données personnelles et sensibles** :
  - **Les inférences** : Même en l'absence de données protégées ou sensibles, les systèmes d'IA peuvent être capables de déduire ces données et d'établir des corrélations à partir de variables indirectes qui ne sont ni personnelles ni sensibles, telles qu'un historique d'achats ou des données de localisation (Kosinski, Stillwell et Graepel, 2013[2]).
  - **Les variables indirectes indésirables** : Les stratégies politiques et techniques de protection de la vie privée et de non-discrimination ont tendance à limiter les données collectées, à interdire l'utilisation de certaines données ou à supprimer certaines données pour en empêcher l'utilisation. Or, un système d'IA peut baser une prévision sur des données indirectes ayant un lien étroit

avec des données interdites et non-collectées. Qui plus est, la seule façon de détecter ces données indirectes est de collecter également les données sensibles ou personnelles telles que la race. Si de telles données sont collectées, alors il devient essentiel de garantir qu'elles seront toujours utilisées de façon appropriée.

- **Le paradoxe données-vie privée** : dans le cas de nombreux systèmes d'IA, augmenter la quantité des données d'entraînement peut améliorer la précision de leurs prévisions et contribuer à réduire les risques de biais dus à des échantillons faussés. Cependant, plus le volume de données collectées est élevé, plus la vie privée des individus concernés est menacée.

Certains types d'IA – qui reçoivent souvent l'appellation de « boîtes noires » – posent de nouveaux défis par rapport aux progrès technologiques précédents (Encadré 4.1). Au vu de ces défis, l'OCDE – s'appuyant sur les travaux du Groupe d'experts sur l'intelligence artificielle à l'OCDE (AIGO) – a défini les grandes priorités à suivre pour une IA centrée sur l'humain. Premièrement, celle-ci doit contribuer à une croissance inclusive et durable ainsi qu'au bien-être. Deuxièmement, elle doit respecter les valeurs centrées sur l'humain et l'équité. Troisièmement, son utilisation et le fonctionnement de ses systèmes doivent être transparents. Quatrièmement, les systèmes d'IA doivent être fiables et sûrs. Cinquièmement, il convient de déterminer les responsabilités quant aux résultats des prévisions établies par l'IA et des décisions qui en ont découlé. Ces mesures sont considérées comme essentielles pour les prévisions qui emportent des enjeux élevés. Elles sont aussi importantes pour les recommandations commerciales ou pour les utilisations de l'IA de moindre conséquence.

Croissance inclusive et durable et bien-être

L'IA recèle un formidable potentiel à mettre au service des Objectifs de développement durable

L'IA peut contribuer au bien commun et à la réalisation des Objectifs de développement durable (ODD) des Nations Unies dans des domaines tels que l'éducation, la santé, le transport, l'agriculture et les villes durables, entre autres. De nombreuses organisations publiques et privées, dont la Banque mondiale, plusieurs institutions des Nations Unies et l'OCDE, œuvrent pour que l'IA soit mise au service de la réalisation des ODD.

Développer une IA équitable et ouverte à tous devient une priorité croissante

Développer une IA équitable et ouverte à tous devient une priorité croissante. Cela est d'autant plus vrai dans la mesure où l'on redoute que l'IA aggrave les inégalités ou creuse les écarts existant au sein des pays et entre pays développés et pays en développement. Ces écarts résultent de la concentration des ressources en IA – technologies, compétences, ensembles de données et puissance de calcul – dans quelques entreprises et pays. D'autres inquiétudes concernent le fait que l'IA puisse perpétuer des préjugés (Talbot et al., 2017[3]). Certains craignent qu'elle ait un impact différent sur les populations vulnérables et sous-représentées, à savoir, entre autres, les personnes moins instruites, les personnes peu qualifiées, les femmes et les personnes âgées, en particulier dans les pays à revenu faible ou intermédiaire (Smith et Neupane, 2018[4]). Le Centre canadien de recherches pour le développement international a récemment recommandé la mise en place d'un fonds mondial baptisé « l'IA au service du développement ». Ce fonds permettrait la création», dans les pays à revenu faible ou intermédiaire, de « centres d'excellence en IA chargés d'appuyer l'élaboration et

l'exécution de politiques inclusives fondées sur des données probantes (Smith et Neupane, 2018[4]). L'objectif est de veiller à ce que les retombées de l'IA soient réparties de manière équitable et conduisent à des sociétés plus égalitaires. Les initiatives en faveur d'une IA inclusive visent à garantir un large partage des gains économiques générés par l'IA au sein de la société, pour que personne ne soit laissé de côté.

L'IA inclusive et durable intéresse tout particulièrement des pays comme l'Inde (NITI, 2018[5]) ; des entreprises comme Microsoft[1] ; et des groupes d'universitaires comme le Berkman Klein Center à Harvard. Ainsi, Microsoft a lancé, entre autres projets, l'application mobile « Seeing AI ». Cette application à l'usage des personnes malvoyantes scanne et reconnaît tous les éléments de leur environnement direct, et en fournit une audiodescription. Par ailleurs, Microsoft investit actuellement deux millions de dollars des États-Unis dans des initiatives permettant d'utiliser l'IA pour répondre à des enjeux de durabilité, par exemple en matière de préservation de la biodiversité et de lutte contre le changement climatique (Heiner et Nguyen, 2018[6]).

Valeurs centrées sur l'humain et équité

Droits de l'homme et codes d'éthique

Le droit international des droits de l'homme consacre des normes éthiques

Le droit international des droits de l'homme consacre des normes éthiques. L'IA peut favoriser le respect des droits de l'homme tout comme elle peut créer de nouveaux risques de violation, délibérée ou accidentelle, de ces droits. Avec les structures juridiques et autres structures institutionnelles connexes, le droit relatif aux droits de l'homme peut aussi constituer l'un des outils au service d'une IA centrée sur l'humain (Encadré 4.2).

Encadré 4.2. Les droits de l'homme et l'IA

Le droit international des droits de l'homme internationaux renvoie à un corpus juridique international, incluant la Charte internationale des droits de l'homme[1], ainsi qu'aux systèmes régionaux de protection des droits de l'homme élaborés au cours des 70 dernières années à travers le monde. Les droits de l'homme fournissent une série de normes minimales universelles fondées, entre autres, sur les valeurs de dignité humaine, d'autonomie et d'égalité, dans le cadre de l'État de droit. Ces normes et les mécanismes juridiques qui y sont associés font que les pays sont tenus en droit de respecter et protéger les droits de l'homme et d'en garantir la pleine jouissance. Ils exigent en outre que ceux qui ont été privés de leurs droits ou dont les droits ont été violés disposent d'un recours effectif.

Les droits de l'homme incluent notamment le droit à l'égalité, le droit à la non-discrimination, le droit à la liberté d'association, le droit à la vie privée ainsi que divers droits économiques, sociaux et culturels tels que le droit à l'éducation ou le droit à la santé.

Des instruments intergouvernementaux récents, tels que les *Principes directeurs des Nations Unies relatifs aux entreprises et aux droits de l'homme* (HCDH, 2011[7]) traite aussi du rôle des acteurs privés dans le contexte des droits de l'homme, les investissant d'une « responsabilité » quant au respect de ces droits. En outre, la version mise à jour en 2011 des *Principes directeurs de l'OCDE à l'intention des entreprises multinationales* (OCDE, 2011[8]), recueil de recommandations adressées par les gouvernements aux entreprises, contient un chapitre consacré aux droits de l'homme.

Les droits de l'homme recoupent des préoccupations éthiques plus larges et d'autres domaines de réglementation en rapport avec l'IA, tels que la protection des données personnelles ou la législation relative à la sécurité des produits. Toutefois, ces autres préoccupations et questions ont souvent une portée différente.

1. La Charte internationale des droits de l'homme comprend la Déclaration universelle des droits de l'homme, le Pacte international relatif aux droits civils et politiques, et le Pacte international relatif aux droits économiques, sociaux et culturels.

L'IA promet de faire grandir le respect des droits de l'homme

Compte tenu de l'ampleur potentielle de ses applications et utilisations, l'IA promet de faire progresser la protection et le respect des droits de l'homme. Elle pourrait ainsi servir à analyser les ressorts des pénuries alimentaires pour mieux lutter contre la faim, à améliorer les diagnostics et les traitements médicaux ainsi qu'à accroître la disponibilité et l'accessibilité des services de santé, et à dévoiler en plein jour les discriminations.

L'IA pourrait aussi desservir la cause des droits de l'homme

L'IA peut aussi poser plusieurs problèmes dans le domaine des droits de l'homme, ces problèmes étant souvent mentionnés lors des débats dont elle fait l'objet et, plus généralement, des débats d'éthique. Certains systèmes d'IA, ou l'usage qui est en fait, pourraient constituer une violation, accidentelle ou non, des droits de l'homme. L'aspect accidentel est particulièrement étudié. Les algorithmes d'apprentissage automatique qui prédisent la récidive, par exemple, peuvent présenter un biais non détecté. Néanmoins, il arrive aussi que des technologies d'IA soient associées à des atteintes délibérées aux droits de l'homme. Ainsi en est-il lorsqu'elles servent, par exemple, à traquer des dissidents politiques, à museler la liberté d'expression ou encore à restreindre la participation des individus à la vie politique. Dans ces cas-là, la violation en elle-même ne réside pas tout entière dans l'utilisation de l'IA. Toutefois, elle pourrait être aggravée par la technicité et l'efficacité de celle-ci.

L'utilisation de l'IA peut également poser des problèmes inédits lorsque ses effets sur les droits de l'homme ne sont pas voulus ou sont difficiles à déceler. Cela peut tenir à l'utilisation de données d'entraînement de mauvaise qualité, à la façon dont le système est conçu ou à la complexité des interactions entre le système d'IA et son environnement. L'exacerbation par les algorithmes des discours haineux ou des incitations à la violence sur l'internet en est un exemple. Un autre exemple est l'amplification non intentionnelle des fausses nouvelles, qui peut avoir des répercussions sur le droit à prendre part à la vie politique et aux affaires publiques. L'ampleur et l'incidence probables du préjudice seront fonction de celles que peuvent avoir les décisions prises par un système d'IA donné. Par exemple, une décision prise par un système d'IA de recommandation d'actualités a une incidence potentielle plus faible que la décision d'un algorithme qui prédit le risque de récidive des détenus en liberté conditionnelle.

Les codes d'éthique de l'IA complètent les cadres relatifs aux droits de l'homme

Les codes d'éthique peuvent parer au risque que l'IA puisse ne pas être centrée sur l'humain ou ne pas être en adéquation avec les valeurs humaines. Les entreprises privées comme les gouvernements ont adopté un grand nombre de codes d'éthique en lien avec l'IA.

Par exemple, l'entreprise DeepMind, qui appartient à Google, a créé en octobre 2017 une unité consacrée à l'éthique (DeepMind Ethics & Society)[2]. L'unité a pour but d'aider les

technologues à comprendre les incidences éthiques de leur travail et d'aider la société à décider en quoi l'IA peut lui être profitable. En outre, l'unité financera des recherches externes sur, entre autres, le biais algorithmique, l'avenir du travail ou les armes létales autonomes. L'entreprise Google a elle aussi annoncé la mise en place d'une série de principes éthiques destinés à guider ses recherches, le développement de ses produits et ses décisions commerciales[3]. Elle a publié un livre blanc sur la gouvernance de l'IA, dans lequel elle met en évidence les points à éclaircir avec les gouvernements et les sociétés civiles[4]. La philosophie de Microsoft en matière d'IA est de « développer l'ingéniosité humaine grâce à une technologie intelligente » (Heiner et Nguyen, 2018[9]). L'entreprise a lancé des projets visant à garantir un développement inclusif et durable.

Avec ses mécanismes institutionnels et son architecture globale, le droit relatif aux droits de l'homme fournit l'orientation et l'assise nécessaires pour garantir un développement et une utilisation de l'IA en société qui soient éthiques et centrés sur l'humain.

Le recours aux cadres relatifs aux droits de l'homme dans le contexte de l'IA offre des avantages

Le recours aux cadres relatifs aux droits de l'homme dans le contexte de l'IA offre différents avantages, notamment de par les institutions en place, la jurisprudence, le langage universel et la reconnaissance internationale qui entourent ces cadres :

- **Les institutions en place** : Une vaste infrastructure internationale, régionale et nationale a été mise en place au fil du temps dans le domaine des droits de l'homme. Elle est composée d'organisations intergouvernementales, de tribunaux, d'organisations non gouvernementales, d'universités, ainsi que d'autres institutions et communautés dans le cadre desquelles il est possible d'invoquer les droits de l'homme et d'exercer un recours.
- **La jurisprudence** : En tant que normes juridiques, les valeurs protégées par les droits de l'homme reçoivent leur traduction concrète, et sont rendues juridiquement contraignantes, dans des situations spécifiques grâce à la jurisprudence et au travail d'interprétation des institutions internationales, régionales et nationales.
- **Un langage universel** : Les droits de l'homme fournissent un langage universel pour une problématique internationale. Associé à l'infrastructure relative aux droits de l'homme, ce langage peut aider à autonomiser un plus large éventail de parties prenantes. Celles-ci peuvent ainsi participer au débat sur la place de l'IA dans la société aux côtés d'acteurs intervenant directement dans le cycle de vie de cette technologie.
- **Une reconnaissance internationale** : Les droits de l'homme bénéficient d'une reconnaissance et d'une légitimité importantes au niveau international. Qu'un acteur passe seulement pour les enfreindre et il y aura probablement de lourdes conséquences, puisque sa réputation en sera sans doute passablement écornée.

Une approche de l'IA basée sur les droits de l'homme peut aider à identifier les risques, les priorités, les groupes vulnérables et à proposer des solutions

- **Identification des risques** : Les cadres relatifs aux droits de l'homme peuvent aider à identifier les risques de préjudice. En particulier, ils peuvent servir à mettre en œuvre des études sur la diligence raisonnable en matière de droits de l'homme , par exemple des études d'impact sur les droits de l'homme (Encadré 4.3).

- **Exigences fondamentales** : En tant que normes minimales, les droits de l'homme définissent des exigences fondamentales inviolables. Par exemple, dans le cadre de la réglementation relative à l'expression sur les réseaux sociaux, la jurisprudence en matière de droits de l'homme aide à faire des discours haineux une limite à ne pas franchir.
- **Identification des situations à haut risque** : Les droits de l'homme peuvent s'avérer utiles pour repérer les situations ou activités à haut risque. Dans de tels cas, il convient de redoubler d'attention à moins que l'on estime qu'il n'est pas approprié de recourir à l'IA.
- **Identification des groupes ou des communautés vulnérables** : Les droits de l'homme peuvent aider à identifier les groupes ou communautés vulnérables ou à risque en lien avec l'IA. Certains individus ou communautés peuvent être sous-représentés en raison, par exemple, d'une utilisation limitée des smartphones.
- **Réparation** : En tant que normes juridiques assorties d'obligations, les droits de l'homme peuvent assurer une réparation à ceux à qui l'on a fait du tort. Ces réparations incluent, par exemple, une cessation d'activité, l'élaboration de nouveaux processus ou politiques, des excuses ou une indemnité pécuniaire.

Encadré 4.3. Les études d'impact sur les droits de l'homme

Les études d'impact sur les droits de l'homme peuvent aider à mettre en évidence des risques que les acteurs intervenant au cours du cycle de vie de l'IA n'auraient pas nécessairement prévus sans cela. À cette fin, elles portent davantage sur les effets connexes sur l'homme que sur l'optimisation de la technologie ou de ses produits. Ces études, ou des processus similaires, pourraient garantir le respect des droits de l'homme dès la conception de la technologie et tout au long de son cycle de vie.

Les études d'impact sur les droits de l'homme mesurent un grand nombre des effets que la technologie peut avoir sur les droits de l'homme, et ce dans le cadre d'une démarche de grande ampleur nécessitant beaucoup de ressources. Il est probablement plus simple de partir du système d'IA en question. De cette façon, l'étude ne porte que sur un nombre limité de domaines où l'on a le plus de chances de constater des problèmes en matière de droits. Les organisations du secteur peuvent contribuer à la réalisation études d'impact pour le compte de petites et moyennes entreprises (PME) ou pour celui d'entreprises non technologiques qui utilisent des systèmes d'IA sans forcément maîtriser la technologie. La *Global Network Initiative* est l'une de ces organisations qui œuvrent au respect de la liberté d'expression et à la protection de la vie privée. Elle aide des entreprises à planifier des études sur les droits de l'homme et à les intégrer dans leurs projets de nouveaux produits (https://globalnetworkinitiative.org/).

Les études d'impact sur les droits de l'homme présentent l'inconvénient d'être généralement exécutées entreprise par entreprise, alors même que les systèmes d'IA peuvent impliquer de nombreux acteurs. De ce fait, il peut s'avérer inefficace de n'étudier qu'une seule composante. Microsoft a été la première grande entreprise technologique à mener à bien une étude d'impact de l'IA en 2018.

D'autre part, la mise en œuvre d'une approche de l'IA basée sur les droits de l'homme se heurte à d'importantes difficultés, lesquelles sont liées au fait que les droits de l'homme s'adressent aux États, que leur garantie dépend des pays et territoires, qu'ils sont mieux adaptés pour remédier à des préjudices majeurs causés à un petit groupe d'individus et qu'ils peuvent coûter cher aux entreprises :

- **Les droits de l'homme s'adressent aux États, non aux acteurs privés**, or les acteurs du secteur privé jouent un rôle clé dans la recherche sur l'IA comme dans le développement et le déploiement de systèmes fondés sur cette technologie. Cette difficulté n'est pas propre à l'IA. Plusieurs initiatives intergouvernementales cherchent à combler le fossé entre les secteurs public et privé. Au-delà de ces efforts, il est de plus en plus généralement admis que les entreprises ont tout intérêt à respecter les droits de l'homme[5].
- **La garantie des droits de l'homme dépend des pays et territoires**. En général, la partie requérante doit démontrer qu'elle a qualité pour agir dans un pays ou sur un territoire donné. Cette démarche n'est peut-être pas optimale lorsque sont mis en causes de grandes entreprises multinationales et des systèmes d'IA couvrant de multiples pays et territoires.
- **Les droits de l'homme sont mieux adaptés pour remédier à des préjudices majeurs causés à un petit groupe d'individus**, qu'à des préjudices moins importants subis par un grand nombre d'individus. En outre, les droits de l'homme et leurs structures peuvent sembler opaques aux non-initiés.
- **Dans certains contextes, les droits de l'homme ont la réputation de coûter cher aux entreprises**. En conséquence, les démarches qui mettent en avant l'éthique, la protection des consommateurs ou la conduite responsable des entreprises, ainsi que les arguments économiques en faveur du respect des droits de l'homme, semblent prometteuses.

Certains des défis généraux posés par l'IA, tels que la transparence et l'explicabilité, concernent aussi le respect des droits de l'homme (voir ci-après la section « Transparence et explicabilité »). Sans transparence, il est difficile de repérer les violations des droits de l'homme ou d'étayer une plainte pour violation. Il en va de même pour ce qui est de demander réparation, de déterminer les liens de causalité et d'établir les responsabilités.

La protection des données personnelles

L'IA défie les notions de « données personnelles » et de consentement

L'IA est de plus en plus capable d'établir des liens entre différents ensembles de données et de faire coïncider différents types d'information, ce qui a de graves conséquences. Les données conservées séparément étaient autrefois considérées comme non personnelles (ou, s'étant vu retirer tout élément d'identification, elles avaient été « anonymisées »). Cependant, l'IA est capable de croiser ces données non personnelles avec d'autres données et de les réattribuer ensuite aux individus concernés, ce qui en fait à nouveau des données personnelles (ou « désanonymisées »). Ainsi, la corrélation algorithmique fragilise la distinction entre les données personnelles et les autres données. Les données non personnelles peuvent de plus en plus servir à ré-identifier des individus ou à déduire des informations sensibles les concernant, au-delà de celles qu'ils avaient divulgués de leur plein gré à l'origine (Cellarius, 2017[10]). Par exemple, en 2007, des chercheurs avaient déjà utilisé des données dites anonymes pour associer la liste des films loués sur Netflix avec des avis publiés sur le site IMDB. Ce faisant, ils ont identifié les personnes ayant loué des films et ont eu accès à l'historique complet de leurs locations. L'augmentation du nombre de données collectées et les progrès technologiques vont de plus en plus permettre d'établir ce type de rapprochements. Il devient difficile de déterminer quelles données peuvent être considérées comme non personnelles et le resteront.

Il est toujours plus difficile de distinguer les données sensibles des données non sensibles, comme l'illustre le règlement général de l'Union européenne sur la protection des données (RGPD). Certains algorithmes parviennent à déduire des informations sensibles à partir de données « non sensibles », ainsi ceux qui déterminent l'état émotionnel d'individus à la manière dont ils tapent sur leur clavier (Privacy International et Article 19, 2018[11]). L'utilisation de l'IA pour identifier ou ré-identifier des données initialement non personnelles ou anonymisées représente aussi un problème sur le plan juridique. Les garde-fous en place, tels que la *Recommandation du Conseil de l'OCDE concernant les Lignes directrices régissant la protection de la vie privée et les flux transfrontières de données de caractère personnel* (ci-après les « Lignes directrices relatives à la vie privée »), s'appliquent aux données personnelles (Encadré 4.4). En conséquence, il n'est pas clairement établi si, ou à quel moment, ces cadres incluent dans leur périmètre les données qui, dans certaines circonstances, seraient, ou pourraient être, identifiables (Office of the Victorian Information Commissioner, 2018[12]). Une interprétation extrême pourrait élargir considérablement le champ de la protection de la vie privée, mais rendrait du même coup cette protection difficile à assurer dans les faits.

Encadré 4.4. Les Lignes directrices de l'OCDE relatives à la vie privée

La *Recommandation du Conseil concernant les Lignes directrices régissant la protection de la vie privée et les flux transfrontières de données de caractère personnel* (ci-après dénommée les « Lignes directrices relatives à la vie privée ») a été adoptée en 1980 et actualisée en 2013 (OCDE, 2013[13]). Elle contient des définitions de termes pertinents en ce domaine, et notamment celle des « données de caractère personnel », entendues comme « toute information relative à une personne physique identifiée ou identifiable (personne concernée) ». Elle définit également les principes devant régir le traitement de ces données. Ces principes ont trait à la limitation en matière de collecte (avec, lorsqu'il y a lieu, le consentement des individus comme garantie), à la qualité des données, à la spécification des finalités, à la limitation de l'utilisation, aux garanties de sécurité, à la transparence, à la participation individuelle et à la responsabilité. En outre, la Recommandation dispose que, lors de la mise en œuvre des Lignes directrices relatives à la vie privée, les membres doivent veiller à ce que les personnes concernées ne fassent l'objet d'aucune discrimination déloyale. La mise en œuvre des Lignes directrices relatives à la vie privée devait faire l'objet d'une révision en 2019 pour que soient pris en considération, entre autres, les avancées récentes, notamment celles réalisées dans le domaine de l'IA.

L'IA défie également les principes de protection des données personnelles concernant la limitation en matière de collecte, la limitation de l'utilisation et la spécification des finalités

Pour entraîner et optimiser les systèmes d'IA, les algorithmes d'apprentissage automatique ont besoin d'énormes quantités de données, ce qui incite à en maximiser la collecte plutôt qu'à la freiner. Avec l'utilisation croissante des dispositifs fondés sur l'IA et de l'internet des objets (IdO), cette collecte à la fois est plus abondante, plus fréquente et plus simple. Ces données sont reliées à d'autres données, parfois plus ou moins à l'insu des personnes concernées ou sans leur consentement.

Les tendances identifiées et l'évolution de « l'apprentissage » sont difficiles à anticiper. En conséquence, la collecte et l'utilisation de données peuvent aller au-delà de ce que savait initialement la personne concernée, de ce qui lui avait été communiqué et de ce à quoi elle

avait consenti (Privacy International et Article 19, 2018[11]). Cela est potentiellement incompatible avec les principes de limitation en matière de collecte, de limitation de l'utilisation et de spécification des finalités énoncés dans les Lignes directrices relatives à la vie privée (Cellarius, 2017[10]). Les deux premiers principes reposent en partie sur le consentement de la personne concernée (selon le cas, étant donné qu'il n'est pas possible de recueillir le consentement dans certaines situations). Ce consentement est le point de départ de la collecte de données à caractère personnel ou de leur utilisation à des fins autres que celles initialement indiquées. Les technologies d'IA telles que l'apprentissage profond, qui sont difficiles à comprendre ou à surveiller, sont également difficiles à expliquer aux personnes concernées. Cela constitue un défi pour les entreprises. Elles indiquent qu'il est compliqué de concilier la vitesse à laquelle l'IA accède à des données, les analyse et les utilise, qui augmente de manière exponentielle, avec ces principes de protection des données (OCDE, 2018[14]).

Ces difficultés sont exacerbées par l'association des technologies liées à l'IA avec les progrès de l'IdO, c'est-à-dire la connexion à internet d'un nombre croissant d'appareils et d'objets avec le temps. Le fait que les technologies d'IA et celles de l'IdO soient de plus en plus souvent associées (avec, par exemple, des dispositifs de l'IdO dotés d'IA, ou des algorithmes d'IA utilisés pour analyser les données de l'IdO) entraîne la collecte constante de données toujours plus nombreuses, et notamment de données personnelles. Ces données peuvent être de plus en plus facilement croisées entre elles et analysées. D'une part, les appareils qui recueillent des informations sont toujours plus nombreux (comme les caméras de surveillance ou les véhicules autonomes), d'autre part, la technologie liée à l'IA s'est améliorée (c'est le cas, par exemple, de la reconnaissance faciale). Combinées, ces deux tendances risquent de donner lieu à des résultats plus intrusifs que chaque facteur pris séparément (Office of the Victorian Information Commissioner, 2018[12]).

L'IA peut aussi renforcer la participation et le consentement des individus

L'IA a le potentiel de renforcer les données personnelles. Par exemple, des initiatives visant à élaborer des systèmes d'IA reposant sur les principes de la protection de la vie privée dès la conception et de la protection de la vie privée par défaut sont en cours au sein de plusieurs organismes de normalisation technique. Pour la plupart, ces organismes utilisent et adaptent des lignes directrices relatives à la vie privée, dont celles de l'OCDE. En outre, l'IA est utilisée pour offrir aux individus des services personnalisés adaptés à leurs besoins, basés sur leurs préférences de confidentialité telles qu'acquises au fil du temps (Office of the Victorian Information Commissioner, 2018[12]). Ces services peuvent aider les individus à s'y retrouver parmi les différentes politiques de traitement des données personnelles et à s'assurer que leurs préférences sont prises en considération partout. Dans ce cas, l'IA facilite le consentement éclairé et la participation des individus. À titre d'exemple, une équipe de chercheurs a mis au point Polisis, structure automatisée qui utilise les classificateurs d'un réseau neuronal pour analyser les politiques de confidentialité (Harkous, 2018[15]).

Équité et éthique

Les algorithmes d'apprentissage automatique peuvent refléter les biais implicites de leurs données d'entraînement

À ce jour, les initiatives stratégiques relatives à l'IA donnent une place prépondérante aux questions d'éthique, d'équité et/ou de justice. La propension des algorithmes d'apprentissage automatique à refléter et à reproduire les biais implicites de leurs données d'entraînement, tels que les biais raciaux et les associations stéréotypées, suscitent de nombreuses inquiétudes.

Parce que les artefacts technologiques incarnent souvent des valeurs sociales, les débats sur l'équité doivent établir clairement à quelles sociétés les technologies doivent profiter, qui doit être protégé et grâce à quelles valeurs fondamentales (Flanagan, Howe et Nissenbaum, 2008[16]). Des disciplines telles que la philosophie, le droit et l'économie sont aux prises depuis des décennies avec diverses conceptions de l'équité qui correspondent à autant d'éclairages différents, illustrant toute la diversité des interprétations que l'on peut donner de cette notion et des implications qu'elle peut avoir dans le champ politique.

Les notions philosophiques, juridiques et informatiques de l'équité et d'une IA éthique varient

La philosophie met l'accent sur les concepts de bonne et mauvaise conduites, de bien et de mal, et de morale. Trois grandes théories philosophiques sont dignes d'attention dans le contexte d'une IA éthique (Abrams et al., 2017[17]) :

- **L'approche basée sur les droits fondamentaux de l'homme**, associée à Emmanuel Kant, définit les principes formels de l'éthique, qui sont des droits spécifiques tels que le respect de la vie privée ou la liberté. Elle protège ces principes au moyen de réglementations que les systèmes d'IA doivent respecter.
- **L'approche utilitariste**, suivie par Jeremy Bentham et John Stuart Mill, met l'accent sur les politiques publiques qui maximisent le bien-être humain en se basant sur des analyses de rentabilité économique. S'agissant de l'IA, l'approche utilitariste soulève la question de savoir *qui* doit voir son bien-être maximisé (par exemple, les individus, la famille, la société ou les institutions/gouvernements), la réponse pouvant influer sur la conception des algorithmes.
- **L'approche fondée sur l'éthique de la vertu**, inspirée de la philosophie d'Aristote, est axée sur les valeurs et normes éthiques dont une société a besoin afin d'aider les individus dans leurs efforts quotidiens pour vivre une vie qui vaut la peine d'être vécue. Cette approche soulève la question de savoir quelles sont les valeurs et les normes éthiques qui garantissent une protection.

Dans le droit, les termes « égalité » et « justice » sont souvent utilisés pour désigner les concepts d'équité. Les deux grandes approches juridiques de l'équité sont l'équité individuelle et l'équité de groupe.

- **L'équité individuelle** correspond à la notion d'égalité devant la loi. Elle implique que tous les individus doivent être traités sur un pied d'égalité et ne pas subir de discriminations au regard de leurs spécificités. L'égalité fait partie des droits humains internationaux.
- **L'équité de groupe** privilégie l'équité du résultat. Elle veille à ce que le résultat ne diffère pas de façon systématique pour les personnes qui, sur la base d'une caractéristique protégée (telles que la race ou le genre), appartiennent à des groupes différents. L'équité de groupe soutient que les différences et les contextes historiques peuvent conduire des groupes distincts à réagir diversement face à une situation donnée. L'approche de l'équité de groupe diffère considérablement selon les pays. Certains utilisent, par exemple, la discrimination positive.

Les concepteurs de systèmes d'IA ont réfléchi à la façon de traduire l'équité dans leurs systèmes. Aux différentes définitions de l'équité correspondent différentes approches possibles (Narayanan, 2018[18]) :

- **L'approche basée sur « l'ignorance »**, dans le cadre de laquelle un système d'IA doit ignorer tout facteur identifiable, va de pair avec l'approche juridique de l'équité individuelle. Dans ce cas, le système d'IA ne prend pas en considération les données concernant des caractéristiques sensibles ou interdites de traitement, telles que le sexe, la race et l'orientation sexuelle (Yona, 2017[19]). Toutefois, de nombreux autres facteurs peuvent être en corrélation avec une caractéristique dont le traitement est protégé/interdit (comme le sexe), et les supprimer pourrait réduire la précision d'un système d'IA.
- **L'équité basée sur la connaissance** tient compte des différences entre les groupes et vise à traiter des individus similaires de la même manière. Le défi consiste néanmoins à déterminer qui traiter sur un pied d'égalité avec qui. Pour cerner quels individus devraient être considérés comme similaires aux fins d'une tâche particulière, il faut connaître des caractéristiques sensibles.
- **Les approches basées sur l'équité de groupe** s'attachent à garantir que les résultats ne diffèrent pas systématiquement pour les personnes appartenant à des groupes distincts. On craint en effet que les systèmes d'IA puissent être inéquitables, en perpétuant ou en renforçant les biais traditionnels, car ils reposent souvent sur des ensembles de données qui portent la marque des pratiques passées.

Des notions de l'équité différentes donnent des résultats différents pour les divers groupes de la société et les divers types de parties prenantes. Ils ne peuvent tous être atteints simultanément. Ce sont des considérations, et parfois des choix, politiques qui doivent éclairer les choix en matière de conception technologique susceptibles de nuire à des groupes spécifiques.

L'application de l'IA aux ressources humaines donne une illustration des biais qu'elle peut introduire et des problèmes qui en résultent

Dans le domaine des ressources humaines, l'utilisation de l'IA perpétue les biais de recrutement, ou aide au contraire à mettre au jour et à réduire ceux qui ont des effets préjudiciables. Une étude menée par l'université Carnegie Mellon au sujet des tendances observées en ce qui concerne les offres d'emplois publiées sur l'internet a montré qu'une annonce publiée pour un poste de cadre bien rémunéré était présentée 1 816 fois à des hommes et 311 fois seulement à des femmes (Simonite, 2018[20]). Ainsi, un domaine de collaboration potentiel entre les humains et l'IA est la recherche de la transparence des applications de l'IA utilisées pour le recrutement et l'évaluation. Ces applications ne devraient pas codifier de biais, par exemple en disqualifiant automatiquement les candidatures issues de la diversité lorsqu'il s'agit de pourvoir un emploi dans un domaine jusque-là fermé à celle-ci (OCDE, 2018[21]).

Plusieurs approches peuvent aider à réduire la discrimination dans les systèmes d'IA

Les approches proposées pour réduire la discrimination dans les systèmes d'IA incluent la sensibilisation ; les politiques et pratiques organisationnelles relatives à la diversité ; les normes ; les solutions techniques permettant de détecter et de corriger les biais algorithmiques ; et les approches d'auto-réglementation ou de réglementation. Par exemple, dans le cadre des systèmes de prévision policière, certains proposent le recours à des études ou à des déclarations d'impact algorithmique. Il s'agirait, pour les services de police, d'évaluer l'efficacité, les avantages et les éventuels effets discriminatoires de l'ensemble des options technologiques qui s'offrent à eux (Selbst, 2017[22]). La responsabilité et la transparence sont importantes pour atteindre l'équité. Cependant, même combinées, elles ne la garantissent pas (Weinberger, 2018[23]) ; (Narayanan, 2018[18]).

Les efforts visant à atteindre l'équité dans les systèmes d'IA peuvent nécessiter des compromis

On attend des systèmes d'IA qu'ils soient « équitables ». Cela doit se traduire, par exemple, par le fait que seuls les prévenus les plus dangereux restent en prison ou que seul le prêt le plus approprié au regard de la capacité de remboursement soit proposé. Les **erreurs de type I** (ou faux positifs) signalent la classification erronée d'une personne ou d'un comportement. Par exemple, les systèmes peuvent prédire à tort qu'un prévenu récidivera alors qu'il ne le fera pas. De même, ils peuvent se tromper en prédisant une maladie qui n'a pas lieu d'être. Les **erreurs de type II** (ou faux négatifs) se rencontrent dans les cas où un système d'IA prédit à tort, par exemple, qu'un prévenu ne récidivera pas. Un autre exemple serait un test qui indiquerait, à tort, l'absence d'une maladie.

Les approches de l'équité de groupe tiennent compte de points de départ différents selon les groupes. Elles tentent de rendre compte des différences sur le plan mathématique en garantissant une « précision égale » ou un taux d'erreur identique entre tous les groupes. Par exemple, elles classeraient à tort le même pourcentage d'hommes et de femmes en tant que récidivistes (ou égaliseraient les faux positifs et les faux négatifs).

Égaliser les faux positifs et les faux négatifs entraîne une difficulté. Les faux négatifs sont souvent considérés comme plus indésirables et risqués que les faux positifs parce que plus préjudiciables (Berk et Hyatt, 2015[24]). Par exemple, le coût pour une banque d'un prêt fait à un individu dont un système d'IA a prédit qu'il pourrait rembourser – mais qui ne le peut pas – est supérieur au gain tiré de ce prêt. Un individu qu'un mauvais diagnostic a déclaré indemne d'une maladie alors qu'il en est bel et bien atteint peut endurer de grandes souffrances. Égaliser les faux positifs et les faux négatifs peut aussi entraîner des effets indésirables, tels que le fait d'incarcérer des femmes qui ne présentent aucune menace pour la sécurité pour parvenir à libérer la même proportion d'hommes et de femmes (Berk et Hyatt, 2015[24]). Certaines approches visent, par exemple, à égaliser les faux positifs et les faux négatifs en même temps. Cependant, il est difficile de satisfaire à différentes notions d'équité simultanément (Chouldechova, 2016[25]).

Les responsables politiques pourraient réfléchir à un traitement approprié des données sensibles dans le contexte de l'IA

Il pourrait être opportun de revenir sur la question du traitement à réserver aux données sensibles. Dans certains cas, des organisations peuvent avoir besoin de garder et d'utiliser des données sensibles pour assurer que leurs algorithmes ne reconstruisent pas ces données sans que l'on y prenne garde. Une autre priorité d'action est de surveiller chaînes de réaction imprévues. Ainsi, lorsque la police se rend dans des quartiers identifiés par des algorithmes comme ayant une criminalité élevée, cela pourrait conduire à une collecte de données faussées et introduire, par la suite, un biais dans l'algorithme – et dans la société – contre ces quartiers (O'Neil, 2016[26]).

Transparence et explicabilité

La transparence sur l'utilisation de l'IA et le fonctionnement des systèmes d'IA est essentielle

Le terme « transparence » n'a pas la même signification sur les plans technique et politique. Pour les responsables de l'élaboration des politiques, la transparence concerne traditionnellement la façon dont une décision est prise, les participants au processus et les facteurs entrant dans

la prise de décision (Kosack et Fung, 2014[27]). Sous cet angle, les mesures de transparence pourraient révéler comment l'IA est actuellement utilisée dans le cadre d'une prévision, d'une recommandation ou d'une décision. Elles pourraient en outre consister à informer l'usager que son interlocuteur est un système d'IA lorsque tel est le cas.

Pour les technologues, la transparence d'un système d'IA concerne principalement les questions liées aux processus. Il s'agit de permettre aux individus de comprendre comment un système d'IA est développé, entraîné et mis en place. Il peut s'agir également d'expliciter les facteurs qui influent sur une prévision ou une décision spécifique. En général, cela ne passe pas par le partage de code ou d'ensembles de données précis. Dans de nombreux cas, les systèmes sont trop complexes pour que ces éléments apportent une transparence digne de ce nom (Wachter, Mittelstadt et Russell, 2017[28]). En outre, la divulgation de ces renseignements pourrait entraîner celle de secrets commerciaux ou de données sensibles d'utilisateurs.

Plus généralement, on considère qu'il est important de faire connaître et comprendre les systèmes de raisonnement employés en IA pour que cette technologie soit acceptée de tous et utile à tous.

Les approches de la transparence dans les systèmes d'IA

Des experts du *Berkman Klein Center Working Group on Explanation and the Law* (Groupe de travail du Centre Berkman Klein sur l'explication et la législation), de l'Université de Harvard, définissent des approches visant à améliorer la transparence des systèmes d'IA, et notent que chacune implique des compromis (Doshi-Velez et al., 2017[29]). Une approche supplémentaire réside dans la transparence de l'optimisation, c'est-à-dire la transparence sur les objectifs d'un système d'IA et les résultats obtenus en lien avec ceux-ci. Ces approches sont : i) les garanties théoriques ; ii) les données empiriques ; et iii) l'explication (Tableau 4.1).

Tableau 4.1. Approches visant à améliorer la transparence et la responsabilité dans les systèmes d'IA

Approche	Description	Contextes bien adaptés	Contextes peu adaptés
Garanties théoriques	Dans certaines situations, il est possible de donner des garanties théoriques à propos d'un système d'IA étayées par des preuves.	L'environnement est entièrement observable (ex., le jeu de Go) et le problème comme la solution peuvent être formalisés.	La situation ne peut pas être décrite avec précision (la plupart des situations en conditions réelles).
Preuves statistiques/ probabilité	Les données empiriques mesurent la performance globale d'un système, montrant si le système est bénéfique ou néfaste, sans expliquer les décisions particulières.	Les résultats peuvent être entièrement formalisés ; il est acceptable d'attendre de voir les résultats négatifs pour les mesurer ; les problèmes peuvent n'apparaître que dans les agrégats.	L'objectif ne peut pas être entièrement formalisé ; il est possible d'établir les responsabilités à l'égard d'une décision donnée.
Explication	Les humains peuvent interpréter des informations concernant la logique suivie par un système pour traiter un ensemble particulier d'entrées et atteindre une conclusion spécifique.	Les problèmes ne sont pas intégralement spécifiés, les objectifs ne sont pas clairs et les entrées peuvent être erronées.	D'autres formes de responsabilité sont possibles.

Source : adapté de Doshi-Velez et al. (2017[29]), « Accountability of AI under the law: The role of explanation », https://arxiv.org/pdf/1711.01134.pdf.

Certains systèmes offrent des garanties théoriques sur leurs contraintes d'exploitation

Dans certains cas, il est possible d'apporter des **garanties théoriques**, qui indiqueront que le système fonctionnera de manière visible dans le cadre de contraintes bien précises. Les garanties théoriques s'appliquent aux situations dans lesquelles l'environnement est entièrement observable et le problème comme la solution peuvent être intégralement formalisés, comme dans le jeu de Go. Dans de telles situations, certains types de résultats ne peuvent pas être obtenus, même si un système d'IA traite de nouveaux genres de données. Par exemple, un système pourrait être conçu pour suivre, de manière prouvée, les processus définis d'un commun accord pour un vote et le décompte des voix. Dans ce cas, il n'est pas forcément nécessaire d'apporter des explications ou des preuves : le système n'a pas besoin d'expliquer comment il est parvenu à un résultat parce que les types de résultats qui suscitent l'inquiétude sont mathématiquement impossibles. Il est possible de réaliser une étude dès le départ pour déterminer si ces contraintes sont suffisantes.

Des preuves statistiques de la performance globale peuvent être fournies dans certains cas

Dans certains cas, il peut être suffisant de se baser sur les **preuves statistiques** de la performance globale d'un système. Apporter la preuve qu'un système d'IA donné accroît de manière sensible tel bienfait ou tel préjudice pour la société ou les individus peut constituer une garantie de responsabilité suffisante. Par exemple, un système autonome d'atterrissage pour les aéronefs peut causer moins d'incidents de sécurité que des pilotes humains, ou un outil d'aide au diagnostic clinique réduire la mortalité. Les preuves statistiques pourraient constituer un mécanisme de responsabilité approprié pour de nombreux systèmes d'IA, parce que ce mécanisme protège les secrets commerciaux en plus d'être capable de repérer les préjudices répandus mais à faible risque qui n'apparaissent que dans les agrégats (Barocas et Selbst, 2016[30] ; Crawford, 2016[31]). Les questions de biais ou de discrimination peuvent être vérifiées statistiquement : par exemple, un système d'approbation de prêt présenterait un biais s'il approuvait davantage de prêts pour les hommes que pour les femmes lorsque les autres facteurs sont neutralisés. Le taux d'erreur acceptable et l'incertitude tolérée varient selon l'application. Par exemple, le taux d'erreur jugé acceptable pour un outil de traduction ne le sera peut-être pas pour la conduite autonome ou des examens médicaux.

La transparence de l'optimisation est la transparence des objectifs et des résultats d'un système

Une autre approche de la transparence des systèmes d'IA propose que la gouvernance porte son attention non plus sur les moyens mais sur les finalités d'un système. Il s'agit non plus d'exiger l'explicabilité du fonctionnement interne d'un système mais de mesurer ses résultats, c'est-à-dire ce pour quoi le système est « optimisé ». Cela nécessiterait une déclaration concernant ce pour quoi un système d'IA est optimisé, sachant que les optimisations sont imparfaites, qu'elles entraînent des compromis et doivent être limitées par des « contraintes majeures » telles que la sûreté et l'équité. Cette approche préconise d'utiliser les systèmes d'IA pour faire ce pour quoi ils sont optimisés. Elle s'appuie sur les cadres éthiques et juridiques existants, ainsi que sur des débats sociaux et des processus politiques si besoin est pour fournir des informations sur les domaines pour lesquels les systèmes d'IA devraient être optimisés (Weinberger, 2018[1]).

L'explication concerne un résultat précis d'un système d'IA

L'explication est indispensable pour les situations dans lesquelles une anomalie doit être déterminée dans une instance spécifique – situation qui risque de devenir de plus en plus fréquente à mesure que des systèmes d'IA sont mis en place pour formuler des recommandations ou prendre des décisions actuellement laissées à la discrétion de l'homme (Burgess, 2016[32]). Le RGPD exige que les personnes concernées reçoivent des informations utiles concernant la logique sous-jacente, ainsi que l'importance et les conséquences prévues des systèmes de prise de décision automatisée. En général, l'explication n'a pas besoin de présenter le processus de prise de décision du système dans sa totalité. La plupart du temps, il suffit de répondre à l'une des questions ci-dessous (Doshi-Velez et al., 2017[29]) :

1. **Les principaux facteurs d'une décision** : Pour toutes sortes de décisions, concernant, par exemple, les audiences relatives à la garde des enfants, les conditions à remplir pour obtenir un prêt ou une mise en liberté provisoire, divers facteurs doivent être pris en considération (ou au contraire formellement proscrits). Dresser une liste des facteurs qui ont compté dans une prévision établie par l'IA – classés, de préférence, par ordre d'importance – peut aider à garantir que les bons facteurs ont été pris en compte.
2. **Les facteurs déterminants, c'est-à-dire les facteurs qui influent de manière décisive sur le résultat** : Il arrive qu'il soit important de savoir si un facteur donné a orienté un résultat. Le fait de changer un facteur donné, tel que la race dans le cadre d'admissions à l'université, peut montrer si le facteur a été utilisé correctement.
3. **Pourquoi deux cas apparemment similaires donnent-ils des résultats différents, ou inversement** ? Il est possible d'évaluer la cohérence et l'intégrité des prévisions basées sur l'IA. Par exemple, le revenu doit être pris en considération lorsqu'il est décidé d'octroyer ou non un prêt, mais il ne saurait être déterminant dans des situations par ailleurs similaires où il n'a pas lieu d'entrer en ligne de compte.

L'explication fait l'objet de recherches actives mais elle entraîne des coûts, et pourrait même nécessiter des compromis

Des recherches techniques sont actuellement menées par des entreprises, des organismes de normalisation, des organisations à but non lucratif et des institutions publiques en vue de la création de systèmes d'IA capables d'expliquer leurs prévisions. Les entreprises travaillant dans des domaines particulièrement réglementés, tels que la finance, la santé et les ressources humaines, cherchent activement à éliminer les éventuels risques financiers, juridiques et d'atteinte à la réputation liés aux prévisions établies par des systèmes d'IA. Par exemple, en 2016, la banque américaine Capital One a constitué une équipe de recherche chargée de trouver des moyens d'améliorer l'explicabilité des techniques d'IA (Knight, 2017[33]). Des entreprises telles que MondoBrain ont conçu des interfaces utilisateur pour aider à expliquer les facteurs significatifs (Encadré 4.5). Des organisations à but non lucratif, telles qu'OpenAI, cherchent des moyens de mettre au point une IA explicable et de vérifier les décisions prises par l'IA. Des recherches financées par les pouvoirs publics sont par ailleurs en cours. Ainsi, la DARPA finance 13 groupes de recherche différents, qui travaillent sur diverses façons d'améliorer l'explicabilité de l'IA.

Dans de nombreux cas, il est possible de générer au moins un de ces types d'explications concernant les résultats des systèmes d'IA. Toutefois, les explications ont un coût. Concevoir un système destiné à fournir une explication peut s'avérer complexe et onéreux. Exiger des explications pour tous les systèmes d'IA peut ne pas être approprié selon leur finalité et peut désavantager les PME en particulier. Les systèmes d'IA doivent souvent être conçus *ex ante* pour fournir un certain type d'explication. Chercher des explications après coup

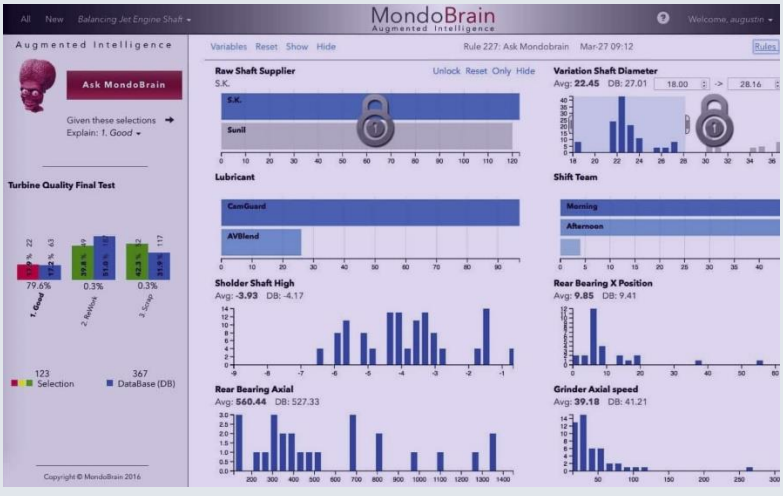
nécessite généralement davantage de travail, potentiellement de recréer l'intégralité du système de décision. Par exemple, un système d'IA ne peut pas fournir une explication pour tous les facteurs majeurs qui ont pesé sur un résultat si sa conception ne lui permet d'en fournir que pour un seul. De même, un système d'IA chargé de dépister les affections cardiaques ne peut être interrogé au sujet des différences de diagnostic entre hommes et femmes si les données ayant servi à son entraînement n'étaient pas ventilées par sexe, et ce même s'il tient bel et bien compte de ce paramètre par le biais de variables indirectes, comme d'autres affections plus fréquentes chez les femmes que chez les hommes.

Encadré 4.5. Régler les problèmes d'explicabilité grâce à des interfaces utilisateur mieux conçues

Certaines entreprises ont commencé à inclure l'explicabilité dans leurs solutions pour que les utilisateurs comprennent mieux les processus d'IA exécutés en arrière-plan. MondoBrain est l'une d'entre elles. Basée en France, elle combine intelligence humaine, collective et artificielle pour fournir une solution de réalité augmentée aux entreprises. Grâce à des tableaux de bord interactifs de visualisation des données, elle évalue l'ensemble des données existantes au sein d'une entreprise (à partir des logiciels de planification des ressources de l'entreprise, de gestion des programmes de l'entreprise ou de gestion des relations avec la clientèle, par exemple) et formule des recommandations normatives sur la base des requêtes des clients (Graphique 4.1). Elle utilise un algorithme d'apprentissage automatique pour éliminer les variables commerciales qui ne présentent pas d'intérêt par rapport à la requête et pour extraire les variables ayant le plus d'impact.

Les couleurs des feux de signalisation guident les utilisateurs à chaque étape de la requête, facilitant leur compréhension du processus de décision. Chaque décision est automatiquement enregistrée et devient vérifiable et traçable. Cela donne un compte rendu complet mais simple de toutes les étapes qui ont conduit à la recommandation commerciale finale.

Graphique 4.1. Illustration des outils de visualisation des données visant à améliorer l'explicabilité



Source : www.mondobrain.com.

Dans certains cas, un compromis doit être trouvé entre explicabilité et précision. Pour être explicables, les variables d'une solution doivent potentiellement être réduites à un ensemble

suffisamment petit pour pouvoir être appréhendé par l'homme. Cela peut s'avérer sous-optimal dans le cadre de problèmes complexes et de grande ampleur. Par exemple, certains modèles d'apprentissage automatique utilisés pour établir un diagnostic médical peuvent prédire avec précision la probabilité d'une maladie, mais sont trop complexes pour être accessibles à l'esprit humain. Dans ce genre de cas, il convient de comparer les préjudices que peut causer un système moins précis qui offre des explications claires avec ceux d'un système plus précis dans lequel les erreurs sont plus difficiles à détecter. Par exemple, la prévision de la récidive peut nécessiter des modèles simples et explicables dans lesquels les erreurs sont décelables (Dressel et Farid, 2018[34]). Dans des domaines tels que les prévisions climatiques, on acceptera plus facilement des modèles plus complexes qui offrent de meilleures prévisions mais sont moins explicables. À plus forte raison s'il existe d'autres mécanismes de responsabilité vis-à-vis des résultats, tels que des données statistiques permettant de détecter un éventuel biais ou une éventuelle erreur.

Robustesse, sûreté et sécurité

Ce qu'il faut entendre par robustesse, sûreté et sécurité

La robustesse peut s'entendre comme la capacité à supporter ou surmonter des conditions défavorables (OCDE, 2019[35]), notamment des risques de sécurité numérique. Les systèmes d'IA pourront être dits « sécurisés » dans la mesure où leur utilisation dans des conditions normales ou prévisibles, y compris si elle est abusive, ne fera jamais peser un risque de sécurité démesuré, quel que soit le stade de leur cycle de vie (OCDE, 2019[36]). Les questions de robustesse et de sécurité de l'IA sont interdépendantes. À titre d'exemple, la sécurité numérique peut avoir une incidence sur la sécurité des produits si les dispositifs connectés, comme les voitures autonomes ou les appareils électroménagers fonctionnant grâce à l'IA ne sont pas suffisamment sécurisés ; des pirates pourraient en prendre le contrôle et en modifier les paramètres à distance.

La gestion des risques dans les systèmes d'IA

Le niveau de protection requis dépend d'une analyse risques-avantages

Il conviendrait de mettre les préjudices susceptibles d'être causés par un système d'IA en regard des coûts qu'il faudrait supporter pour faire la transparence sur ces systèmes et définir les responsabilités y afférentes. Les préjudices potentiels pourraient être des risques pesant sur les droits individuels, la vie privée, l'équité et la robustesse. Toutes les utilisations de l'IA ne s'accompagnent pas des mêmes risques cependant, et exiger une explication, par exemple, génère aussi un certain coût. En matière de gestion des risques, un large consensus semble se dégager autour de l'idée selon laquelle plus les enjeux sont importants plus il faut faire preuve de transparence et de responsabilité, tout particulièrement lorsqu'il y va de la vie ou de la liberté des personnes.

Les stratégies de gestion des risques ont leur place tout au long du cycle de vie des systèmes d'IA

Les organisations ont recours à la gestion des risques pour isoler, évaluer, hiérarchiser et traiter les risques susceptibles d'altérer le comportement d'un système et les résultats attendus de son utilisation. Cette démarche peut également servir à déterminer quels risques pèsent sur les différentes parties prenantes et comment les maîtriser tout au long du cycle de vie du système d'IA considéré (voir au chapitre 1 la section consacrée au cycle de vie des systèmes d'IA).

Les acteurs de l'IA – ceux qui jouent un rôle actif au cours du cycle de vie d'un système d'IA – évaluent et atténuent les risques à l'échelle de ce système pris dans son ensemble, ainsi qu'au cours de chaque phase de son cycle de vie. La gestion des risques dans les systèmes d'IA suit les étapes suivantes, dont l'importance varie selon le stade atteint dans le cycle de vie :

1. **Objectifs** : Définir les objectifs, les fonctions ou les propriétés du système, en contexte. Fonctions et propriétés peuvent changer suivant la phase du cycle de vie.
2. **Parties prenantes et acteurs** : Identifier les parties prenantes et les acteurs concernés, autrement dit ceux qui sont directement ou indirectement intéressés par les fonctions ou les propriétés du système à chaque étape du cycle de vie.
3. **Évaluation des risques** : Évaluer les effets potentiels – avantages et risques – du système pour les parties prenantes et les acteurs. Ces effets varieront en fonction des parties prenantes et des acteurs concernés comme selon la phase de son cycle de vie atteinte par le système d'IA considéré.
4. **Atténuation des risques** : Identifier des stratégies d'atténuation adaptées et proportionnées aux risques. Celles-ci devraient tenir compte de différents paramètres tels que les buts et objectifs de l'entité, les parties prenantes et acteurs concernés, la probabilité d'occurrence du risque et les avantages potentiels.
5. **Mise en œuvre** : Appliquer les stratégies d'atténuation des risques.
6. **Suivi, évaluation et reddition de comptes** : Suivre la mise en œuvre de la stratégie, évaluer ses résultats et en rendre compte.

L'utilisation de la gestion des risques et la consignation des décisions prises à chaque étape du cycle de vie peut contribuer à la transparence d'un système d'IA et à la responsabilisation de l'entité à l'égard de ce système.

Il convient d'apprécier côté à côté l'ampleur du préjudice global et le risque immédiat

Considérées de manière isolée, quelques-unes des utilisations possibles des systèmes d'IA présentent un faible niveau de risque. Elles pourraient cependant nécessiter davantage de robustesse de la part de ces systèmes en raison de leurs effets sur la société. Un système qui, par son fonctionnement, causerait un préjudice mineur à un grand nombre de personnes n'en causerait pas moins un préjudice global significatif pour la collectivité. Supposons, par exemple, qu'un petit nombre d'outils fondés sur l'IA soient intégrés dans une multitude de services et de secteurs et servent pour les demandes de prêt, la souscription de contrats d'assurance ou les enquêtes de moralité. Une seule erreur, un seul biais, introduits dans un système seraient susceptibles d'entraîner une cascade de réponses négatives (Citron et Pasquale, 2014[37]). Ces réponses négatives, prises une à une, ne prêteront probablement pas à conséquence. Leur accumulation, en revanche, pourrait avoir un effet délétère. Il semble par conséquent souhaitable que les décideurs prennent en compte, dans leurs débats, l'ampleur du préjudice global, en plus de considérer le risque immédiat.

La robustesse face aux risques de sécurité numériques associés à l'IA

L'IA permet des attaques plus sophistiquées et d'une envergure potentiellement accrue

L'utilisation de l'IA à des fins malveillantes est appelée à se développer à mesure que celle-ci devient moins onéreuse et plus accessible, et parallèlement à son emploi au service de la sécurité numérique (voir, au chapitre 3, la section sur l'IA dans la sécurité numérique). Les auteurs de cyberattaques s'emploient à renforcer leurs capacités en matière d'IA. La rapidité croissante et la sophistication des attaques ne laissent pas d'inquiéter[6]. Dans ce contexte, on voit se renforcer les menaces existantes tandis qu'il en surgit de nouvelles et que le caractère même des menaces évolue.

Les systèmes d'IA contemporains présentent un certain nombre de vulnérabilités. Des individus malintentionnés peuvent manipuler les données servant à entraîner l'un de ces systèmes (par exemple dans le cas d'un « empoisonnement des données »). Ils peuvent aussi bien découvrir les caractéristiques servant, dans un modèle de sécurité numérique, à détecter les logiciels malveillants et, cette information une fois connue, créer un code malveillant ou causer de manière intentionnelle une classification erronée d'éléments d'information (par exemple en donnant des « exemples contradictoires », Encadré 4.6) (Brundage et al., 2018[38]). Les technologies d'IA devenant de plus en plus accessible, davantage de personnes peuvent les utiliser pour mener des attaques sophistiquées d'une envergure supérieure aux attaques de naguère. La fréquence et l'efficacité des attaques de sécurité numérique nécessitant une préparation méticuleuse, comme c'est le cas avec le harponnage (*spear phishing*), pourraient bien augmenter avec leur automatisation, rendue possible par les algorithmes d'apprentissage automatique.

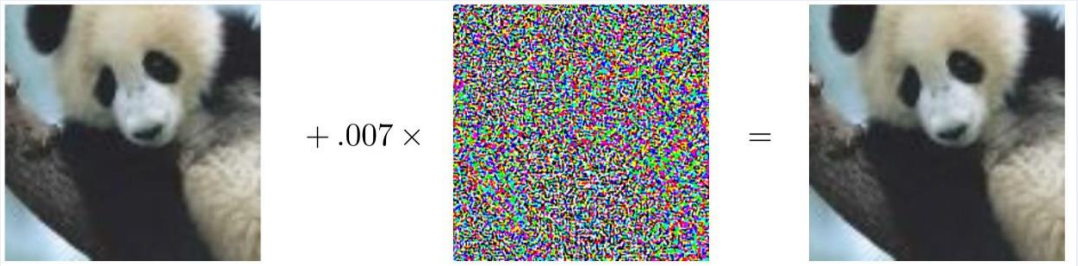
Encadré 4.6. Du danger des exemples contradictoires pour l'apprentissage automatique

On appelle **exemples contradictoires** les éléments introduits à dessein dans des modèles d'apprentissage automatique par des personnes malintentionnées afin d'amener ces modèles à commettre des erreurs présentant toutes les apparences de la fiabilité. Il s'agit d'un réel problème pour la robustesse et la sûreté des systèmes d'IA car plusieurs modèles d'apprentissage automatique, y compris les réseaux neuronaux à la pointe de la technologie, leur sont vulnérables.

Ces exemples contradictoires peuvent être subtils. Ainsi, dans le Graphique 4.2, une modification imperceptible, ou « élément contradictoire », a été ajoutée à l'image d'un panda. Cette modification est destinée à tromper le modèle de classification d'images. Il s'ensuit que l'algorithme confond un panda avec un gibbon avec un niveau de confiance proche de 100 %.

Des recherches récentes ont démontré qui plus est que l'on pouvait créer des exemples contradictoires à partir d'une image imprimée sur papier ordinaire et photographiée à l'aide d'un smartphone avec un niveau de résolution normal. Ces images peuvent être dangereuses : un simple autocollant apposé sur un « stop » pourrait induire une voiture autonome à interpréter ce panneau comme une priorité ou à le confondre avec n'importe quel autre panneau de signalisation.

Graphique 4.2. Trompé par une légère modification, un algorithme confond un panda avec un gibbon



Sources : Goodfellow, Shlens et Szegedy (2015[39]), « Explaining and harnessing adversarial examples », https://arxiv.org/pdf/1412.6572.pdf ; Kurakin, Goodfellow et Bengio (2017[40]), « Adversarial examples in the physical world », https://arxiv.org/abs/1607.02533.

La sûreté

Les systèmes d'apprentissage automatique et les systèmes autonomes bousculent les cadres d'action en place en matière de sûreté

La gamme des équipements dotés d'une intelligence artificielle s'étoffe rapidement – depuis les robots jusqu'aux voitures autonomes en passant par les produits et services grand public comme les appareils et les systèmes d'alarme domestique intelligents. Ces équipements présentent des avantages non négligeables sur le plan de la sûreté mais soulèvent dans le même temps des problèmes d'ordre juridique et pratique en rapport avec les cadres relatifs à la sécurité des produits (OCDE, 2018[21]). Ces cadres tendent à régir des produits « finis » et tangibles davantage que des logiciels, or de nombreux système d'IA continuent d'apprendre et évoluent tout au long de leur cycle de vie[7]. Les produits fondés sur l'IA peuvent aussi être « autonomes » ou « semi-autonomes », autrement dit prendre et appliquer des décisions dans lesquels l'être humain n'intervient pas ou alors seulement de manière limitée.

Aux différents types d'applications de l'IA correspondront vraisemblablement des actions adaptées de la part des pouvoirs publics (Freeman, 2017[41]). D'une manière générale, les systèmes d'IA appellent les décideurs à mener une quadruple réflexion. Il faut premièrement rechercher le meilleur moyen de garantir la sécurité des produits. En d'autres termes, les produits ne doivent pas présenter des risques démesurés pour la sécurité dans des conditions d'utilisation normales ou prévisibles, y compris en cas d'utilisation abusive, et ce tout au long de leur cycle de vie. Cela concerne en particulier les cas où les données disponibles pour entraîner le système d'IA sont peu nombreuses (Encadré 4.7). Deuxièmement, il convient de se demander qui doit être tenu pour responsable, et jusqu'à quel point, des dommages causés par un système d'IA. Il y a lieu en parallèle de se demander quelles sont les parties prenantes qui peuvent concourir à la sûreté des appareils autonomes. Les utilisateurs, les fabricants de produits et de capteurs, les producteurs de logiciels, les concepteurs, les fournisseurs d'infrastructure et les entreprises d'analyse de données pourraient être du nombre. Il importe, troisièmement, de réfléchir au choix du ou des types de responsabilité à appliquer – devrait-il s'agir d'une responsabilité objective ou d'une responsabilité en cas de faute et quel devrait être le rôle dévolu à l'assurance. L'opacité de certaines systèmes d'IA n'aide pas à trancher cette question. Quatrièmement, il y a lieu pour les décideurs de s'interroger sur la manière de donner effet à la législation, sur ce qu'il faut considérer comme un « défaut » dans un produit fondé sur l'IA, sur la définition de la charge de la preuve et sur les voies de recours existantes.

La Directive européenne de 1985 relative à la responsabilité du fait des produits défectueux (Directive 85/374/CEE) pose le principe de « responsabilité sans faute » ou de « responsabilité stricte », en vertu duquel le producteur est responsable des dommages causés au consommateur par un produit défectueux, même en l'absence de négligence ou de faute de sa part. La Commission européenne a entrepris de réviser cette directive. D'après les premières conclusions, le modèle demeure globalement adapté (Ingels, 2017[42]). Il reste que les technologies reposant actuellement sur l'IA et celles que l'on peut attendre dans l'avenir remettent en cause les notions de « produit », de « sûreté », de « défaut » et de « dommage », ce qui tend à rendre plus délicate la question de la charge de la preuve.

Dans le domaine des véhicules autonomes, la sûreté est la préoccupation première des autorités. Des travaux de fond doivent être réalisés au sujet des essais auxquels soumettre ces véhicules en vue de garantir que leur fonctionnement est bien sécurisé. Ces travaux doivent porter notamment sur les régimes de licence qui évaluent la possibilité d'une expérimentation préalable des systèmes installés à bord ou traiter de la nécessité pour ces

systèmes de contrôler la vigilance des conducteurs humains qui peuvent être appelés à reprendre la main en cas de besoin. L'octroi de licences constitue dans certains cas un réel problème pour les entreprises qui souhaitent procéder à des essais de véhicules. Les pouvoirs publics sont d'autre part plus ou moins favorables à la réalisation de tels essais. D'aucuns ont appelé à l'application d'un régime de responsabilité stricte à l'égard des constructeurs de véhicules autonomes, la responsabilité dépendant alors du caractère contrôlable ou non du risque. Il serait reconnu par exemple qu'un simple passager d'une voiture sans chauffeur ne saurait être tenu pour fautif ou manquer à un devoir de vigilance. Les juristes sont d'avis que même le concept de « détenteur déclaré » ne serait pas applicable car le détenteur doit être en mesure de maîtriser le risque (Borges, 2017[43]). L'idée a été émise que les compagnies d'assurance pourraient prendre en charge le risque de dommage du fait des voitures autonomes à partir d'une classification des véhicules déclarés établie sur la base d'évaluations des risques.

Encadré 4.7. Les données synthétiques au service d'une IA plus sûre et plus précise : le cas des véhicules autonomes

L'utilisation des données synthétiques devient de plus en plus courante dans le domaine de l'apprentissage automatique car elle permet de simuler des scénarios difficilement observables ou reproductibles en conditions réelles. Selon les explications de Philipp Slusallek, directeur scientifique du Centre allemand de recherche sur l'IA, il s'agit par exemple de s'assurer par ce moyen qu'une voiture autonome ne percutera pas un enfant qui traverserait la rue en courant.

La « réalité numérique » – un environnement simulé répliquant les caractéristiques pertinentes du monde réel – pourrait avoir quatre effets. Premièrement, elle pourrait fournir les données synthétiques à partir desquelles apprendre aux systèmes d'IA à faire face à des situations complexes. Deuxièmement, elle permettrait la validation des caractéristiques de fonctionnement et le recalibrage des données synthétiques par rapport aux données obtenues en conditions réelles. Troisièmement, elle pourrait servir à l'organisation d'examens, comme celui du permis de conduire pour les conducteurs de véhicule autonome. Elle permettrait en quatrième lieu d'explorer le processus décisionnel suivi par le système et de découvrir les conséquences potentielles des autres choix possibles. Cette méthode a ainsi permis à Google d'entraîner ses voitures autonomes en leur faisant parcourir, en simulation, plus de 4.8 millions de kilomètres par jour (soit plus de 500 allers-retours entre New-York et Los Angeles).

Sources : Golson (2016[44]), « Google's self-driving cars rack up 3 million simulated miles every day », https://www.theverge.com/2016/2/1/10892020/google-self-driving-simulator-3-million-miles ; Slusallek (2018[45]), *Artificial Intelligence and Digital Reality: Do We Need a CERN for AI?*, https://www.oecd-forum.org/channels/722-digitalisation/posts/28452-artificial-intelligence-and-digital-reality-do-we-need-a-cern-for-ai.

Les normes de sécurité au travail demanderont sans doute à être mises à jour

Il est probable qu'entre autres conséquences directes sur les conditions de travail, l'IA va rendre nécessaire l'introduction de nouveaux protocoles de sécurité. L'adoption ou la révision de normes sectorielles et d'accords d'entreprise à portée technologique vont devenir inévitables pour garantir que les conditions de fiabilité et de sûreté nécessaires à la productivité sont bien réunies en milieu de travail. Le Comité économique et social européen (CESE) préconise que les « parties prenantes œuvrent ensemble en faveur de systèmes d'IA complémentaires et de leur mise en place conjointe sur le lieu de travail » (CESE, 2017[46]).

Responsabilité

L'utilisation croissante de l'IA doit s'accompagner d'un effort en matière de responsabilité, garant du bon fonctionnement des systèmes

La notion de **responsabilité** désigne essentiellement le fait de savoir attribuer à chacun, organisation ou individu, la part qui lui revient dans le bon fonctionnement des systèmes d'IA. Les critères sur lesquels elle repose sont le respect des valeurs humaines et de l'équité, de la transparence, de la robustesse et de la sûreté. La responsabilité dépend du rôle de chacun des acteurs de l'IA, du contexte et de l'état de la technologie. Du point de vue des responsables de l'action publique, elle dépend de mécanismes remplissant plusieurs offices. Ces mécanismes identifient la partie qui est comptable de telle recommandation ou de telle décision. Ils corrigent ladite recommandation ou décision avant sa mise à exécution. Il est possible également par ce moyen de contester la décision, ou d'en faire appel, à un stade ultérieur, ou même de récuser le système dont elle est issue (Helgason, 1997[47]).

Dans la pratique, la responsabilité dans les systèmes d'IA dépend souvent du fonctionnement d'un système donné au regard d'indicateurs de précision ou d'efficience, auxquels viennent aujourd'hui s'ajouter de plus en plus fréquemment des indicateurs d'équité, de sûreté et de robustesse. Les seconds cependant restent moins utilisés que les premiers. Comme pour tous les indicateurs, le suivi et l'évaluation peuvent s'avérer coûteux. Aussi le type et la fréquence des relevés doivent-ils être proportionnés aux risques et avantages potentiels.

Le niveau de responsabilité requis dépend du niveau de risque

Les stratégies à suivre sont fonction du contexte et des circonstances. À titre d'exemple, une responsabilité relativement élevée sera sans doute attendue, en matière d'utilisation de l'IA, de la part du secteur public, en particulier dans l'exercice de fonctions régaliennes, telles la sécurité ou l'exécution des lois, d'où peuvent découler des préjudices importants. Des mécanismes de responsabilité formels sont par ailleurs souvent requis à l'égard des applications développées par le secteur privé dans les domaines du transport, de la finance et des soins de santé, qui sont strictement encadrés. Dans d'autres domaines soumis à un contrôle moins sévère, l'utilisation de l'IA s'accompagne plus rarement de semblables mécanismes. Dans ces cas, les stratégies techniques prennent d'autant plus d'importance en matière de transparence et de responsabilité. Elles doivent garantir que les systèmes conçus et exploités par des acteurs du secteur privé respectent un certain nombre de normes sociétales et de contraintes légales.

Certaines applications ou décisions pourraient requérir l'intervention d'un « élément humain » chargé d'apprécier le contexte social dans lequel elles s'inscrivent et leurs potentiels effets indésirables. Lorsqu'une décision emporte des conséquences significatives sur le quotidien des individus, il est communément admis qu'elle ne devrait pas être prise uniquement sur la base d'un résultat fourni par l'IA (par exemple un score). Le Règlement général sur la protection des données préconise ainsi qu'il y ait en pareil cas une intervention humaine. À titre d'exemple, les individus doivent être informés lorsque l'IA est utilisée pour rendre un jugement, accorder ou refuser un prêt, décider de l'orientation d'élèves ou d'étudiants ou sélectionner des candidats à un poste. Lorsque les enjeux sont importants, des mécanismes de responsabilité formels sont souvent exigés. À titre d'exemple, un magistrat qui s'appuie sur l'IA pour prononcer une condamnation constituera l'« élément humain » intervenant dans le processus. Cependant, l'existence d'autres mécanismes de responsabilité – dont la possibilité de faire appel du jugement comme dans une procédure traditionnelle – contribue à garantir que les recommandations formulées par l'IA soient bien prises comme un

élément d'appréciation parmi d'autres (Wachter, Mittelstadt et Floridi, 2017[48]). En l'absence de risque particulier, par exemple lorsqu'il s'agit de recommander un restaurant, la machine seule suffira. Il n'y aura sans doute pas lieu en l'occurrence de multiplier les strates au risque de générer des coûts superflus.

Cadre d'action applicable à l'IA

Des politiques nationales doivent être mises en œuvre pour promouvoir des systèmes d'IA dignes de confiance. Ces mesures peuvent entraîner des effets bénéfiques et équitables pour les individus et pour la planète, en particulier dans des domaines prometteurs actuellement sous-investis par le marché. La création d'un environnement réglementaire propice à l'avènement d'une IA à laquelle on puisse se fier sans crainte suppose, entre autres choses, de favoriser l'investissement public et privé dans les activités de recherche-développement connexes et de faire acquérir aux individus les compétences qui leur permettront de s'adapter avec succès à l'évolution des emplois. Les paragraphes qui suivent sont consacrés à quatre domaines d'action essentiels à la promotion et au développement d'une IA qui mérite toute notre confiance.

Investissement dans la recherche et le développement en matière d'IA

L'investissement à long terme dans la recherche publique peut aider à façonner l'innovation en matière d'IA

L'OCDE s'intéresse au rôle des politiques d'innovation dans la transformation numérique et l'adoption de l'IA (OCDE, 2018[49]). À ce titre, elle étudie notamment le rôle des politiques en faveur de la recherche publique, du transfert de connaissances et de la création conjointe, à l'appui du développement des outils et des infrastructures de recherche pour l'IA. L'intelligence artificielle oblige les décideurs à réévaluer le niveau d'intervention idoine des pouvoirs publics dans la recherche connexe pour relever les défis sociétaux (OCDE, 2018[14]). En outre, les établissements de recherche dans tous les domaines devront se doter de systèmes d'IA robustes pour rester compétitifs, en particulier dans des domaines comme la science biomédicale et la biologie. Des instruments émergents, à l'instar des plateformes de partage des données et des installations de superinformatique, peuvent aider à stimuler la recherche dans l'IA et pourraient appeler de nouveaux investissements. Le Japon, par exemple, consacre plus de 120 millions USD par an à la construction d'une infrastructure de calcul hautes performances pour les universités et les centres publics de recherche.

L'IA est considérée comme une technologie générique susceptible d'avoir des incidences sur de nombreux secteurs (Agrawal, Gans et Goldfarb, 2018[50] ; Brynjolfsson, Rock et Syverson, 2018[51]). Elle est également vue comme l'« invention d'une méthode d'invention » (Cockburn, Henderson et Stern, 2018[52]), déjà largement utilisée par les chercheurs et les inventeurs pour faciliter l'innovation. Sans compter que des secteurs entièrement nouveaux pourraient voir le jour grâce à des percées scientifiques faisant fond sur l'IA. D'où l'importance de la recherche fondamentale et la nécessité d'inscrire les politiques de recherche dans une vision à long terme (OCDE, 2018[53]).

Favoriser l'instauration d'un écosystème numérique propice à l'IA

Technologies et infrastructure d'IA

Des progrès significatifs ont été réalisés ces dernières années dans les technologies liées à l'IA. On les doit à la maturité des techniques de modélisation statistique, comme les réseaux neuronaux, en particulier les réseaux neuronaux profonds (on parle d'« apprentissage

profond »). Nombre des outils employés pour gérer et utiliser l'IA sont des ressources à code source libre, qui relèvent du domaine public. Cela facilite leur adoption et permet de corriger les bogues logiciels à l'aide de solutions issues de contributions participatives. TensorFlow (de Google), Michelangelo (d'Uber) et Cognitive Toolkit (de Microsoft) en sont des exemples. Par ailleurs, des entreprises et des chercheurs partagent publiquement des ensembles de données d'entraînement après curation et des outils d'apprentissage afin de favoriser la diffusion des technologies liées à l'IA.

Une partie des avancées récentes de l'IA s'explique par l'accélération exponentielle des temps de traitement et la Loi de Moore (selon laquelle le nombre de transistors sur un circuit intégré à haute densité double tous les deux ans environ). Grâce à la conjugaison de ces deux phénomènes, les algorithmes d'IA peuvent traiter rapidement des volumes considérables de données. À mesure que les projets d'IA passent du concept à l'application commerciale, les besoins de ressources spécialisées et onéreuses d'infonuagique et de processeurs graphiques vont croissant. L'essor des systèmes d'IA s'accompagne également d'une augmentation fulgurante de la puissance de calcul requise. Selon une estimation, l'expérience la plus importante menée récemment, AlphaGo Zero, a nécessité une puissance de calcul 300 000 fois supérieure à celle utilisée pour mener à bien l'expérience la plus importante six ans plus tôt (OpenAI, 16 mai 2018[54]). De même, les prouesses du programme d'échecs et de Go, AlphaGo Zero, ont mobilisé une puissance de calcul qui dépasserait celle des dix superordinateurs les plus puissants du monde conjugués (OCDE, 2018[53]).

Accessibilité et utilisation des données

L'accessibilité et le partage des données peuvent accélérer ou, selon le cas, freiner les progrès de l'IA

Les technologies actuelles d'apprentissage automatique requièrent, pour s'entraîner et évoluer, des données fiables ayant fait l'objet d'une curation. L'accès à des ensembles de données de qualité s'avère par conséquent essentiel au développement de l'IA. Les facteurs liés à l'accessibilité et au partage des données susceptibles d'accélérer ou, selon le cas, de freiner les progrès de l'IA, sont les suivants (OCDE, 2019[55]) :

- **Normes** *:* Les normes sont nécessaires pour permettre l'interopérabilité et la réutilisation des données d'une application à l'autre, favoriser l'accessibilité et garantir que les données puissent être trouvées, intégrées à des catalogues et/ou interrogées et réutilisées.
- **Risques** : Les risques liés au partage des données, qui pèsent sur les individus, les organisations et les pays, peuvent aller de la violation de la confidentialité et de la vie privée, au non-respect des droits de propriété intellectuelle (DPI), en passant par la menace des intérêts commerciaux ou la compromission de la sécurité nationale et de la sécurité numérique.
- **Coûts des données** : La collecte, l'accès, le partage et la réutilisation nécessitent des investissements en amont et en aval. Outre ceux liés à l'acquisition des données, des investissements supplémentaires doivent être consacrés à leur nettoyage, à leur curation, à la maintenance des métadonnées, au stockage et au traitement des données, et à la sécurisation de l'infrastructure informatique.
- **Incitations** : Les approches fondées sur le marché peuvent encourager à ouvrir l'accès aux données et à les partager avec des marchés et des plateformes qui commercialisent des données et proposent des services à valeur ajoutée, comme les infrastructures de paiement et d'échange de données.

- **Incertitudes quant à la propriété des données** : Les cadres juridiques – régimes de propriété intellectuelle et droit (cyber)criminel, de la concurrence et de la protection de la vie privée –, conjugués à la multiplicité des parties intervenant dans la création des données, créent des incertitudes quant à la « propriété des données ».
- **Autonomisation des utilisateurs, y compris des agents intelligents** : Donner aux utilisateurs les moyens d'agir et faciliter la portabilité des données – tout en mettant en place des mécanismes efficaces de consentement et de choix pour les personnes concernées par les données – peuvent inciter les individus et les entreprises à partager des données personnelles ou professionnelles. D'aucuns soulignent en outre que les agents intelligents qui connaissent les préférences des individus pourraient les aider à négocier des dispositifs complexes de partage des données avec d'autres systèmes d'IA (Neppel, 2017[56]).
- **Tiers de confiance** *:* Les tierces parties peuvent aider à instaurer la confiance et faciliter le partage et la réutilisation des données entre l'ensemble des parties prenantes. Les intermédiaires de données peuvent agir en tant qu'autorités de certification. Les plateformes de confiance spécialisées dans le partage des données, à l'image des fiducies de données, fournissent des données de qualité. Sans oublier les comités d'évaluation éthique, qui veillent au respect des intérêts légitimes des tierces parties.
- **Représentativité des données** : Les systèmes d'IA établissent des prévisions d'après des schémas identifiés dans les ensembles de données d'entraînement. C'est pourquoi, dans une optique à la fois d'exactitude et d'équité, les ensembles de données d'entraînement doivent être inclusifs, divers et représentatifs afin de ne pas sous-représenter ni sur-représenter des groupes particuliers.

Les politiques publiques peuvent favoriser l'accessibilité et le partage des données à l'appui du développement de l'IA

Plusieurs stratégies sont envisageables pour renforcer l'accessibilité et le partage des données (OCDE, 2019[55]) :

- **Favoriser l'accès aux données du secteur public**, qu'il s'agisse de données publiques ouvertes, de données géographiques (des cartes, par exemple) ou de données liées aux transports.
- **Faciliter le partage des données dans le secteur privé**, soit selon un principe de volontariat, soit en application de dispositions obligatoires, auquel cas le partage des données se fait exclusivement avec des utilisateurs de confiance. Certains domaines appellent une attention particulière, à l'instar des « données d'intérêt général », des données relevant d'industries de réseau comme les transports et l'énergie, pour l'interopérabilité des services, et de la portabilité des données à caractère personnel.
- **Développer les capacités statistiques/analytiques**, en mettant en place des centres de technologie qui fournissent un soutien et des conseils en matière d'utilisation et d'analyse des données.
- **Définir des stratégies nationales en matière de données**, afin d'assurer la cohérence des cadres nationaux de gouvernance des données et leur compatibilité avec les stratégies nationales en matière d'IA.

Des approches techniques voient le jour pour remédier aux contraintes liées aux données

Certains algorithmes d'apprentissage automatique, tels que ceux appliqués à la reconnaissance d'images, affichent des performances supérieures aux capacités humaines moyennes. Toutefois, pour y parvenir, ils devaient jusqu'à présent être entraînés à l'aide de bases de données colossales contenant des millions d'images étiquetées. Les besoins en données ont encouragé la recherche active dans des techniques d'apprentissage automatique qui requièrent moins de données pour entraîner les systèmes d'IA. Plusieurs méthodes peuvent aider à parer au manque de données.

- **L'apprentissage profond par renforcement** est une technique d'apprentissage automatique qui allie des réseaux neuronaux profonds et l'apprentissage par renforcement (voir chapitre 1, sous-section « Volet 2 : Techniques d'apprentissage automatique »). Ce faisant, il apprend à privilégier un comportement donné menant au résultat recherché (Mousave, Schukat et Howley, 2018[57]). Des « agents » intelligents rivalisent en exécutant des actions dans un environnement complexe et reçoivent soit une « récompense », soit une « pénalité », selon que l'action a mené au résultat souhaité ou non. Les agents ajustent leurs actions à la lumière de ces « retours d'information »[8].
- **L'apprentissage par transfert ou le pré-entraînement** (Pan et Yang, 2010[58]) réutilise des modèles qui ont été entraînés, en vue d'exécuter des tâches différentes dans le même domaine. Par exemple, certaines couches d'un modèle entraîné à reconnaître des images de chats pourraient être réutilisées pour détecter des images de robes bleues. La taille de l'échantillon d'images s'avérerait alors bien inférieur à ce qu'exigent les algorithmes d'apprentissage automatique traditionnels (Jain, 2017[59]).
- **L'apprentissage fondé sur des données augmentées**, ou synthétisation de données, peut créer artificiellement des données à l'aide de simulations ou d'interpolations à partir de données existantes. Cette technique permet d'accroître le volume de données et, partant, d'améliorer l'apprentissage. Elle s'avère particulièrement intéressante lorsque les contraintes liées au respect de la vie privée limitent l'utilisation des données ou pour simuler des scénarios qui se produisent extrêmement rarement dans la réalité (Encadré 4.7)[9].
- **Les modèles d'apprentissage hybride** peuvent modéliser l'incertitude en alliant différents types de réseaux neuronaux profonds et des approches probabilistes ou bayésiennes. Cette modélisation de l'incertitude a pour objectif d'améliorer les performances et l'explicabilité, et de réduire la probabilité d'obtenir des erreurs de prévisions (Kendall, 23 mai 2017[60]).

Les préoccupations quant à la protection de la vie privée, la confidentialité et la sécurité pourraient avoir pour effet de limiter l'accessibilité et le partage des données. De là peut naître un décalage entre la rapidité d'apprentissage des systèmes d'IA et la disponibilité des ensembles de données utilisés pour les entraîner. Les progrès récents des techniques de cryptographie, comme le calcul multipartite sécurisé (CMS) et le chiffrement homomorphe, pourraient permettre de réaliser des analyses de données tout en garantissant le respect des droits connexes. De fait, les systèmes d'IA pourraient alors opérer sans collecter des données sensibles ni devoir y accéder (Encadré 4.8). Les modèles d'IA sont par ailleurs de plus en plus à même de travailler avec des données chiffrées[10]. Toutefois, dans la mesure où ces solutions nécessitent une puissance de calcul considérable, il peut s'avérer difficile de les déployer à grande échelle (Brundage et al., 2018[38]).

Encadré 4.8. Les nouveaux outils cryptographiques permettent d'exécuter des calculs tout en préservant la vie privée

Les progrès de la cryptographie ouvrent la voie à des applications prometteuses dans le domaine de l'IA. Par exemple, un modèle d'apprentissage automatique pourrait être entraîné à l'aide d'une combinaison de données issues de diverses organisations. Ce faisant, les données de l'ensemble des participants resteraient confidentielles. Une telle solution lèverait les obstacles liés aux problématiques de respect de la vie privée et de confidentialité. Les techniques de chiffrement qui permettent ce type de traitement ne sont pas nouvelles : si le chiffrement homomorphe existe depuis des années, le calcul multipartite sécurisé remonte quant à lui à plusieurs décennies. Pour autant, elles n'étaient pas jusqu'à présent suffisamment efficientes pour une utilisation pratique. Grâce aux progrès récents des algorithmes et de la mise en œuvre, elles deviennent peu à peu des outils fonctionnels capables d'exécuter des analyses sur des ensembles de données réels.

- **Chiffrement homomorphe** : Technique permettant d'exécuter des calculs sur des données chiffrées, sans avoir besoin de disposer des données non chiffrées.
- **Calcul multipartite sécurisé (CMS)** : Technique permettant de calculer une fonction de données collectées à partir de diverses sources sans que les informations relatives aux données de l'une des sources ne soient révélées à aucune des autres sources. Les protocoles de CMS permettent à diverses parties de calculer conjointement des algorithmes dont les entrées restent des données privées.

Sources : Brundage et al. (2018[38]), *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*, https://arxiv.org/ftp/arxiv/papers/1802/1802.07228.pdf ; Dowlin (2016[61]), *CryptoNets: Applying Neural Networks to Encrypted Data with High Throughput and Accuracy*, https://www.microsoft.com/en-us/research/wp-content/uploads/2016/04/CryptonetsTechReport.pdf.

Autre solution : les modèles d'IA pourraient mettre à profit la technologie des chaînes de blocs, qui utilise elle aussi des outils cryptographiques pour sécuriser le stockage des données (Encadré 4.9). Les solutions alliant l'IA et la technologie des chaînes de blocs pourraient contribuer à accroître la disponibilité des données, tout en minimisant les risques d'atteinte à la vie privée et de sécurité liés au traitement de données non chiffrées.

Encadré 4.9. La technologie des chaînes de blocs permet une vérification d'identité respectueuse de la vie privée dans le cadre de l'IA

Kairos, une entreprise qui édite une solution de reconnaissance faciale, a intégré la technologie des chaînes de blocs dans son portefeuille. Elle l'allie à la biométrie faciale pour offrir aux utilisateurs une meilleure protection de leur vie privée. Un algorithme compare l'image d'une personne avec des repères faciaux jusqu'à obtenir une correspondance exacte. Celle-ci est convertie en une chaîne unique et aléatoire de nombres, ce qui permet de ne pas avoir à conserver l'image d'origine. Cette chaîne de blocs biométrique se fonde sur le principe que les entreprises ou les administrations n'ont pas besoin de savoir qui est l'utilisateur pour en vérifier l'identité.

Source : https://kairos.com/.

Concurrence

L'OCDE a étudié l'impact de la transformation numérique sur la concurrence, ainsi que les implications en termes d'action des pouvoirs publics (OCDE, 2019[62]). Cette sous-section expose certaines incidences potentielles propres à l'IA. Elle met en évidence les effets pro-concurrentiels largement reconnus de l'IA, qui facilite l'entrée de nouveaux acteurs. Elle souligne en outre que les politiques de la concurrence tendent à accorder davantage d'attention aux grands acteurs de l'IA, du fait de leur rôle d'opérateurs de plateformes en ligne et de détenteurs de volumes considérables de données. Elles s'intéressent peu en revanche à l'utilisation de l'IA à proprement parler.

Une question liée plus précisément à l'IA se pose : existe-t-il un effet de réseau fondé sur les données ? Si tel est le cas, l'utilité pour chaque utilisateur de recourir à certains types de plateformes augmente dès lors que d'autres l'utilisent également. Par exemple, en recourant à l'une de ces plateformes, ils contribuent à apprendre aux algorithmes qui la sous-tendent à mieux servir les utilisateurs (Heiner et Nguyen, 2018[9]). Par ailleurs, les données se caractérisent par des rendements d'échelle décroissants : l'amélioration des prévisions devient de plus en plus marginale à mesure que les données dépassent un certain seuil. Certains s'interrogent par conséquent sur le risque que l'IA pose des problèmes de concurrence à long terme (Bajari et al., 2018[63] ; OCDE, 2016[64] ; Varian, 2018[65]).

Des économies d'échelle pourraient être dégagées en termes de valeur conférée par des données supplémentaires. Si une entreprise affichant une qualité de données légèrement supérieure à celle de ses concurrents voit sa part de marché bondir, cela pourrait donner lieu à un effet de rétroaction positif. Plus de clients rime avec plus de données, ce qui renforce le cycle et permet à l'entreprise de consolider progressivement sa position sur le marché. Des économies d'échelle peuvent également être réalisées eu égard à l'expertise nécessaire pour bâtir des systèmes d'IA efficaces.

D'un autre côté, on s'inquiète du risque que les algorithmes favorisent la collusion en permettant aux entreprises de surveiller les conditions de marché, les prix et la réaction des concurrents aux variations de prix. Elles pourraient alors disposer d'outils nouveaux ou plus perfectionnés pour coordonner leurs stratégies, fixer les prix et mettre en place des ententes. D'aucuns spéculent en outre sur le risque que des algorithmes d'apprentissage profond plus sophistiqués puissent conduire à des résultats équivalents sans même que soient conclues des ententes formelles entre les concurrents, donc sans intervention humaine. Ce qui ne manquerait pas de poser des difficultés aux autorités de contrôle. Le droit de la concurrence exige en effet que, pour qu'une entente soit constatée et punie, il faut que des éléments probants attestent d'un accord ou d'un consentement entre les parties (OCDE, 2017[66]).

Propriété intellectuelle

Cette sous-section traite de certaines incidences potentielles de l'IA sur la propriété intellectuelle. Elle montre qu'il s'agit là d'un domaine en rapide évolution, qui commence à peine à faire l'objet de travaux analytiques fondés sur des données probantes. Les règles de propriété intellectuelle contribuent généralement à renforcer le degré et le rythme des découvertes, des inventions et de la diffusion des nouvelles technologies liées à l'IA. Elles sont comparables en ce sens aux règles applicables aux autres technologies protégées par des droits de propriété intellectuelle (DPI). Si elles doivent préserver les intérêts des inventeurs, des auteurs, des artistes et des propriétaires de marques, les politiques en matière de propriété intellectuelle doivent également tenir compte du potentiel de l'IA en tant que ressource à l'appui de nouvelles innovations.

La protection de l'IA par des DPI autres que ceux associés aux secrets commerciaux pourrait soulever de nouvelles questions quant aux incitations susceptibles d'encourager les innovateurs à divulguer leurs innovations liées à l'IA, notamment les algorithmes et leur entraînement. Le Bureau du Parlement européen a examiné, lors d'une conférence, trois types de brevets envisageables pour l'IA (OEB, 2018[67]). Le premier type a trait à l'« intelligence artificielle fondamentale » (« *core AI* » en anglais) ; il est souvent lié aux algorithmes qui, en tant que méthodes mathématiques, ne sont pas brevetables. Pour le deuxième type – qui porte sur les modèles entraînés et l'apprentissage automatique –, les demandes portant sur des variations et des fourchettes de paramètres pourraient poser problème. Troisièmement, des brevets pourraient être déposés sur l'IA en tant qu'outil dans un domaine d'application, défini d'après ses effets techniques. D'autres organisations internationales et des pays de l'OCDE étudient eux aussi les incidences de l'IA dans le domaine de la propriété intellectuelle[11].

La diffusion de l'IA soulève une autre problématique : des ajustements doivent-ils être apportés aux systèmes de protection de la propriété intellectuelle, dans un monde où les systèmes d'IA peuvent eux-mêmes créer des inventions (OCDE, 2018[68]) ? De fait, certains systèmes d'IA sont d'ores et déjà en mesure de produire des inventions brevetables, dans des domaines tels que la chimie, les produits pharmaceutiques et les biotechnologies. De nombreuses inventions portent par exemple sur la création de combinaisons originales de molécules pour former de nouveaux composés, ou sur l'identification de nouvelles propriétés de molécules existantes. Par exemple, KnIT, un outil d'apprentissage automatique mis au point par IBM, est parvenu à reconnaître des kinases – des enzymes catalysant le transfert de groupes de phosphates vers des substrats spécifiques. Ces kinases présentaient des propriétés particulières parmi un ensemble de kinases connues, qui ont été testées à titre expérimental. Les propriétés particulières de ces molécules ont été révélées par un logiciel, et des brevets ont été déposés pour protéger cette découverte. Ces aspects (ainsi que d'autres) liés à l'IA et aux DPI sont examinés par des organismes spécialisés de la zone OCDE, à l'instar de l'Office européen des brevets, du United States Patent and Trademark Office, ou de l'Organisation mondiale de la propriété intellectuelle. On pourrait également s'intéresser aux questions liées à la protection des droits d'auteur attachés aux données traitées par l'IA.

Petites et moyennes entreprises

Les politiques et programmes destinés à aider les petites et moyennes entreprises (PME) à opérer une transition vers l'IA revêtent un caractère prioritaire croissant. Il s'agit là d'un domaine en rapide évolution, qui commence à faire l'objet de travaux analytiques fondés sur des données probantes. Plusieurs pistes pourraient favoriser la mise en place d'écosystèmes numériques propices à l'adoption et l'utilisation de l'IA par les PME :

- Renforcer les compétences – un volet essentiel dans la mesure où les PME peinent à rivaliser pour attirer des spécialistes de l'IA par trop rares.
- Encourager la réalisation d'investissements ciblés dans des secteurs verticaux de choix. Les politiques visant à stimuler les investissements dans des applications de l'IA spécifiques dans le secteur français de l'agriculture, par exemple, pourraient bénéficier à l'ensemble des acteurs, y compris aux PME qui n'auraient pas les épaules pour investir seules (OCDE, 2018[14]).
- Aider les PME à accéder aux données, notamment via la création de plateformes d'échange de données.

- Faciliter l'accès des PME aux technologies liées à l'IA, notamment par le biais du transfert de technologies depuis les établissements publics de recherche, ainsi que l'accès à la puissance de calcul et aux plateformes infonuagiques (Allemagne, 2018[69]).
- Améliorer les mécanismes de financement afin d'aider les PME spécialisées dans l'IA à se développer, moyennant par exemple la création d'un fonds d'investissement public, ainsi que l'assouplissement et le relèvement des plafonds de financement des dispositifs en faveur de l'investissement dans les entreprises à forte intensité de savoir (RU, 2017[70]). La Commission européenne s'attache pour sa part à soutenir les PME européennes, notamment dans le cadre du projet AI4EU, une plateforme d'IA à la demande.

Cadre d'action à l'appui de l'innovation dans l'IA

L'OCDE analyse les changements qu'il conviendrait d'opérer dans les politiques d'innovation et d'autres politiques intéressant l'IA, dans le contexte de l'IA et d'autres transformations numériques (OCDE, 2018[49]). L'Organisation s'intéresse notamment aux moyens de renforcer l'adaptabilité, la réactivité et la souplesse des instruments d'action et des expérimentations connexes. Les pouvoirs publics peuvent recourir à l'expérimentation pour fournir des environnements contrôlés afin de tester les systèmes d'IA. De tels environnements pourraient intégrer des bacs à sable réglementaires, des centres d'innovation et des laboratoires des politiques. Les expérimentations de politiques peuvent se faire en « mode startup » : elles peuvent alors être déployées, évaluées et modifiées, transposées à une échelle supérieure ou inférieure, ou abandonnées rapidement, selon le cas.

Une autre solution pour parvenir à une prise de décision plus rapide et efficace consiste à recourir aux outils numériques pour concevoir les politiques (y compris les politiques d'innovation) et suivre la réalisation des objectifs y afférents. Par exemple, certains pays utilisent la modélisation multi-agents pour anticiper l'impact de diverses variantes des politiques sur les différents types d'entreprises.

Les pouvoirs publics peuvent encourager les acteurs de l'IA à mettre au point des mécanismes d'autoréglementation tels que des codes de conduite, des normes volontaires et des pratiques optimales. Ces dispositifs peuvent les guider tout au long du cycle de vie des systèmes d'IA, notamment pour le suivi, la communication, l'évaluation et le traitement des effets néfastes ou de l'utilisation abusive de ces systèmes.

Ils peuvent également instaurer et favoriser les mécanismes de surveillance des systèmes d'IA dans les secteurs public et privé, en tant que de besoin. Ces mécanismes peuvent intégrer des examens de conformité, des audits, des évaluations et des dispositifs de certification. Ils peuvent également être utiles lors de l'examen des besoins particuliers des PME et des contraintes auxquelles elles sont confrontées.

Se préparer à la transformation des emplois et renforcer les compétences

Emplois

L'IA devrait compléter le travail humain dans certaines tâches, le remplacer dans d'autres, et ouvrir la voie à de nouveaux types d'emplois

L'OCDE a réalisé une étude approfondie de l'impact de la transformation numérique sur l'emploi, ainsi que des incidences sur l'action des pouvoirs publics (OCDE, 2019[62]). Si

l'IA est un domaine en rapide évolution dans lequel les travaux analytiques fondés sur des données probantes ne font que commencer, on s'attend néanmoins à ce qu'elle modifie la nature du travail à mesure qu'elle se diffuse dans les différents secteurs. Elle est appelée à compléter le travail humain dans certaines tâches, le remplacer dans d'autres, et ouvrir la voie à de nouveaux types d'emplois. Cette section expose un certain nombre de mutations qui devraient intervenir sur les marchés du travail sous l'effet de l'IA, ainsi que les considérations intéressant l'action des pouvoirs publics, soulevées par la transition vers une économie de l'IA.

L'IA devrait stimuler la productivité

L'IA devrait stimuler la productivité de deux façons. D'une part, certaines activités menées à bien jusqu'à présent par des hommes vont être automatisées. D'autre part, avec l'avènement des machines autonomes, les systèmes fonctionneront et s'adapteront aux conditions moyennant un contrôle humain réduit, voire nul (OCDE, 2018[68] ; Autor et Salomons, 2018[71]). Des travaux de recherche portant sur 12 économies développées ont révélé que l'augmentation de la productivité du travail imputable à l'IA pourrait aller jusqu'à 40 % d'ici à 2035 par rapport aux niveaux de référence attendus (Purdy et Daugherty, 2016[72]). Les exemples sont légion. Le système Watson d'IBM assiste les conseillers des caisses du Crédit Mutuel à répondre aux questions des clients avec une rapidité augmentée de 60 %[12]. Lors de soldes en 2017, l'agent conversationnel d'Alibaba a traité plus de 95 % des demandes des clients. Les chargés de clientèle ont ainsi pu se charger des problématiques plus complexes ou personnelles (Zeng, 2018[73]). En théorie, l'augmentation de la productivité des travailleurs devrait se traduire par des hausses de salaires, puisque chaque employé produit davantage de valeur ajoutée.

La constitution d'équipes mêlant ressources humaines et IA permet de limiter les erreurs et d'ouvrir le champ des possibilités pour les travailleurs. Elles s'avèrent d'ailleurs plus productives que l'IA ou les travailleurs humains pris séparément (Daugherty et Wilson, 2018[74]). Ainsi, dans les usines de BMW, la constitution d'équipes mixtes de ce type a eu pour effet d'accroître la productivité manufacturière de 85 % par rapport à des équipes non intégrées. Les exemples ne se limitent pas aux activités industrielles : les robots de Walmart, par exemple, gèrent les stocks, de sorte que le personnel des magasins peut se concentrer sur l'assistance à la clientèle. Et lorsqu'un radiologue s'appuie sur des modèles d'IA pour réaliser des radiographies pulmonaires en cas de suspicion de tuberculose, la précision des diagnostics est de 100 % – soit un taux supérieur à celui atteint en cas de recours à l'IA ou d'intervention humaine seuls (Lakhani et Sundaram, 2017[75]).

L'IA peut également aider à améliorer et accélérer des tâches déjà automatisées. Elle permet ainsi aux entreprises de produire davantage, à moindre coût. Si la réduction des coûts est répercutée sur les prix aux entreprises ou aux individus, on peut s'attendre à une hausse de la demande de biens. Ce qui stimulerait la demande de main-d'œuvre à la fois au sein de l'entreprise concernée – dans des postes de production, par exemple – et dans les secteurs en aval, dans le cas de biens intermédiaires.

L'IA devrait modifier la physionomie des tâches automatisables – voire accélérer les mutations

L'automatisation n'est pas un phénomène nouveau, mais l'IA devrait modifier la physionomie des tâches susceptibles d'être automatisées, voire accélérer les mutations. Contrairement aux ordinateurs, les technologies liées à l'IA ne sont pas strictement préprogrammées et basées sur des règles. La diffusion des ordinateurs s'est traduite par une réduction des

emplois routiniers nécessitant un niveau de qualification intermédiaire. En revanche, les nouvelles applications faisant appel à l'IA sont de plus en plus à même d'exécuter des tâches relativement complexes impliquant de formuler des prévisions (voir chapitre 3). Ces tâches peuvent aller de la transcription à la traduction, en passant par la conduite de véhicules, l'établissement de diagnostics médicaux, ou le traitement des questions des clients (Graetz et Michaels, 2018[76] ; Michaels, Natraj et Van Reenen, 2014[77] ; Goos, Manning et Salomons, 2014[78])[13].

L'OCDE a réalisé des mesures préliminaires afin d'estimer la capacité des technologies à répondre aux questions de l'Évaluation des compétences des adultes (PIAAC) liées aux compétences à l'écrit et en calcul (Elliott, 2017[79]). Ces travaux ont montré qu'en 2017, les systèmes d'IA étaient à même de répondre aux questions portant sur les compétences à l'écrit à un niveau équivalent à celui de 89 % des adultes des pays de l'OCDE. En d'autres termes, seuls 11 % des adultes affichaient un niveau supérieur à celui que l'IA parvenait à reproduire en termes de maîtrise de la langue. L'étude tablait sur une augmentation de la pression économique en faveur de l'application des capacités informatiques pour certaines compétences à l'écrit et en calcul. Ce qui se traduirait par une baisse de la demande de travailleurs humains pour exécuter des tâches mobilisant des compétences à l'écrit de niveau faible à intermédiaire, à l'inverse des tendances observées récemment. L'étude soulignait en outre la difficulté de concevoir des politiques d'éducation pour les adultes disposant d'un niveau supérieur à celui des capacités informatiques actuelles. Elle proposait par conséquent de nouveaux outils et mesures d'incitation pour promouvoir les compétences des adultes ou associer des politiques de développement des compétences et d'autres interventions, dans des domaines tels que la protection sociale et le dialogue social (OCDE, 2018[14]).

Les incidences de l'IA sur les emplois dépendront de sa rapidité de diffusion dans différents secteurs

Les incidences de l'IA sur les emplois dépendront par ailleurs du rythme de développement et de diffusion des technologies liées à l'IA dans différents secteurs au cours des décennies à venir. Les véhicules autonomes devraient avoir des conséquences sur les emplois liés aux services de transport et de livraison. Des constructeurs de camions bien établis, à l'instar de Volvo et Daimler, par exemple, sont en compétition avec des startups comme Kodiak et Einride pour développer et tester des véhicules sans conducteur (Stewart, 2018[80]). Selon le Forum international des transports, les camions autonomes pourraient se multiplier sur les routes au cours des dix prochaines années. Quelque 50 à 70 % des 6.4 millions d'emplois de chauffeurs routiers professionnels aux États-Unis et en Europe pourraient disparaître d'ici à 2030 (FIT, 2017[81]). En parallèle, de nouveaux emplois seront toutefois créés pour fournir des services de support pour ce parc croissant de camions sans conducteur. De plus, les camions autonomes pourraient contribuer à réduire les frais d'exploitation liés au fret routier d'environ 30 %, notamment du fait de la diminution des coûts de main-d'œuvre. Cela pourrait entraîner la disparition d'entreprises de transport traditionnelles et, par ricochet, une baisse plus rapide encore des emplois de chauffeurs routiers.

Les technologies liées à l'IA devraient avoir des incidences sur les tâches exigeant traditionnellement un niveau de qualification plus élevé

Les technologies liées à l'IA exécutent des tâches de prévision généralement dévolues à des travailleurs très qualifiés – des juristes au personnel médical. Un robot avocat a par exemple aidé des automobilistes à contester des amendes de stationnement d'une valeur totale de 12 millions USD (Dormehl, 2018[82]). En 2016, les systèmes Watson d'IBM et

DeepMind Health ont obtenu de meilleurs résultats que des médecins humains dans le diagnostic de cancers rares (Frey et Osborne, 2017[83]). L'IA a également fait preuve d'une meilleure capacité à prévoir les variations des cours en bourse que les professionnels de la finance (Mims, 2010[84]).

L'IA peut compléter l'homme et créer de nouveaux types de travail

L'IA complète les travailleurs humains et devrait créer des possibilités d'emplois. Les domaines concernés sont ceux qui complètent les prévisions et exploitent les compétences comme la pensée critique, la créativité et l'empathie (EOP, 2016[85] ; OCDE, 2018[21]).

- **Scientifiques des données et experts en apprentissage automatique** : On a besoin de spécialistes pour créer et nettoyer les données, et programmer et développer les applications d'IA. Toutefois, bien que les données et l'apprentissage automatique donnent lieu à l'apparition de certaines tâches nouvelles, celles-ci ne devraient pas être pléthoriques pour les travailleurs.
- **Actions** : Certaines actions revêtent par nature davantage de valeur lorsqu'elles sont exécutées par des hommes (qu'il s'agisse d'athlètes de haut niveau, de professionnels de la petite enfance, ou de commerciaux) plutôt que par des machines. Beaucoup pensent que les humains vont se concentrer sur les emplois qui améliorent la qualité de vie, comme la garde d'enfants, le coaching sportif ou l'accompagnement des malades en fin de vie.
- **Jugement pour déterminer l'objet des prévisions** : Le facteur le plus important est probablement la capacité de jugement – à savoir le processus de détermination de l'intérêt d'une action particulière dans un environnement donné. Lorsque l'on recourt à l'IA pour établir des prévisions, un humain doit décider de ce que l'on va prévoir et de l'usage qui en sera fait. Énoncer des dilemmes, interpréter des situations ou extraire le sens d'un texte nécessitent entre autres des qualités de jugement et d'équité (OCDE, 2018[14]). En science, par exemple, l'IA peut compléter les personnes chargées du raisonnement conceptuel nécessaire pour bâtir les cadres de recherche et définir le contexte d'expériences spécifiques.
- **Jugement pour décider de l'usage à faire des prévisions** : Une décision ne peut être prise uniquement à partir d'une prévision. Par exemple, la décision somme toute banale d'emporter ou non un parapluie avant de sortir sera prise en tenant compte des prévisions sur les risques de précipitations, mais pas seulement : elle dépendra également, pour une large part, de préférences personnelles, selon que la personne déteste être mouillée ou ne souhaite pas s'embarrasser d'un parapluie. Cet exemple vaut pour de nombreuses décisions importantes. En cybersécurité, une prévision quant au caractère hostile d'une nouvelle requête doit être évaluée au regard du risque de refuser une requête amicale ou de laisser une requête hostile obtenir des informations non autorisées.

Les prévisions quant à l'impact net de l'IA sur la quantité de travail varient sensiblement

Au cours des cinq dernières années, des estimations divergentes ont été réalisées sur les conséquences globales de l'automatisation sur les pertes d'emplois (Winick, 2018[86] ; MGI, 2017[87] ; Frey et Osborne, 2017[83]). Par exemple, une étude de Frey et Osborne estimait que 47 % des emplois aux États-Unis étaient menacés de suppression au cours des 10 à 15 prochaines années. Adoptant une approche axée sur les tâches, le McKinsey Global

Institute a pour sa part estimé en 2017 qu'environ un tiers des activités dans 60 % des emplois étaient exposées à un risque d'automatisation. Néanmoins, l'automatisation des emplois identifiés était imputable non pas seulement au développement et au déploiement de l'IA, mais aussi à d'autres évolutions technologiques.

Sans compter qu'il est difficile de prévoir les futures créations d'emplois dans de nouveaux domaines. Selon une étude, l'IA devrait être à l'origine de deux millions de créations nettes d'emplois d'ici à 2025 (Gartner, 2017[88]). Des emplois devraient être créés à la fois dans le sillage de l'émergence de nouveaux métiers et par des canaux plus indirects. Par exemple, l'IA devrait contribuer à réduire les coûts de production des biens et des services et à en accroître la qualité. Ce qui devrait conduire à une hausse de la demande et, par ricochet, des emplois.

Les dernières estimations en date de l'OCDE tiennent compte de l'hétérogénéité des tâches dans des postes très spécifiques, en s'appuyant sur les données du Programme pour l'évaluation internationale des compétences des adultes (PIAAC). Si l'on se fonde sur les technologies existantes, 14 % des emplois sont fortement menacés d'automatisation dans les pays de l'OCDE ; et 32 % des travailleurs devraient voir leurs emplois sensiblement évoluer (Nedelkoska et Quintini, 2018[89]). Les travailleurs les plus jeunes et les plus âgés sont les groupes les plus exposés au risque d'automatisation. Une récente analyse de l'OCDE laisse entrevoir une baisse de l'emploi dans des métiers considérés comme exposés à un risque élevé d'automatisation dans 82 % des régions de 16 pays européens. Elle met en outre en lumière une augmentation plus marquée des emplois faiblement exposés dans 60 % des régions, augmentation qui compense les pertes d'emplois. Ces travaux tendent à confirmer que l'automatisation pourrait entraîner une mutation de la répartition des emplois, sans pour autant provoquer une baisse généralisée du niveau d'emploi (OCDE, 2018[90]).

L'IA est appelée à modifier la nature du travail

L'adoption de l'IA devrait modifier la nature du travail. L'IA pourrait contribuer à rendre le travail plus intéressant en favorisant l'automatisation des tâches répétitives et en ouvrant la voie à un travail plus flexible, voire à un meilleur équilibre entre vie professionnelle et vie privée. La créativité et l'ingéniosité humaines peuvent être conjuguées à l'augmentation des ressources en termes de puissance de calcul, de données et d'algorithmes pour créer de nouvelles tâches et activités faisant appel à leur tour à la créativité (Kasparov, 2018[91]).

Plus généralement, l'IA pourrait accélérer l'évolution du fonctionnement des marchés du travail en stimulant l'efficience. Aujourd'hui, les techniques axées sur l'IA, couplées aux données massives, promettent d'aider les entreprises à définir les rôles des travailleurs – et de contribuer à mettre en correspondance les travailleurs et les emplois. IBM, par exemple, utilise l'IA pour optimiser la formation de ses employés, leur recommandant des modules d'après leurs performances passées, leurs objectifs professionnels et les besoins en compétences de l'entreprise. De même, des sociétés comme KeenCorp et Vibe ont mis au point des techniques d'analyse de texte afin d'aider les entreprises à analyser les communications des employés pour faciliter l'établissement de mesures afférentes par exemple à leur état d'esprit, à leur productivité ou aux effets de réseau (Deloitte, 2017[92]). Grâce à ces informations, l'IA pourrait aider les entreprises à optimiser la productivité des travailleurs.

Les paramètres de changement organisationnel devront être définis

Il devient de plus en plus impératif de mettre en œuvre des normes sectorielles nouvelles ou révisées et des accords technologiques entre les directions et les employés afin de garantir un lieu de travail fiable, sûr et productif. Le Comité économique et social européen

(CESE) « préconise que les parties prenantes œuvrent ensemble en faveur de systèmes d'IA complémentaires et de leur mise en place conjointe sur le lieu de travail » (CESE, 2017[46]). Il importe en outre de favoriser une certaine flexibilité, tout en préservant l'autonomie des travailleurs et la qualité des emplois, y compris en termes de partage des bénéfices. La convention collective conclue récemment entre le syndicat de branche allemand *IG Metall* et les employeurs (*Gesamtmetall)* illustre la faisabilité économique de la mise en place de temps de travail variables. Elle montre en effet que, selon les besoins organisationnels et personnels dans le nouveau monde du travail, les employeurs et les syndicats peuvent parvenir à des accords sans que cela passe par une révision de la législation en matière de protection de l'emploi (Byhovskaya, 2018[93]).

L'utilisation de l'IA pour soutenir les fonctions des marchés du travail – avec des garanties – s'avère également prometteuse

L'IA contribue d'ores et déjà à accroître l'efficience de la mise en correspondance des offres et des demandes d'emploi, ainsi que de la formation. Elle peut aider à orienter les demandeurs d'emploi, y compris ceux dont l'emploi a été supprimé, vers les programmes de valorisation de la main-d'œuvre dont ils ont besoin en vue d'acquérir les qualifications nécessaires pour accéder aux métiers émergents ou en expansion. Dans de nombreux pays de l'OCDE, employeurs et services publics de l'emploi ont déjà recours à des plateformes électroniques pour pourvoir les emplois (OCDE, 2018[90]). À l'avenir, l'IA et d'autres technologies numériques permettront de mettre en place des approches innovantes et personnalisées des processus de recherche d'emploi et de recrutement, et de renforcer l'efficience de l'appariement des offres et des demandes d'emploi. C'est ainsi que la plateforme LinkedIn utilise l'IA pour aider les recruteurs à identifier les bons candidats et aiguiller les candidats vers les emplois les mieux adaptés à leur recherche. Elle s'appuie pour ce faire sur les données relatives au profil et à l'activité de ses 470 millions d'utilisateurs enregistrés (Wong, 2017[94]).

Les technologies liées à l'IA qui exploitent les données massives peuvent également éclairer les pouvoirs publics, les employeurs et les travailleurs sur les conditions des marchés du travail locaux. Ces informations les aident à identifier et prévoir les besoins en compétences, orienter les ressources de formation et guider les travailleurs vers les emplois. Plusieurs pays, tels la Finlande, la République tchèque et la Lettonie, mènent actuellement des projets en vue de développer les informations sur les marchés du travail (OCDE, 2018[90]).

Instaurer une gouvernance de l'utilisation des données des travailleurs

Si l'IA doit s'appuyer sur des ensembles de données volumineux pour être productive, des risques émergent dès lors que ces données concernent les travailleurs individuels, en particulier si les systèmes d'IA qui analysent ces données présentent un fonctionnement opaque. Or la planification des ressources humaines et de la productivité s'appuieront de plus en plus sur les données des salariés et les algorithmes. Les décideurs et les parties prenantes pourraient par conséquent s'intéresser à la manière dont la collecte et le traitement des données influent sur les perspectives et les conditions d'emploi. Les données peuvent être recueillies à partir des applications, des empreintes, des technologies prêt-à-porter et des capteurs en temps réel indiquant la localisation et le lieu de travail des employés. Dans le domaine des services à la clientèle, les logiciels basés sur l'AI analysent l'intonation des employés. Toutefois, selon les témoignages de certains salariés, ils ne tiennent pas compte des modèles de voix et il est difficile d'en contester les résultats (UNI, 2018[95]).

En revanche, les accords sur l'utilisation des données des employés et le droit à la déconnexion font leur apparition dans certains pays. L'opérateur de télécommunications Orange France Télécom et cinq centres syndicaux ont été parmi les premiers à prendre des engagements en faveur de la protection des données des employés. Les dispositions portent notamment sur la transparence quant à l'utilisation des données, la formation et l'introduction de nouveaux équipements. Pour parer aux lacunes réglementaires sur la gestion des données des travailleurs, des mesures pourraient être prises pour mettre en place des organes de gouvernance des données dans les entreprises et établir la responsabilité au regard de l'utilisation des données (personnelles), ainsi que des droits en matière de portabilité, d'explication et de suppression des données (UNI, 2018[95]).

Gérer la transition vers l'IA

Des politiques doivent être mises en place pour gérer la transition vers l'IA, notamment dans le domaine de la protection sociale

Les changements organisationnels n'intervenant pas au même rythme que le développement des technologies, des perturbations et des turbulences pourraient se manifester sur les marchés du travail (OCDE, 2018[14]). Les projections optimistes sur le long terme ne signifient pas pour autant que la transition vers une économie de plus en plus irriguée par l'IA se fera sans heurts : certains secteurs vont vraisemblablement croître, d'autres décliner. Des emplois sont menacés de disparaître, tandis que de nouveaux se créent. Par conséquent, l'enjeu phare de l'action publique sur les questions d'IA et d'emploi sera de gérer la transition, en agissant sur les politiques en matière de protection sociale, d'assurance maladie, d'imposition progressive du travail et du capital, et d'éducation. Les analyses de l'OCDE mettent par ailleurs en évidence la nécessité de prêter attention aux politiques de concurrence et autres politiques susceptibles d'influer sur les phénomènes de concentration, le pouvoir de marché et la répartition des revenus (OCDE, 2019[62]).

Compétences requises pour utiliser l'IA

La mutation des emplois s'accompagne d'une évolution des compétences nécessaires aux travailleurs

La mutation des emplois s'accompagne d'une évolution des compétences nécessaires aux travailleurs (OCDE, 2017[96] ; Acemoglu et Restrepo, 2018[97] ; Brynjolfsson et Mitchell, 2017[98]). Cette sous-section expose quelques-unes des répercussions possibles de l'IA sur les compétences, soulignant qu'il s'agit là d'un domaine en rapide évolution, qui commence à peine à faire l'objet de travaux analytiques fondés sur des données probantes. Des ajustements devront vraisemblablement être apportés aux politiques d'éducation afin d'étendre l'apprentissage tout au long de la vie, la formation et le développement des compétences. Comme pour les autres technologies, l'IA devrait créer une demande de compétences dans trois domaines. Premièrement, on aura besoin de **compétences spécialisées** pour programmer et développer les applications d'IA. Ces compétences sont requises dans différents domaines, depuis la recherche fondamentale, l'ingénierie et le développement d'applications en lien avec l'IA, jusqu'à la science des données et la pensée computationnelle. Deuxièmement, les individus devront disposer de **compétences génériques** pour exploiter l'IA, afin par exemple de pouvoir travailler dans des équipes mixtes IA/travailleurs humains dans les ateliers de fabrication ou pour les activités de contrôle qualité. Troisièmement, l'IA nécessitera des **compétences complémentaires**. Il s'agira notamment de mettre à profit des qualités humaines comme la pensée critique ; la créativité, l'innovation et l'entrepreneuriat ; ou encore l'empathie (EOP, 2016[85] ; OCDE, 2018[21]).

Des initiatives devront être mises en place pour développer et renforcer les compétences en IA nécessaires pour parer à la pénurie actuelle dans ce domaine

La pénurie de compétences en IA devrait s'accentuer et pourrait devenir plus prégnante encore avec l'essor de la demande de spécialistes dans des domaines comme l'apprentissage automatique. Les PME, les universités publiques et les centres de recherche sont d'ores et déjà en concurrence avec les entreprises dominantes pour attirer les talents. Des initiatives visant à développer et renforcer les compétences en IA voient peu à peu le jour dans les secteurs public, privé et universitaire. Par exemple, à Singapour, le gouvernement a mis en place un programme de recherche sur cinq ans sur la gouvernance de l'IA et l'utilisation des données à la Singapore Management University. Son Centre de gouvernance de l'IA et des données (Centre for AI & Data Governance) mène des recherches intéressant le secteur industriel, ciblées sur l'IA et l'industrie, la société et la commercialisation. Côté universitaire, le Massachusetts Institute of Technology (MIT) s'est engagé à consacrer 1 milliard USD à la création du Schwarzman College of Computing. L'objectif est de doter les étudiants et les chercheurs de tous horizons des compétences dont ils ont besoin pour utiliser l'informatique et l'IA en vue de faire progresser leur discipline, et inversement.

La pénurie de compétences en IA a également poussé certains pays à rationaliser les règles d'immigration pour les experts hautement qualifiés. Ainsi, le Royaume-Uni a doublé le nombre de visas *Tier 1* (*Exceptional Talent*), qu'il a porté à 2 000 par an, et rationalisé les règles permettant aux meilleurs étudiants et chercheurs de travailler dans le pays (RU, 2017[99]). Dans la même veine, le Canada a fixé à deux semaines les délais de traitement des demandes de permis de travail émanant de personnes hautement qualifiées et mis en place des dispenses de permis pour les missions de recherche de courte durée. Ces mesures s'inscrivent dans le cadre de la Stratégie en matière de compétences mondiales, adoptée par le Canada en 2017 pour attirer des travailleurs hautement qualifiés et des chercheurs étrangers (Canada, 2017[100]).

Compétences génériques requises pour exploiter l'IA

Tous les pays de l'OCDE évaluent les compétences et tentent d'anticiper les besoins immédiats et à moyen et long termes. La Finlande a ainsi proposé de mettre en place un Programme d'intelligence artificielle qui prévoit un « compte de compétences » ou un programme de bons d'échange pour bénéficier de formations continues afin de stimuler la demande d'éducation et de formation (Finlande, 2017[101]). Le Royaume-Uni promeut pour sa part une main-d'œuvre diversifiée formée à l'IA et investit environ 406 millions GBP (530 millions USD) dans le développement des compétences. Le pays concentre ses efforts sur la science, la technologie, l'ingénierie et les mathématiques, ainsi que sur la formation des enseignants en sciences informatiques (RU, 2017[99]).

Les professionnels doivent désormais disposer d'une double expertise (ce que certains appellent en anglais des « *bilinguals* »), à savoir être spécialisés dans une discipline comme l'économie, la biologie ou le droit, tout en disposant de compétences dans les techniques d'IA telles que l'apprentissage automatique. De la même veine, le MIT a annoncé en octobre 2018 l'évolution la plus importante de sa structure depuis 50 ans : créer une école informatique indépendante de la filière Ingénierie, avec des interconnexions avec tous les autres départements universitaires. On y enseignera aux étudiants l'art d'appliquer les techniques d'IA et d'apprentissage automatique aux défis qui se posent dans leurs propres disciplines. Il s'agit là d'un véritable tournant dans la façon dont le MIT enseigne les sciences informatiques. L'établissement consacre un milliard USD à la création de cette nouvelle école au sein du MIT (MIT, 2018[102]).

Compétences complémentaires

Les compétences non techniques font l'objet d'une attention accrue. Des travaux de recherche ont montré qu'elles peuvent avoir trait au jugement, à l'analyse et à la communication interpersonnelle (Agrawal, Gans et Goldfarb, 2018[103] ; Deming, 2017[104] ; Trajtenberg, 2018[105]). En 2021, l'OCDE intègrera au Programme international pour le suivi des acquis des élèves (PISA) un module destiné à tester les compétences en matière de créativité et de pensée critique. Les résultats aideront à établir une référence pour l'évaluation de la créativité dans les différents pays, afin d'étayer l'action des pouvoirs publics et des partenaires sociaux.

Mesure

La mise en œuvre d'une IA centrée sur l'humain et digne de confiance dépend du contexte. Toutefois, pour tenir leur engagement en ce sens, les décideurs devront définir des objectifs et des mesures permettant d'évaluer les performances des systèmes d'IA, dans des domaines tels que la fiabilité, l'efficience, la réalisation des objectifs sociétaux, l'équité et la robustesse.

Références

Abrams, M. et al. (2017), *Artificial Intelligence, Ethics and Enhanced Data Stewardship*, The Information Accountability Foundation, Plano, Texas. [17]

Acemoglu, D. et P. Restrepo (2018), *Artificial Intelligence, Automation and Work*, National Bureau of Economic Research, Cambridge, MA, http://dx.doi.org/10.3386/w24196. [97]

Agrawal, A., J. Gans et A. Goldfarb (2018), « Economic Policy for Artificial Intelligence », *National Bureau of Economic Research, Cambridge, MA,* 24690, http://dx.doi.org/10.3386/w24690. [50]

Agrawal, A., J. Gans et A. Goldfarb (2018), *Prediction Machines: The Simple Economics of Artificial Intelligence*, Harvard Business School Press. [103]

Agrawal, G. (dir. pub.) (2018), « Artificial intelligence and the modern productivity paradox: A clash of expectations and statistics », *National Bureau of Economic Research,* 24001, https://www.nber.org/papers/w24001. [51]

Allemagne (2018), « Key points for a federal government strategy on artificial intelligence », communiqué de presse, 18 juillet, BMWI, https://www.bmwi.de/Redaktion/EN/Pressemitteilungen/2018/20180718-key-points-for-federal-government-strategy-on-artificial-intelligence.html. [69]

Autor, D. et A. Salomons (2018), « Is automation labor-displacing? Productivity growth, employment, and the labor share », *document de travail*, n° 24871, National Bureau of Economic Research, Cambridge, MA, http://dx.doi.org/10.3386/w24871. [71]

Bajari, P. et al. (2018), « The impact of big data on firm performance: An empirical investigation », *document de travail*, n° 24334, National Bureau of Economic Research, Cambridge, MA, http://dx.doi.org/10.3386/w24334. [63]

Barocas, S. et A. Selbst (2016), « Big data's disparate impact », *California Law Review*, vol. 104, pp. 671-729, http://www.californialawreview.org/wp-content/uploads/2016/06/2Barocas-Selbst.pdf. [30]

Berk, R. et J. Hyatt (2015), « Machine learning forecasts of risk to inform sentencing decisions », *Federal Sentencing Reporter*, vol. 27/4, pp. 222-228, http://dx.doi.org/10.1525/fsr.2015.27.4.222. [24]

Borges, G. (2017), *Liability for Machine-Made Decisions: Gaps and Potential Solutions*, presentation at the "AI: Intelligent Machines, Smart Policies" conference, Paris, 26-27 October, http://www.oecd.org/going-digital/ai-intelligent-machines-smart-policies/conference-agenda/ai-intelligent-machines-smart-policies-borges.pdf. [43]

Brundage, M. et al. (2018), *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*, Future of Humanity Institute, University of Oxford, Centre for the Study of Existential Risk, University of Cambridge, Centre for a New American Security, Electronic Frontier Foundation and Open AI, https://arxiv.org/ftp/arxiv/papers/1802/1802.07228.pdf. [38]

Brynjolfsson, E. et T. Mitchell (2017), « What can machine learning do? Workforce implications », *Science*, vol. 358/6370, pp. 1530-1534, http://dx.doi.org/10.1126/science.aap8062. [98]

Burgess, M. (2016), « Holding AI to account: Will algorithms ever be free of bias if they are created by humans? », *WIRED,* 11 janvier, https://www.wired.co.uk/article/creating-transparent-ai-algorithms-machine-learning. [32]

Byhovskaya, A. (2018), *Overview of the National Strategies on Work 4.0: A Coherent Analysis of the Role of the Social Partners*, Comité économique et social européen, Bruxelles, https://www.eesc.europa.eu/sites/default/files/files/qe-02-18-923-en-n.pdf. [93]

Canada (2017), « Le gouvernement du Canada lance la Stratégie en matière de compétences mondiales », *communiqué de presse*, Immigration, Réfugiés et Citoyenneté Canada, 12 juin, https://www.canada.ca/fr/immigration-refugies-citoyennete/nouvelles/2017/06/le_gouvernement_ducanadalancelastrategieenmatieredecompetencesmo.html. [100]

Cellarius, M. (2017), *Artificial Intelligence and the Right to Informational Self-determination*, Forum de l'OCDE, OCDE, Paris, https://www.oecd-forum.org/users/75927-mathias-cellarius/posts/28608-artificial-intelligence-and-the-right-to-informational-self-determination. [10]

CESE (2017), *L'intelligence artificielle – Les retombées de l'intelligence artificielle pour le marché unique (numérique), la production, la consommation, l'emploi et la société*, Comité économique et social européen, Bruxelles, https://webapi2016.eesc.europa.eu/v1/documents/eesc-2016-05369-00-00-ac-tra-fr.docx/content. [46]

Chouldechova, A. (2016), « Fair prediction with disparate impact: A study of bias in recidivism prediction instruments », *arXiv, Cornell University*, vol. 07524, https://arxiv.org/abs/1610.07524. [25]

Citron, D. et F. Pasquale (2014), « The scored society: Due process for automated predictions », *Washington Law Review*, vol. 89, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2376209. [37]

Cockburn, I., R. Henderson et S. Stern (2018), « The impact of artificial intelligence on innovation », *document de travail*, n° 24449, National Bureau of Economic Research, Cambridge, MA, http://dx.doi.org/10.3386/w24449. [52]

Crawford, K. (2016), « Artificial intelligence's white guy problem », *New York Times,* 26 June, https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html?_r=0. [31]

Daugherty, P. et H. Wilson (2018), *Human Machine: Reimagining Work in the Age of AI*, Harvard Business Review Press, Cambridge, MA. [74]

Deloitte (2017), *HR Technology Disruptions for 2018: Productivity, Design and Intelligence Reign*, Deloitte, http://marketing.bersin.com/rs/976-LMP-699/images/HRTechDisruptions2018-Report-100517.pdf. [92]

Deming, D. (2017), « The growing importance of social skills in the labor market », *The Quarterly Journal of Economics*, vol. 132/4, pp. 1593-1640, http://dx.doi.org/10.1093/qje/qjx022. [104]

Dormehl, L. (2018), « Meet the British whiz kid who fights for justice with robo-lawyer sidekick », *Digital Trends* 3 mars, https://www.digitaltrends.com/cool-tech/robot-lawyer-free-acess-justice/. [82]

Doshi-Velez, F. et al. (2017), « Accountability of AI under the law: The role of explanation », *arXiv, Cornell University,* 21 novembre, https://arxiv.org/pdf/1711.01134.pdf. [29]

Dowlin, N. (2016), *CryptoNets: Applying Neural Networks to Encrypted Data with High Throughput and Accuracy*, Microsoft Research, https://www.microsoft.com/en-us/research/wp-content/uploads/2016/04/CryptonetsTechReport.pdf. [61]

Dressel, J. et H. Farid (2018), « The accuracy, fairness and limits of predicting recidivism », *Science Advances*, vol. 4/1, http://advances.sciencemag.org/content/4/1/eaao5580. [34]

Elliott, S. (2017), *Computers and the Future of Skill Demand*, La recherche et l'innovation dans l'enseignement, Éditions OCDE, Paris, https://dx.doi.org/10.1787/9789264284395-en. [79]

EOP (2016), *Artificial Intelligence, Automation and the Economy*, Executive Office of the President, Gouvernement des États-Unis, https://www.whitehouse.gov/sites/whitehouse.gov/files/images/EMBARGOED AI Economy Report.pdf. [85]

Finlande (2017), *Finland's Age of Artificial Intelligence - Turning Finland into a Leader in the Application of AI*, page web, Ministère finlandais de l'Emploi et de l'Économie, https://tem.fi/en/artificial-intelligence-programme. [101]

FIT (2017), « Driverless trucks: New report maps out global action on driver jobs and legal issues », *International Transport Forum*, https://www.itf-oecd.org/driverless-trucks-new-report-maps-out-global-action-driver-jobs-and-legal-issues. [81]

Flanagan, M., D. Howe et H. Nissenbaum (2008), « Embodying values in technology: Theory and practice », dans van den Hoven, J. et J. Weckert (dir. pub.), *Information Technology and Moral Philosophy*, Cambridge University Press, Cambridge, http://dx.doi.org/10.1017/cbo9780511498725.017. [16]

Freeman, R. (2017), *Evolution or Revolution? The Future of Regulation and Liability for AI*, presentation at the "AI: Intelligent Machines, Smart Policies" conference, Paris, 26-27 October, http://www.oecd.org/going-digital/ai-intelligent-machines-smart-policies/conference-agenda/ai-intelligent-machines-smart-policies-freeman.pdf. [41]

Frey, C. et M. Osborne (2017), « The future of employment: How susceptible are jobs to computerisation? », *Technological Forecasting and Social Change*, vol. 114, pp. 254-280, http://dx.doi.org/10.1016/j.techfore.2016.08.019. [83]

Gartner (2017), « Gartner says by 2020, artificial intelligence will create more jobs than it eliminates », Gartner, communiqué de presse, 13 décembre, https://www.gartner.com/en/newsroom/press-releases/2017-12-13-gartner-says-by-2020-artificial-intelligence-will-create-more-jobs-than-it-eliminates. [88]

Golson, J. (2016), « Google's self-driving cars rack up 3 million simulated miles every day », *The Verge,* 1 février, https://www.theverge.com/2016/2/1/10892020/google-self-driving-simulator-3-million-miles. [44]

Goodfellow, I., J. Shlens et C. Szegedy (2015), « Explaining and harnessing adversarial examples », *arXiv*, vol. 1412.6572, Cornell University, https://arxiv.org/pdf/1412.6572.pdf. [39]

Goos, M., A. Manning et A. Salomons (2014), « Explaining job polarization: Routine-biased technological change and offshoring », *American Economic Review*, vol. 104/8, pp. 2509-2526, http://dx.doi.org/10.1257/aer.104.8.2509. [78]

Graetz, G. et G. Michaels (2018), « Robots at work », *Review of Economics and Statistics*, vol. 100/5, pp. 753-768, http://dx.doi.org/10.1162/rest_a_00754. [76]

Harkous, H. (2018), « Polisis: Automated analysis and presentation of privacy policies using deep learning », *arXiv, Cornell University,* 29 juin, https://arxiv.org/pdf/1802.02561.pdf. [15]

HCDH (2011), *Principes directeurs des Nations Unies relatifs aux entreprises et aux droits de l'homme*, Haut-Commissariat des Nations Unies aux droits de l'homme, https://www.ohchr.org/documents/publications/guidingprinciplesbusinesshr_fr.pdf. [7]

Heiner, D. et C. Nguyen (2018), « Amplify Human Ingenuity with Intelligent Technology », *Shaping human-centered artificial intellilgence, A.Ideas Series*, Réseau du Forum, OCDE, Paris, https://www.oecd-forum.org/users/86008-david-heiner-and-carolyn-nguyen/posts/30653-shaping-human-centered-artificial-intelligence. [6]

Heiner, D. et C. Nguyen (2018), « Amplify Human Ingenuity with Intelligent Technology », *Shaping Human-Centered Artificial Intelligence, A.Ideas Series*, The Forum Network, OCDE, Paris, https://www.oecd-forum.org/users/86008-david-heiner-and-carolyn-nguyen/posts/30653-shaping-human-centered-artificial-intelligence. [9]

Helgason, S. (1997), *Vers un principe de responsabilité fondée sur la performance : éléments de discussion*, Service de la gestion publique, Éditions OCDE, Paris, http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=PUMA/PAC(97)8&docLanguage=Fr. [47]

Ingels, H. (2017), *Artificial Intelligence and EU Product Liability Law*, presentation at the "AI: Intelligent Machines, Smart Policies" conference, Paris, 26-27 October, http://www.oecd.org/going-digital/ai-intelligent-machines-smart-policies/conference-agenda/ai-intelligent-machines-smart-policies-ingels.pdf. [42]

Jain, S. (2017), « NanoNets : How to use deep learning when you have limited data, Part 2 : Building object detection models with almost no hardware », *Medium* 30 janvier, https://medium.com/nanonets/nanonets-how-to-use-deep-learning-when-you-have-limited-data-f68c0b512cab. [59]

Kasparov, G. (2018), *Deep Thinking: Where Machine Intelligence Ends and Human Creativity Begins*, Public Affairs, New York. [91]

Kendall, A. (23 mai 2017), « Deep learning is not good enough, we need Bayesian deep learning for safe AI », Alex Kendall Blog, https://alexgkendall.com/computer_vision/bayesian_deep_learning_for_safe_ai/. [60]

Knight, W. (2017), « The financial world wants to open AI's black boxes », *MIT Technology Review,* 13 April, https://www.technologyreview.com/s/604122/the-financial-world-wants-to-open-ais-black-boxes/. [33]

Kosack, S. et A. Fung (2014), « Does transparency improve governance? », *Annual Review of Political Science*, vol. 17, pp. 65-87, https://www.annualreviews.org/doi/pdf/10.1146/annurev-polisci-032210-144356. [27]

Kosinski, M., D. Stillwell et T. Graepel (2013), « Private traits and attributes are predictable from digital records of human behavior », *PNAS,* 11 mars, http://www.pnas.org/content/pnas/early/2013/03/06/1218772110.full.pdf. [2]

Kurakin, A., I. Goodfellow et S. Bengio (2017), « Adversarial examples in the physical world, », *arXiv, Cornell University* 02533, https://arxiv.org/abs/1607.02533. [40]

Lakhani, P. et B. Sundaram (2017), « Deep learning at chest radiography: Automated classification of pulmonary tuberculosis by using convolutional neural networks », *Radiology*, vol. 284/2, pp. 574-582, http://dx.doi.org/10.1148/radiol.2017162326. [75]

Matheson, R. (2018), *Artificial intelligence model "learns" from patient data to make cancer treatment less toxic*, MIT News, 9 août 2018, http://news.mit.edu/2018/artificial-intelligence-model-learns-patient-data-cancer-treatment-less-toxic-0810. [107]

MGI (2017), *Jobs Lost, Jobs Gained: Workforce Transitions in a Time of Automation*, McKinsey Global Institute, New York. [87]

Michaels, G., A. Natraj et J. Van Reenen (2014), « Has ICT polarized skill demand? Evidence from eleven countries over twenty-five years », *Review of Economics and Statistics*, vol. 96/1, pp. 60-77, http://dx.doi.org/10.1162/rest_a_00366. [77]

Mims, C. (2010), « AI that picks stocks better than the pros », *MIT Technology Review,* 10 June, https://www.technologyreview.com/s/419341/ai-that-picks-stocks-better-than-the-pros/. [84]

MIT (2018), « Cybersecurity's insidious new threat: Workforce stress », *MIT Technology Review,* 7 August, https://www.technologyreview.com/s/611727/cybersecuritys-insidious-new-threat-workforce-stress/. [102]

Mousave, S., M. Schukat et E. Howley (2018), « Deep reinforcement learning: An overview », *arXiv,* 1806.08894, https://arxiv.org/abs/1806.08894. [57]

Narayanan, A. (2018), « Tutorial: 21 fairness definitions and their politics », https://www.youtube.com/watch?v=jIXIuYdnyyk. [18]

Nedelkoska, L. et G. Quintini (2018), « Automation, skills use and training », *Documents de travail de l'OCDE sur les questions sociales, l'emploi et les migrations*, n° 202, Éditions OCDE, Paris, https://dx.doi.org/10.1787/2e2f4eea-en. [89]

Neppel, C. (2017), *AI: Intelligent Machines, Smart Policies*, exposé présenté à la conférence AI: Intelligent Machines, Smart Policies, Paris, les 26-27 octobre 2007, http://oe.cd/ai2017. [56]

NITI (2018), *National Strategy for Artificial Intelligence #AIforall*, NITI Aayog, juin 2018, http://niti.gov.in/writereaddata/files/document_publication/NationalStrategy-for-AI-Discussion-Paper.pdf. [5]

OCDE (2019), *Enhanced Access to Data and Sharing of Data (EASD)*, Groupe de travail sur la sécurité et la vie privée dans l'économie numérique , DSTI/CDEP/SPDE(2017)13/REV3. [55]

OCDE (2019), *Going Digital: Shaping Policies, Improving Lives*, Éditions OCDE, Paris, https://dx.doi.org/10.1787/9789264312012-en. [62]

OCDE (2019), *Recommandation du Conseil sur l'intelligence artificielle*, OCDE, Paris, https://legalinstruments.oecd.org/api/print?ids=648&lang=fr. [36]

OCDE (2019), *Scoping Principles to Foster Trust in and Adoption of AI – Proposal by the Expert Group on Artificial Intelligence at the OECD (AIGO)*, Éditions OCDE, Paris, http://oe.cd/ai. [35]

OCDE (2018), « AI: Intelligent machines, smart policies: Conference summary », *Documents de travail de l'OCDE sur l'économie numérique*, n° 270, Éditions OCDE, Paris, https://dx.doi.org/10.1787/f1a650d9-en. [14]

OCDE (2018), « Approaches to market openness in the digital age », dans *« Perspectives on innovation policies in the digital age », dans OECD Science, Technology and Innovation Outlook 2018 : Adapting to Technological and Societal Disruption*, Éditions OCDE, Paris, https://dx.doi.org/10.1787/sti_in_outlook-2018-8-en. [49]

OCDE (2018), *Job Creation and Local Economic Development 2018: Preparing for the Future of Work*, Éditions OCDE, Paris, https://dx.doi.org/10.1787/9789264305342-en. [90]

OCDE (2018), *La prochaine révolution de la production : Conséquences pour les pouvoirs publics et les entreprises*, Éditions OCDE, Paris, https://dx.doi.org/10.1787/9789264280793-fr. [68]

OCDE (2018), *Perspectives de l'économie numérique de l'OCDE 2017*, Éditions OCDE, Paris, https://dx.doi.org/10.1787/9789264282483-fr. [21]

OCDE (2018), *Science, technologie et innovation : Perspectives de l'OCDE 2018 (version abrégée) : S'adapter aux bouleversements technologiques et sociétaux*, Éditions OCDE, Paris, https://dx.doi.org/10.1787/sti_in_outlook-2018-fr. [53]

OCDE (2017), *Algorithms and Collusion: Competition Policy in the Digital Age*, Éditions OCDE, Paris, https://www.oecd.org/fr/concurrence/algorithms-collusion-competition-policy-in-the-digital-age.htm. [66]

OCDE (2017), *Getting Skills Right: Skills for Jobs Indicators*, Getting Skills Right, Éditions OCDE, Paris, https://dx.doi.org/10.1787/9789264277878-en. [96]

OCDE (2016), *Données massives : Adapter la politique de la concurrence à l'ère du numérique (Synthèse)*, Comité de la concurrence, https://one.oecd.org/document/DAF/COMP/M(2016)2/ANN4/FINAL/fr/pdf. [64]

OCDE (2013), *Recommandation du Conseil concernant les Lignes directrices régissant la protection de la vie privée et les flux transfrontières de données de caractère personnel*, OCDE, Paris, https://www.oecd.org/fr/internet/ieconomie/lignesdirectricesregissantlaprotectiondelavieprivee etlesfluxtransfrontieresdedonneesdecaracterepersonnel.htm. [13]

OCDE (2011), *Les principes directeurs de l'OCDE à l'intention des entreprises multinationales*, Editions OCDE, Paris, https://doi.org/10.1787/9789264115439-fr. [8]

OEB (2018), *Patenting Artificial Intelligence - Conference summary*, Office européen des brevets, Munich, 30 mai, http://documents.epo.org/projects/babylon/acad.nsf/0/D9F20464038C0753C125829E0031B814/$FILE/summary_conference_artificial_intelligence_en.pdf. [67]

Office of the Victorian Information Commissioner (2018), « Artificial intelligence and privacy », *Issues Paper*, juin, Office of the Victorian Information Commissioner, https://ovic.vic.gov.au/wp-content/uploads/2018/08/AI-Issues-Paper-V1.1.pdf. [12]

O'Neil, C. (2016), *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, Broadway Books. [26]

OpenAI (16 mai 2018), « AI and compute », OpenAI blog, San Francisco, https://blog.openai.com/ai-and-compute/. [54]

Pan, S. et Q. Yang (2010), « A survey on transfer learning », *IEEE Transactions on Knowledge and Data Engineering*, vol. 22/10, pp. 1345-1359. [58]

Patki, N., R. Wedge et K. Veeramachaneni (2016), « The Synthetic Data Vault », *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, http://dx.doi.org/10.1109/dsaa.2016.49. [106]

Privacy International et Article 19 (2018), *Privacy and Freedom of Expression in the Age of Artificial Intelligence*, https://www.article19.org/wp-content/uploads/2018/04/Privacy-and-Freedom-of-Expression-In-the-Age-of-Artificial-Intelligence-1.pdf. [11]

Purdy, M. et P. Daugherty (2016), « Artificial intelligence poised to double annual economic growth rate in 12 developed economies and boost labor productivity by up to 40 percent by 2035, according to new research by Accenture », Accenture, Press Release, 28 septembre, http://www.accenture.com/futureofAI. [72]

RU (2017), *UK Digital Strategy*, Gouvernement du Royaume-Uni, https://www.gov.uk/government/publications/uk-digital-strategy. [70]

RU (2017), *UK Industrial Strategy: A Leading Destination to Invest and Grow*, Royaume-Uni et Irlande du Nord, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/668161/the_labour_market_story-_skills_use_at_work.pdf. [99]

Selbst, A. (2017), « Disparate impact in big data policing », *Georgia Law Review*, vol. 52/109, http://dx.doi.org/10.2139/ssrn.2819182. [22]

Simonite, T. (2018), « Probing the dark side of Google's ad-targeting system », *MIT Technology Review,* 6 July, https://www.technologyreview.com/s/539021/probing-the-dark-side-of-googles-ad-targeting-system/. [20]

Slusallek, P. (2018), *Artificial Intelligence and Digital Reality: Do We Need a CERN for AI?*, The Forum Network, OCDE, Paris, https://www.oecd-forum.org/channels/722-digitalisation/posts/28452-artificial-intelligence-and-digital-reality-do-we-need-a-cern-for-ai. [45]

Smith, M. et S. Neupane (2018), *Artificial Intelligence and Human Development: Toward a Research Agenda*, Centre de recherches pour le développement international, Ottawa, https://idl-bnc-idrc.dspacedirect.org/handle/10625/56949. [4]

Stewart, J. (2018), « As Uber gives up on self-driving trucks, another startup jumps in », *WIRED,* 8 July, https://www.wired.com/story/kodiak-self-driving-semi-trucks/. [80]

Talbot, D. et al. (2017), « Charting a roadmap to ensure AI benefits all », *Medium,* 30 November, https://medium.com/berkman-klein-center/charting-a-roadmap-to-ensure-artificial-intelligence-ai-benefits-all-e322f23f8b59. [3]

Trajtenberg, M. (2018), « AI as the next GPT: A political-economy perspective », *National Bureau of Economic Research*, vol. 24245, Cambridge, MA, http://dx.doi.org/10.3386/w24245. [105]

UNI (2018), *10 Principles for Workers' Data Rights and Privacy*, UNI Global Union, http://www.thefutureworldofwork.org/docs/10-principles-for-workers-data-rights-and-privacy/. [95]

Varian, H. (2018), « Artificial intelligence, economics and industrial organization », n° 24839, National Bureau of Economic Research, Cambridge, MA, http://dx.doi.org/10.3386/w24839. [65]

Wachter, S., B. Mittelstadt et L. Floridi (2017), « Transparent, explainable and accountable AI for robotics », *Science Robotics,* 31 May, http://robotics.sciencemag.org/content/2/6/eaan6080. [48]

Wachter, S., B. Mittelstadt et C. Russell (2017), « Counterfactual explanations without opening the black box: Automated decisions and the GDPR », *arXiv, Cornell University,* 00399, https://arxiv.org/pdf/1711.00399.pdf. [28]

Weinberger, D. (2018), « Optimization over explanation - Maximizing the benefits of machine learning without sacrificing its intelligence », *Medium,* 28 January, https://medium.com/@dweinberger/optimization-over-explanation-maximizing-the-benefits-we-want-from-machine-learning-without-347ccd9f3a66. [1]

Weinberger, D. (2018), *Playing with AI Fairness*, Google PAIR, 17 septembre, https://pair-code.github.io/what-if-tool/ai-fairness.html. [23]

Winick, E. (2018), « Every study we could find on what automation will do to jobs, in one chart », *MIT Technology Review,* 25 January, https://www.technologyreview.com/s/610005/every-study-we-could-find-on-what-automation-will-do-to-jobs-in-one-chart/. [86]

Wong, Q. (2017), « "At LinkedIn, artificial intelligence is like "oxygen"" », *Mercury News,* 1 June, http://www.mercurynews.com/2017/01/06/at-linkedin-artificial-intelligence-is-like-oxygen. [94]

Yona, G. (2017), « A gentle introduction to the discussion on algorithmic fairness », *Toward Data Science,* 5 October, https://towardsdatascience.com/a-gentle-introduction-to-the-discussion-on-algorithmic-fairness-740bbb469b6. [19]

Zeng, M. (2018), *Alibaba and the Future of Business*, Harvard Business Review, septembre-octobre 2018, https://hbr.org/2018/09/alibaba-and-the-future-of-business. [73]

Notes

[1] Pour en savoir plus, consulter la page https://www.microsoft.com/en-us/ai/ai-for-good.

[2] Voir : https://deepmind.com/applied/deepmind-ethics-society/.

[3] Voir : https://www.blog.google/technology/ai/ai-principles/.

[4] Voir : https://ai.google/static/documents/perspectives-on-issues-in-ai-governance.pdf.

[5] Voir l'Organisation internationale du Travail, les Principes directeurs de l'OCDE à l'intention des entreprises multinationales, ou encore les Principes directeurs des Nations Unies relatifs aux entreprises et aux droits de l'homme.

[6] Pour en savoir plus sur le sujet, voir les pages https://www.dudumimran.com/2018/05/speaking-about-ai-and-cyber-security-at-the-oecd-forum-2018.html et https://maliciousaireport.com/.

[7] Pierre Chalançon, Président du Groupe de réflexion du BIAC sur la protection des consommateurs et Vice-Président des affaires réglementaires, Vorwerk & Co KG, Représentation auprès de l'Union européenne – *Science-Fiction is not a Sound Basis for Legislation.*

[8] Cette technique est notamment utilisée pour entraîner les véhicules autonomes à effectuer des manœuvres complexes, entraîner le programme AlphaGo, ou encore traiter des patients atteints de cancers, en déterminant la dose et la fréquence d'administration minimales efficaces sur la réduction des tumeurs cérébrales (Matheson, 2018[107]).

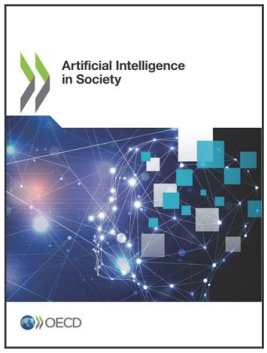
[9] Une récente étude révèle que dans de nombreux cas, les données synthétiques peuvent remplacer utilement les données réelles et aident les chercheurs à s'affranchir des contraintes liées au respect de la vie privée (Patki, Wedge et Veeramachaneni, 2016[106]). Les auteurs montrent en effet que dans 70 % des cas, les résultats générés à l'aide des données synthétiques ne sont pas foncièrement différents de ceux obtenus avec des données réelles.

[10] Des solutions alliant des mécanismes tels que le chiffrement homomorphe complet et des réseaux neuronaux ont été testées avec succès et utilisées à cet effet (Dowlin, 2016[61]).

[11] Voir https://www.wipo.int/about-ip/fr/artificial_intelligence/index.html et https://www.uspto.gov/about-us/events/artificial-intelligence-intellectual-property-policy-considerations.

[12] Voir https://www.ibm.com/watson/stories/creditmutuel/.

[13] À titre d'exemple, Alibaba n'emploie plus de travailleurs temporaires pour gérer les demandes des clients pendant les périodes de forte activité ou les promotions spéciales. Le jour où Alibaba a réalisé son pic de ventes en 2017, l'agent conversationnel a traité plus de 95 % des questions des clients et répondu à quelque 3.5 millions de consommateurs (Zeng, 2018[73]). À mesure que les agents conversationnels prennent en charge des fonctions de service à la clientèle, le rôle des conseillers humains évolue vers le traitement de questions plus complexes ou personnelles.



Extrait de :
Artificial Intelligence in Society

Accéder à cette publication :
https://doi.org/10.1787/eedfee77-en

Merci de citer ce chapitre comme suit :

OCDE (2019), « Considérations de politique publique », dans *Artificial Intelligence in Society*, Éditions OCDE, Paris.

DOI: https://doi.org/10.1787/93d862d5-fr