Strasbourg, 15 octobre 2018

T-PD(2018)09Rev

COMITÉ CONSULTATIF DE LA CONVENTION POUR LA PROTECTION DES PERSONNES À L'ÉGARD DU TRAITEMENT AUTOMATISÉ DES DONNÉES À CARACTÈRE PERSONNEL

(Convention 108)

RAPPORT SUR L'INTELLIGENCE ARTIFICIELLE

Intelligence artificielle et protection des données : enjeux et solutions possibles

Direction générale Droits de l'homme et État de droit

Rapport établi par Alessandro Mantelero, professeur associé de droit privé à l'École polytechnique de Turin, Département Management et Ingénierie de production. Les vues exprimées dans ce document relèvent de la responsabilité de l'auteur.

***

Table des matières

Partie I – État des lieux

I.1 Introduction

Définir le périmètre d'étude du présent rapport n'est pas chose facile, étant donné que les frontières de la protection des données et de l'intelligence artificielle (ci-après IA[1]) sont assez incertaines. D'une part, les technologies à usage intensif de données (comme l'IA) constituent un véritable défi pour l'application de certains principes traditionnels de protection des données, en les rendant plus flous, moins clairs ou plus difficile à appliquer [Conseil de l'Europe, 2017 ; Hildebrandt, 2016 ; Barocas et Nissenbaum, 2015 ; Citron et Pasquale, 2014 ; Mantelero, 2014 ; Rubinstein, 2013 ; Boyd et Crawford, 2012 ; Tene et Polonetsky, 2012]. De l'autre, l'IA est un vaste domaine englobant une diversité d'approches qui cherchent à reproduire les capacités cognitives de l'être humain [Villani, 2018, p. 9].

Protection des données et intelligence artificielle sont nécessairement corrélées. Hormis les scénarios relevant de la science-fiction, l'évolution rapide des applications de l'IA ces dernières années est l'aboutissement d'un processus progressif de « mise en données » (*datafication*) [Mayer-Schönberger et Cukier, 2013, 78 ; Lycett, 2013], de sorte que les données personnelles sont de plus en plus devenues à la fois la source et la cible des applications de l'IA (assistants personnels, appareils domestiques intelligents, etc.).

Dans ce contexte, différentes approches se font jour concernant le développement, l'utilisation et la réglementation de l'intelligence artificielle. En réalité, l'IA est en grande partie non réglementée et n'est généralement pas fondée sur les droits fondamentaux ; de fait, son fonctionnement même repose essentiellement sur le traitement des données.

En ce qui concerne le traitement des données, le cadre international offre différents moyens d'assurer la sauvegarde des droits fondamentaux et notamment du droit à la protection des données à caractère personnel. L'engagement très actif de l'Europe dans le domaine de la protection des données pourrait conduire cette région à assumer un rôle prépondérant face aux enjeux soulevés par l'encadrement du développement de l'IA.

L'adoption d'une perspective axée sur les droits fondamentaux pourrait aussi atténuer le clash annoncé entre un développement axé sur le marché et la technologie et une approche plus inclusive. Au regard de la Convention 108 et, plus généralement, de l'attitude du Conseil de l'Europe en ce qui concerne les droits fondamentaux, une solution aux tensions existantes pourrait être fournie par le cadre réglementaire et par la jurisprudence de la Cour européenne des droits de l'homme.

En termes de politique, la nature structurante des droits fondamentaux a conduit les Parties à la Convention 108 à favoriser le développement de technologies qui s'appuient sur ces droits et ne sont pas simplement dictées par les forces du marché ou par les entreprises de haute technologie. Historiquement, en outre, la protection des données en Europe plonge ses racines dans les appels lancés aux décideurs pour qu'ils prennent en considération les effets potentiellement préjudiciables des technologies de traitement de données.

Cette approche fondée sur les droits a inévitablement des conséquences sur le développement de l'IA, qui devrait se faire en accord avec les valeurs énoncées dans la Convention 108 et la réglementation du Conseil de l'Europe. Les Parties à la Convention devraient dès lors encourager activement les développeurs en intelligence artificielle à adopter une démarche de conception des produits et services **centrée sur les valeurs,** sans rester dans le vague ou afficher une vision excessivement optimiste de l'IA.

---

[1] À l'origine du terme « intelligence artificielle » se trouve l'Américain John McCarthy, un scientifique informatique considéré comme le père de l'IA. Voir M. L. Minsky, N. Rochester et C. E. Shannon, *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*, 31 août 1955, http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html, consulté le 19 juin 2018. Une définition de l'IA est disponible ici : https://www.coe.int/fr/web/human-rights-rule-of-law/artificial-intelligence/glossary.

En même temps, les gouvernements devraient être les premiers à utiliser l'intelligence artificielle d'une manière axée sur la préservation et la protection des données et des droits fondamentaux, empêchant ainsi le développement de systèmes ou techniques d'intelligence artificielle qui limitent les libertés et les droits individuels et collectifs.

C'est pourquoi il est important d'étendre le leadership européen en matière de réglementation de la protection des données à une régulation de l'intelligence artificielle orientée par les valeurs [Villani, 2018, p. 12], fondée sur les trois principes suivants :

- Approche fondée sur les valeurs (englobant les valeurs sociales et éthiques)
- Évaluation et gestion des risques
- Participation

La perspective du Conseil de l'Europe dépasse les frontières de l'UE et recouvre une grande diversité de cultures juridiques et approches réglementaires. Malgré cela, le cadre juridique du Conseil de l'Europe, tout comme la Convention 108 elle-même, fournit un contexte uniforme en termes de valeurs communes.

Le Conseil de l'Europe pourrait être l'un des meilleurs forums pour combiner souci des droits fondamentaux et régulation souple des technologies, en adoptant **une approche fondée sur des principes**. Les principes peuvent avoir une portée plus large et être interprétés spécifiquement pour relever les défis d'un monde en mutation ; en revanche, des dispositions législatives détaillées n'apparaissent pas être à même de réagir suffisamment rapidement aux changements socio-économiques et technologiques.

De plus, une réglementation fondée sur des principes permet de prendre en compte les spécificités locales. Ceci est encore plus pertinent s'agissant des applications de l'intelligence artificielle, qui peuvent avoir une incidence sur les valeurs contextuelles – juridiques, éthiques et sociales [IEEE, 2016].

Bien sûr, la protection des données en soi ne couvre pas tous ces aspects, qui requièrent une approche plus large englobant les droits de l'homme[2] et les questions de société[3] [CEDP, 2018 ; Mantelero, 2018 ; Conseil de l'Europe, 2017]. Cependant, la protection des données peut renforcer et compléter la réponse à ces questions.

L'accent mis sur l'individu, la prise de conscience des conséquences sociales de l'usage des données et le lien avec les droits de la personnalité peuvent élargir l'approche du responsable du traitement : au-delà de la protection des données, il s'agit de garantir les droits fondamentaux et les intérêts collectifs. La protection des données contribue ainsi, à titre complémentaire, à révéler la façon dont sont utilisées les données et les finalités du traitement, éléments essentiels pour mieux appréhender les conséquences potentielles sur divers droits et libertés.

Enfin, l'intelligence artificielle soulève de nombreuses questions propres à chaque secteur concernant ses différents domaines d'application (monde du travail, administration de la justice, répression du crime, relations contractuelles, etc.) et les conséquences de son usage (durabilité, effets sur l'environnement, impact politique, etc.), qu'il convient d'aborder séparément. Étant donné l'approche de la Convention 108, ces aspects ne sont pas examinés dans le présent rapport, qui s'intéresse au dénominateur commun à toutes ces applications : le traitement des données. En termes de retombées potentielles, cette analyse pourrait par conséquent apporter une contribution au débat sur les questions concernant l'IA en général et ses applications concrètes.

[2] Voir Convention modernisée pour la protection des personnes à l'égard du traitement automatisé des données à caractère personnel, préambule et art. 1.

[3] Voir Comité consultatif de la Convention pour la protection des personnes à l'égard du traitement automatisé des données à caractère personnel, *Lignes directrices sur la protection des personnes à l'égard du traitement des données à caractère personnel à l'ère des mégadonnées* (ci-après « Lignes directrices ») du 23 janvier 2017.

I.2 Le développement de l'intelligence artificielle

Au fil des ans, de nombreux rapports et travaux scientifiques ont été publiés sur l'intelligence artificielle et son évolution. Il est inutile ici de retracer l'inégal intérêt scientifique et social manifesté par la société pour cette technologie, des toutes premières études [McCulloch et Pitts, 1943 ; Turing, 1950] aux apports les plus récents. Point n'est besoin non plus de décrire la variété grandissante des applications de l'intelligence artificielle et les résultats obtenus.

Une perspective historique est néanmoins importante pour bien comprendre le présent et l'avenir à court terme de l'intelligence artificielle. Deux questions se posent à cet égard : pourquoi l'intelligence artificielle est-elle au cœur des débats ces dernières années ? Et à quelles formes d'intelligence artificielle peut-on raisonnablement s'attendre dans les années à venir ? Les réponses à ces questions sont essentielles pour s'attaquer à la régulation de l'IA. En effet, il importe de replacer le développement de l'IA dans son contexte et d'éviter les discours commerciaux et médiatiques confus à ce sujet.

Pour commencer, l'IA n'est pas un simple effet de mode. Comme cela fut le cas dans le passé avec l'informatique dématérialisée (*cloud computing*), le Big Data et l'internet des objets (IO), certains vendeurs ont clairement tendance à magnifier les possibilités de l'intelligence artificielle et ce terme est devenu un mot-valise dans des contextes qui n'utilisent pas strictement cette technologie. Derrière cette attention portée à l'IA il y a toutefois un fond de vérité : l'environnement technique particulier qui permet aujourd'hui d'obtenir des résultats inimaginables par le passé.

Ces dix dernières années, la disponibilité grandissante de bande passante pour le transfert et le stockage de données et l'accès aux ressources informatiques – à travers le nouveau paradigme du *cloud computing* – et le processus de mise en données (*datafication*) d'une grande partie de notre vie et de notre environnement ont créé un contexte entièrement nouveau. Cela a débouché sur des avancées majeures de l'intelligence artificielle, en permettant à de nouvelles formes de traitement des données d'extraire davantage d'information et de produire de nouveaux savoirs.

L'analytique des mégadonnées (*Big Data analytics*) et l'apprentissage automatique (*Machine Learning*)[4] constituent les produits les plus récents de ce processus de développement [Autorité norvégienne de protection des données, 2018, p. 5]. Les applications concrètes de ces technologies permettent d'envisager le type d'intelligence artificielle auquel on peut raisonnablement s'attendre dans les prochaines années et montrent que l'on est encore très loin de ce qu'il est convenu d'appeler une « intelligence artificielle généralisée » [Bostrom, 2016 ; Bureau exécutif du président des États-Unis et Conseil national de la science et de la technologie – Comité sur la technologie, 2016, p. 7 ; Autorité norvégienne de protection des données, 2018 ; Cummings et al., 2018].

S'il est vrai que « les algorithmes et l'intelligence artificielle en sont venus à constituer de nouvelles mythologies de notre temps » [CNIL, 2017], le **présent rapport s'intéresse néanmoins aux applications existantes de l'IA et à celles qui devraient être mises en œuvre dans un proche avenir**, laissant de côté des questions épineuses concernant une intelligence artificielle à l'image de l'homme, en termes de responsabilité des machines et de risques pour l'humanité [Bostrom, 2016 ; Kurzweil, 2016]. La Convention 108, dans son texte initial comme dans la version modernisée, fait référence au « traitement automatisé » des données et non pas à leur traitement autonome, soulignant implicitement que l'autonomie est une qualité fondamentale de l'être humain [Commission européenne, 2018].

Il ressort clairement de ce bref état des lieux que **l'intelligence artificielle repose inévitablement sur le traitement des données**. Les algorithmes d'IA ont nécessairement un impact sur l'usage des

[4] La différence entre ces deux technologies peut se résumer comme suit : schémas récurrents / liens. *« C'est là où l'IA peut faire une différence. Alors que les méthodes analytiques classiques nécessitent une programmation pour trouver des correspondances et des liens, l'IA apprend à partir de toutes les données qu'elle voit. Les systèmes informatiques peuvent dès lors constamment intégrer les nouvelles données et ajuster leurs analyses sans intervention humaine. L'IA contribue ainsi à lever les obstacles techniques auxquels se heurtent les méthodes traditionnelles dans l'analyse des mégadonnées »* [Autorité norvégienne de protection des données, 2018, p. 5].

données à caractère personnel. Cela conduit à s'interroger sur l'adéquation de la réglementation actuelle en matière de protection des données face aux défis posés par ces nouveaux paradigmes.

I.3 La perspective adoptée

Les principales menaces liées à l'intelligence artificielle tiennent à des valeurs contestées adoptées par les développeurs en IA et par les usagers, ce dernier groupe englobant les consommateurs et les décideurs qui s'appuient sur l'IA pour étayer leurs choix. Une tendance de fond se dessine vers une société technocratique, axée sur le marché, qui encourage la monétisation des données, des formes de contrôle social et des systèmes d'aide à la décision peu coûteux et rapides [Spiekermann, 2016, p. 152-153], à grande (les villes intelligentes par exemple) ou petite échelle (comme la médecine de précision).

Le renforcement de cette tendance met en question et sape progressivement l'autodétermination de l'individu, des modèles axés sur le respect de la vie privée et des processus où chaque décision est pesée et réfléchie. La boulimie de données, la complexité du traitement et une logique centrée à l'extrême sur les données sont susceptibles de compromettre leur usage démocratique, en supplantant l'individu et la collectivité elle-même, de même que les libertés et l'autodétermination, par une sorte de dictature des données [O'Neil, 2017] imposée par des scientifiques insensibles aux enjeux sociétaux.

Pour éviter que les effets négatifs de l'intelligence artificielle ne l'emportent sur ses avantages et bénéfices [UIT, 2017 ; Information Commissioner's Office (ICO), 2017, p. 15-18 ; Forum économique mondial, 2018], il est nécessaire de souligner **le rôle central de l'être humain dans le développement de la technologie (et de l'IA)**. Cela implique de réaffirmer la prédominance des droits fondamentaux dans ce domaine.

En effet, le droit à la protection des données à caractère personnel peut devenir un tremplin pour aller vers une société de données différente, dans laquelle le développement de l'IA ne serait pas motivé par des intérêts purement économiques ou par une efficience algorithmique déshumanisante.

Un vaste débat s'impose pour renforcer ce paradigme construit autour des droits fondamentaux. Nous devons porter un regard critique sur cette tendance à l'extrême « datafication » de tous les aspects de notre quotidien et affirmer l'importance des droits individuels et collectifs. **Les gouvernements et les citoyens doivent prendre conscience des risques liés à cette mise en données du monde** et des implications potentiellement néfastes des solutions axées sur les données [Rouvroy, 2016].

Comme pour le développement industriel de produits par le passé, **la conscience du risque n'est pas une barrière, mais plutôt un facteur favorisant l'innovation**. Il faut innover de façon responsable, la sauvegarde des droits fondamentaux étant un objectif primordial.

Ceci requiert nécessairement le développement de procédures d'évaluation, l'adoption de modèles participatifs et des autorités de contrôle. Un développement de la technologie axé sur les droits de l'homme pourrait dans un premier temps augmenter les coûts et forcer les développeurs et les entreprises à ralentir la cadence (notamment pour ce qui est des délais de mise sur le marché) ; il convient en effet d'évaluer d'abord l'impact des biens et services sur les droits de l'individu et sur la société. À moyen ou long terme, cependant, cette approche se traduira par une réduction des coûts et une meilleure efficacité (outils de prévision et d'aide à la décision plus précis, confiance accrue [Forum économique mondial, 2018], plaintes moins nombreuses, etc.). En outre, les entreprises et la société ont la maturité suffisante pour considérer que **la responsabilité envers l'individu et la collectivité est l'objectif premier du développement de l'intelligence artificielle**.

Alternativement, si l'IA emprunte une voie différente – à l'instar de technologies antérieures aux premiers stades de leur développement –, le risque, dans un environnement échappant à toute régulation, est que ses évolutions ne soient motivées que par des considérations purement techniques (faisabilité) ou commerciales ou par des intérêts politiques, c'est-à-dire des critères qui ne garantissent pas en eux-mêmes le respect des droits de l'homme.

Toute évolution de l'intelligence artificielle centrée sur les données devrait par conséquent reposer sur les principes de la Convention 108, véritables fondements d'une société numérique florissante. Les principaux piliers de cette approche sont les suivants :

- **Proportionnalité** (le développement de l'IA devrait être inspiré par le principe de proportionnalité[5] : l'efficience ne doit pas l'emporter sur les droits et libertés de l'individu ; l'individu a droit à ne pas être subordonné à des procédés automatisés ; le législateur devrait s'efforcer d'encadrer les applications de l'intelligence artificielle en vue de sauvegarder les intérêts individuels et collectifs)
- **Responsabilité** (au-delà de la simple obligation de rendre des comptes, développeurs et décideurs doivent agir de manière socialement responsable. Cela suppose aussi la création d'organes spécifiques pour appuyer et contrôler leur travail)
- **Gestion des risques** (une intelligence artificielle responsable, c'est évaluer les conséquences potentiellement négatives des applications de l'IA et prendre des mesures appropriées pour prévenir ou atténuer ces conséquences)
- **Participation** (des démarches participatives d'évaluation des risques sont essentielles pour donner la parole aux citoyens. En même temps, la participation citoyenne ne veut pas dire une moindre responsabilité des décideurs)
- **Transparence** (malgré les limitations actuelles qui affectent la transparence de l'intelligence artificielle, un certain degré de transparence peut contribuer à assurer la participation effective des citoyens et à évaluer plus précisément les conséquences des applications de l'IA)

I.4 Les principes et cadres existants

Le cadre réglementaire en vigueur applicable à l'intelligence artificielle et au traitement des données s'appuie principalement sur la Convention 108, même si d'autres instruments juridiques concernant la protection des données (notamment des recommandations[6] et lignes directrices[7]) peuvent aussi être pertinents pour des domaines spécifiques. Dans ce contexte, les lignes directrices sur les mégadonnées adoptées par le Conseil de l'Europe [Conseil de l'Europe, 2017] constituent la première tentative visant à aborder le recours à des outils d'aide à la décision faisant un usage intensif de données et s'inscrivent dans un ensemble plus large de documents et résolutions adoptés par plusieurs institutions européennes pour réguler l'impact des algorithmes sur la société [Comité d'experts du Conseil de l'Europe sur les intermédiaires internet (MSI-NET), 2018 ; CEPD (Le contrôleur européen de la protection des données) – Groupe consultatif sur l'éthique, 2018 ; Parlement européen, 2017 ; Agence des droits fondamentaux de l'Union européenne (FRA), 2018].

L'objet des lignes directrices sur les mégadonnées était de « contribuer à la protection des personnes concernées à l'égard du traitement des données à caractère personnel dans le contexte des mégadonnées en précisant les principes applicables en matière de protection des données et les pratiques correspondantes, en vue de limiter les risques que l'utilisation de mégadonnées comporte pour les droits des personnes concernées. Ces risques sont principalement liés au caractère potentiellement biaisé de l'analyse des données, à la sous-estimation des implications juridiques, sociales et éthiques du recours aux mégadonnées pour prendre des décisions et à la marginalisation d'une participation effective et éclairée des personnes à ces processus ».

Bien que ciblées sur l'analytique des mégadonnées (*Big Data analytics*), ces Lignes directrices couvrent des questions diverses impliquant le recours à de complexes applications à usage intensif de données lors de la prise de décision. C'est pourquoi des considérations relatives au rôle potentiellement positif de l'évaluation des risques (recouvrant des préoccupations éthiques et sociétales), du contrôle, de la minimisation des données, des comités d'experts, d'une approche

[5] Voir Convention modernisée pour la protection des personnes à l'égard du traitement automatisé des données à caractère personnel, art. 5.
[6] Voir par exemple Recommandation CM/Rec(2018)2 du Comité des Ministres aux États membres sur les rôles et les responsabilités des intermédiaires d'internet.
[7] Voir par exemple *Guide pratique sur l'utilisation de données à caractère personnel dans le secteur de la police* (2018), ou *Lignes directrices sur la protection des personnes à l'égard du traitement des données à caractère personnel à l'ère des mégadonnées* (2017).

prudente[8] et de la liberté des décideurs (personnes physiques) peuvent également s'appliquer à l'intelligence artificielle.

Certaines de ces solutions sont examinées plus en détail dans ce rapport (voir *infra*, Partie II). Mais les applications concrètes de l'IA rendent nécessaire une analyse de problématiques nouvelles (comme le rôle de la transparence et des diverses valeurs qui devraient sous-tendre les applications de l'IA) et conduisent à envisager de nouvelles solutions (par exemple élargir la portée de l'analyse d'impact relative à la protection des données ou fixer des limites aux usages potentiels de l'IA). Enfin, les organes de contrôle existants (comme les autorités de contrôle de la protection des données) devraient peut-être reconsidérer l'approche adoptée jusqu'à présent compte tenu des nouveaux enjeux soulevés par l'intelligence artificielle et de leurs conséquences potentielles pour la société.

En ce sens, l'IA – d'une manière analogue[9] aux mégadonnées[10] – rend malaisée l'application de certains principes classiques du traitement des données[11] ; ceci pourrait justifier la recherche de nouvelles solutions applicatives pour assurer la protection des données personnelles et des droits fondamentaux.

I.5 L'autodétermination de l'individu dans le traitement des données

Ces dernières années, les chercheurs dont les travaux portent sur la protection de la vie privée n'ont cessé de souligner la faiblesse, en terme d'autodétermination, du consentement de la personne concernée. La longueur et le caractère excessivement technique des notices d'information sur le traitement des données, un *lock-in* (dépendance) social et technique, une conception peu claire des interfaces et l'insuffisante sensibilisation des intéressés sont certaines des raisons de cette faiblesse.

En outre, un profilage reposant sur l'intelligence artificielle et des pratiques cachées de *nudging* (orienter les comportements) mettent en question tant l'idée de la liberté de choix sur la base de l'accord contractuel que la notion selon laquelle les personnes concernées gardent la maîtrise de leurs données. Enfin, vu la fréquente complexité et opacité des algorithmes d'IA, le consentement obtenu a peu de chances d'être réellement éclairé.

Les juristes ont abordé ces questions en mettant en avant **le rôle de la transparence** [voir entre autres Edwards et Vale, 2017 ; Selbst et Powles, 2017 ; Wachter, Mittelstadt et Floridi, 2017 ; Burrell, 2016 ; Rossi, 2016]**, l'évaluation des risques** [Lignes directrices, 2017 ; Mantelero, 2017] **ou des formes plus souples de consentement**, comme le consentement large [Sheehan, 2011] ou le consentement dynamique [Kaye et al., 2015]. Aucune de ces approches ne fournit certes une réponse définitive au problème du consentement individuel ; dans certains contextes, cependant, ces solutions, seules ou combinées, pourraient renforcer l'autodétermination.

De surcroît, la notion d'autodétermination ne se circonscrit pas à un cas particulier de traitement des données. Elle peut être utilisée au sens large pour faire référence à la liberté de choix concernant l'usage de l'intelligence artificielle et au droit à une version « non intelligente » de biens et services

---

[8] Voir aussi Commission européenne – Groupe européen d'éthique des sciences et des nouvelles technologies, 2018, p. 18 (*« La possibilité d'abus d'utilisation des technologies "autonomes" constitue un défi majeur qui justifie qu'une importance cruciale soit accordée à la sensibilisation au risque et à une approche prudente »).*

[9] Voir aussi à ce propos Autorité norvégienne de protection des données, 2018 *(« Le présent rapport expose plus en détail les avis juridiques et les technologies décrites dans le rapport 2014 sur les données massives, intitulé "*Big Data – data protection principles under pressure"*. Nous nous attacherons ici à décrire l'intelligence artificielle (IA) en apportant davantage de précisions techniques, tout en nous intéressant de plus près à quatre enjeux majeurs de l'IA associés aux principes de protection des données énoncés dans le RGPD : loyauté et absence de discriminations, limitation des finalités, minimisation des données, transparence, droit à l'information »*).

[10] Voir Lignes directrices, section II *(« Compte tenu de la nature des mégadonnées et de leur utilisation, l'application de certains principes traditionnels du traitement de données [principe de minimisation des données ; de finalité ; de loyauté et de transparence ; consentement libre, spécifique et éclairé], pourrait poser des difficultés dans ce scénario technologique »*).

[11] À titre d'illustration, les procédés analytiques peuvent rendre difficile l'identification de la finalité spécifique du traitement au moment de la collecte des données. Par ailleurs, s'agissant des algorithmes d'apprentissage automatique, dont les finalités sont nécessairement spécifiées, il n'est pas toujours possible de prédire et expliquer comment ces objectifs seront atteints. Dans l'un et l'autre cas, la transparence concernant les finalités et les modalités de traitement des données peut rester limitée.

intégrant des technologies d'IA [Commissariat à la protection de la vie privée du Canada, 2016 – *« … à mesure que les appareils et dispositifs intelligents deviendront la norme, on assistera de plus en plus à une "érosion du choix" pour les personnes qui auraient préféré leurs versions "non intelligentes" »*][12]. Cette « option zéro » dépasse la dimension individuelle et renvoie aussi à la façon dont une collectivité décide **du rôle qui doit être dévolu à l'IA : à quel point peut-elle façonner les dynamiques sociales et les comportements collectifs ou orienter des décisions affectant des groupes entiers d'individus ?** [« Principes d'Asilomar », 2017 – *« Contrôle humain : les hommes devraient choisir s'ils veulent ou non – et comment – déléguer des décisions aux systèmes d'IA pour atteindre les objectifs qu'ils se sont fixés »*].

I.6 Minimisation

À l'instar de l'analytique des mégadonnées [Lignes directrices, 2017], la minimisation des données[13] ne va pas sans poser problème. En effet, même si les technologies diffèrent, l'analytique des mégadonnées et les algorithmes d'apprentissage automatique ont besoin de quantités massives de données pour produire des résultats utiles. Partant, seul un certain degré de minimisation est possible.

De plus, comme pour « l'option zéro » mentionnée plus haut, l'adoption de solutions autres que celles intégrant l'IA peut contribuer à réduire la quantité de données collectées en limitant la somme d'information requise (mener une étude auprès d'un échantillon représentatif plutôt qu'auprès d'une grande partie de la population par exemple).

Par ailleurs, une partie des lignes directrices du Conseil de l'Europe sur les mégadonnées (Big Data) peuvent être étendues à l'intelligence artificielle. Les Lignes directrices contiennent un principe qui est également applicable à l'IA : les données devraient être collectées et traitées de façon à « minimiser la présence de données redondantes ou marginales »[14]. Dans le cas de l'IA, ceci concerne principalement les jeux de données d'apprentissage. L'Autorité norvégienne de protection des données a souligné qu'il serait naturel de commencer par un volume restreint de données d'apprentissage, puis de vérifier l'exactitude du modèle lorsqu'il est alimenté par de nouvelles données [Autorité norvégienne de protection des données, 2018]. En outre, les études pourraient aussi porter sur la conception d'algorithmes qui effacent progressivement les données en utilisant des mécanismes d'oubli automatiques [Gama et al., 2013, p. 12-13].

Même si l'apprentissage automatique requiert nécessairement de grandes quantités de données lors de la phase d'apprentissage, il est important d'adopter un paradigme de conception qui **évalue de manière critique la nature et la quantité des données utilisées**, en réduisant les données redondantes ou marginales et en augmentant progressivement la taille du jeu de données d'apprentissage[15]. L'objectif de minimisation peut aussi être atteint si l'algorithme apprend en utilisant des données synthétiques[16] [ministère britannique du Numérique et de la Culture, 2018] issues d'un sous-ensemble de données personnelles qui aura été anonymisé une fois constitué [Barse et al., 2003].

---

[12] Voir aussi Protocole d'amendement à la Convention pour la protection des personnes à l'égard du traitement automatisé des données à caractère personnel (STE n° 108), Rapport explicatif, par. 40.

[13] Voir aussi Convention modernisée pour la protection des personnes à l'égard du traitement automatisé des données à caractère personnel, art. 5.

[14] Lignes directrices, section IV, par. 4.2.

[15] Voir aussi Lignes directrices, section IV, par. 4.3 *(« Lorsque cela est techniquement faisable, les responsables du traitement des données et, le cas échéant, les sous-traitants devraient tester l'adéquation des solutions adoptées dès la conception sur un volume limité de données au moyen de simulations, avant leur utilisation à une plus grande échelle »).*

[16] Les données synthétiques sont générées à partir d'un modèle de données construit sur la base de données réelles. Elles devraient être représentatives des données réelles initiales. Pour une définition des données synthétiques, se reporter au glossaire des termes statistiques de l'OCDE [en anglais seulement : *Glossary of Statistical Terms*, 2007, http://ec.europa.eu/eurostat/ramon/coded_files/OECD_glossary_stat_terms.pdf) – *« Approche de la confidentialité où les données diffusées ne sont pas des données réelles mais des données synthétiques générées à partir d'un ou plusieurs modèles de population »*].

I.7 Biais

Même si une réduction ou une suppression des biais humains est possible grâce à des systèmes d'IA performants, il est aussi possible que les applications faisant un usage intensif de données soient affectées par un biais potentiel. En effet, tant l'apprentissage déterministe que l'apprentissage automatique utilisent les données saisies pour extraire de nouvelles informations (procédés analytiques) ou pour créer et entraîner des modèles d'apprentissage automatique. Le biais peut tenir aux méthodes employées par les scientifiques des données (biais de mesure, biais affectant les méthodologies d'enquête, biais aux stades du nettoyage et du prétraitement) [Veale et Binns, 2017], à l'objet de leur recherche (par exemple biais social dû au biais historique[17] ou sous-représentation de certaines catégories) [Forum économique mondial, 2018, p. 8-9], à leurs sources de données (biais de sélection) ou à la personne responsable de l'analyse (biais de confirmation) [ministère britannique du Numérique et de la Culture, 2018 ; Information Commissioner's Office (ICO), 2017, p. 43-44 ; AI Now Institute, 2016].

Les jeux de données biaisés peuvent affecter négativement les algorithmes, tout particulièrement dans le cas de l'apprentissage automatique (*Machine Learning*), où le biais peut affecter la conception et le développement (l'apprentissage) de l'algorithme. Cette question a déjà été en partie abordée par les lignes directrices du Conseil de l'Europe sur les mégadonnées (Big Data), qui proposent de privilégier des **solutions dès la conception (*by-design*) pour « éviter ainsi tout biais caché potentiel et tout risque de discrimination ou d'impact négatif sur les droits et libertés fondamentales des personnes concernées, lors de la collecte comme de l'analyse »**[18]**.**

Les biais peuvent être dus à des jeux de données biaisés [AI Now Institute, 2017, 4, p. 16-17], mais ils peuvent aussi résulter des décisions des développeurs, qu'elles soient intentionnelles ou non. En ce sens, les prédictions et les performances des machines sont limitées par les décisions et valeurs humaines, et les personnes chargées de la conception, du développement et de la maintenance des systèmes d'IA façonneront ces systèmes selon leur propre vision du monde [AI Now Institute, 2017, p. 18]. C'est pourquoi le développement de l'intelligence artificielle ne peut pas être laissé aux mains des seuls concepteurs : de par leur formation technique, ils pourront en effet être moins sensibles aux conséquences sociétales de leurs décisions.

Des comités d'experts de différents domaines (sciences sociales, droit, éthique, etc.) peuvent constituer le meilleur cadre pour engager le débat et aborder la question de l'impact de l'IA sur l'individu et sur la société (voir *infra*, chapitre II.3.1), compensant ainsi le point de vue limité des développeurs. Des comités pluridisciplinaires pourraient aussi être en mesure de déceler les biais potentiels qui dépendent de l'identité de développeurs, comme les stéréotypes de genre, les biais idéologiques ou la sous-représentation des minorités [AI Now Institute, 2016, p. 5].

Les risques de biais dans les applications d'IA peuvent aussi être réduits grâce à des **démarches participatives d'évaluation des risques** [Mantelero, 2018] reposant non seulement sur la sécurité et la qualité des données (voir *infra*, chapitre II.3.2), mais aussi sur l'engagement actif des groupes potentiellement affectés par les applications d'IA. De telles démarches peuvent contribuer à la détection et à la suppression des biais existants [AI Now Institute, 2016, p. 24].

Cette approche, axée sur une conception responsable de l'intelligence artificielle [Lignes directrices, 2017][19], vise à éviter les conditions biaisées qui peuvent affecter les jeux de données ou les

[17] Voir par exemple l'article intitulé « Amazon ditched AI recruiting tool that favored men for technical jobs », *The Guardian*, 10 octobre 2018, https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine *(« … En 2015, Amazon s'est cependant rendu compte que le système notait les candidats aux postes de développeur de logiciel et aux autres postes techniques de manière sexiste. Ceci s'expliquait par le fait que le modèle informatique utilisé par Amazon s'appuyait sur les CV reçus par le groupe sur une période de dix ans, qui étaient pour la plupart ceux des hommes, reflet de la prédominance masculine dans le secteur des nouvelles technologies » – https://fr.reuters.com/article/technologyNews/idFRKCN1MK26B-OFRIN*).

[18] Lignes directrices, section IV, par. 4.2.

[19] Lignes directrices, section IV, par. 4.2 (*« Les responsables du traitement et, le cas échéant, les sous-traitants devraient soigneusement examiner la conception du traitement de données afin de minimiser la présence de données redondantes ou marginales et d'éviter ainsi tout biais caché potentiel et tout risque de discrimination ou d'impact négatif sur les droits et libertés fondamentales des personnes concernées, lors de la collecte comme de l'analyse »*).

algorithmes. Dans un contexte nécessairement caractérisé par un certain degré d'opacité et de complexité, des évaluations en amont et une conception responsable peuvent être plus efficaces qu'une analyse motivée par la découverte d'un résultat discriminatoire [Selbst, 2017, p. 163 – *« Même si le résultat peut être attribué à un problème de qualité des données, il est souvent assez difficile d'y remédier. Si l'on peut aisément déterminer que quelque chose ne va pas dans les données, comprendre de quoi il s'agit peut s'avérer plus compliqué […] Même si toutes les sources de biais sont repérées, l'ampleur de l'effet de chaque source demeure largement inconnue »* ; Brauneis et al. 2018, p. 131].

Prêter attention aux biais potentiels aux tous premiers stades de la conception [ministère britannique du Numérique et de la Culture, 2018] implique aussi une réflexion plus approfondie sur les jeux de données d'apprentissage et la phase d'apprentissage en général, pour limiter les conséquences négatives du biais historique dans les jeux de données préexistants. Sur ce point, d'aucuns ont proposé de mettre l'accent sur la traçabilité afin de pouvoir connaître l'origine, le développement et l'utilisation des jeux de données d'apprentissage tout au long de leur cycle de vie [AI Now Institute, 2017].

Des tests réalisés avec soin lors de la phase d'apprentissage, avant le déploiement des algorithmes à grande échelle, pourraient révéler des biais cachés. C'est pourquoi les lignes directrices sur les mégadonnées (Big Data) insistent sur le rôle des simulations [Lignes directrices[20] ; AI Now Institute, 2017]. En outre, des biais cachés pourraient aussi tenir à des **biais générés par la machine, qui diffèrent des biais humains** [Cummings, 2018, p. 2 – *« Les machines et les hommes ont des capacités différentes et, tout aussi important, commettent des erreurs différentes basées sur des architectures de prise de décision fondamentalement divergentes »* ; Caruana et al., 2015 ; Szegedy et al., 2013].

Dans le contexte de l'IA, l'**évaluation des biais potentiels peut aussi porter à controverse**, étant donné les multiples variables à prendre en compte et la classification des personnes en groupes qui ne correspondent pas nécessairement aux catégories discriminatoires traditionnelles [Donovan et al., 2018, p. 5]. Les questions qui se posent concernant le **biais des machines ne sauraient être balayées par l'argument selon lequel les décisions humaines sont faillibles,** l'IA étant vue comme un moyen de réduire l'erreur humaine. Quatre raisons montrent que cette comparaison n'est pas valable.

Premièrement, les solutions d'IA sont conçues pour être appliquées en série. Comme dans le régime de la responsabilité du fait des produits défectueux, une mauvaise conception (ici, le biais) affectera inévitablement de nombreuses personnes dans des circonstances identiques ou similaires, tandis qu'une erreur humaine est circonscrite à un cas particulier.

Deuxièmement, même s'il existe des domaines dans lesquels les taux d'erreur de l'IA sont proches, voire inférieurs à ceux du cerveau humain (taux d'erreur de la reconnaissance d'images et de leur étiquetage par exemple) [Artificial Intelligence Index, 2017], pour la plupart des tâches de prise de décision plus complexes, les taux d'erreur sont supérieurs[21] [Cummings et al., 2018, p. 13].

Troisièmement, l'erreur humaine a une dimension socioculturelle. Ceci la distingue d'une erreur des machines en termes d'acceptabilité sociale et d'exonération et a nécessairement une influence sur la propension à adopter des solutions d'IA potentiellement faillibles.

Enfin, la comparaison des conséquences néfastes des décisions humaines et des décisions d'une IA [ministère fédéral des Transports et des Infrastructures numériques, 2017, p. 10 – *« L'octroi de licences à des systèmes automatisés n'est justifiable que si cela promet au moins une diminution des*

[20] Lignes directrices, section IV, par. 4.3 (*« Lorsque cela est techniquement faisable, les responsables du traitement des données et, le cas échéant, les sous-traitants devraient tester l'adéquation des solutions adoptées dès la conception sur un volume limité de données au moyen de simulations, avant leur utilisation à une plus grande échelle. Une telle approche permettrait d'évaluer le préjudice potentiel dans l'utilisation des différents paramètres d'analyse des données et d'apporter des éléments en vue de minimiser l'utilisation des informations et de réduire les incidences négatives potentielles identifiées dans le cadre du processus d'évaluation des risques décrit à la section IV.2 »*).

[21] Voir par exemple Aletras N., Tsarapatsanis D., Preoţiuc-Pietro D., Lampos V., « Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective », *PeerJ Computer Science* 2:e93, 2016, https://doi.org/10.7717/peerj-cs.93.

dommages par rapport à la conduite humaine ; autrement dit, le bilan doit être positif s'agissant des risques »] est essentiellement fondée sur la simple comparaison chiffrée des dommages qui en résultent (par exemple nombre de victimes de voitures conduites par l'homme et nombre de victimes de véhicules entièrement autonomes pilotés par des IA), ce qui est trop réducteur. Pour apprécier les conséquences des décisions humaines et des décisions d'une IA, nous devons **prendre en considération la répartition des effets** (par exemple l'appartenance des individus touchés négativement à différentes catégories, les diverses conditions dans lesquelles le préjudice a été occasionné, la gravité des conséquences, etc.). Une telle approche quantitative paraît en outre en contradiction avec l'approche de précaution [Lignes directrices, 2017], qui requiert l'adoption de politiques de prévention des risques plutôt qu'une simple réduction des dommages.

Partie II - Enjeux et solutions possibles

II.1 Limitations de l'usage de l'intelligence artificielle

La réglementation sur la protection des données, de même que la Convention 108, prévoit des garanties qui peuvent également s'appliquer aux algorithmes (y compris des algorithmes d'IA) utilisés dans les systèmes de prise de décision automatisée. Cependant, la simple existence d'un processus de prise de décision non humain ne saurait suffire pour tracer la ligne rouge entre décisions humaines et décisions automatisées. En effet, la nature soi-disant fiable des solutions d'IA fondées sur les mathématiques peut pousser les personnes qui s'appuient sur des algorithmes pour prendre leurs décisions à faire confiance au tableau des individus et de la société suggéré par les procédés analytiques. Cette attitude peut en outre être renforcée par la menace de sanctions potentielles si une décision a été prise en ignorant les résultats des procédures analytiques. Dès lors, la présence d'un intervenant humain n'est pas en soi suffisante.

L'apparence d'objectivité mathématique bénéficie aux algorithmes d'IA. Si l'on ajoute à cela la complexité de la gestion des données et la position subordonnée de la personne décisionnaire au sein d'une organisation, il peut être difficile pour un décideur humain de prendre une décision autre que celle suggérée par l'algorithme[22].

À la lumière de ce qui précède, il y a lieu de faire une distinction entre les cas où le décideur humain a une liberté effective et ceux où sa marge de manœuvre est nulle. À ce propos, les lignes directrices sur les mégadonnées ont déjà souligné **l'importance de protéger la liberté effective du décideur (personne physique)**[23]**.**

Les comités d'experts (voir *infra*, chapitre II.3.1) peuvent jouer un grand rôle dans l'évaluation des cas de déséquilibre potentiel : cela pourrait aussi faciliter la participation des parties intéressées à l'évaluation (voir *infra*, chapitre II.3.2).

Lorsque les décisions peuvent être déléguées à des systèmes d'IA ou que les décideurs humains n'ont pas la possibilité d'exercer un contrôle suffisant sur les décisions retenues par l'IA, la question plus large qui se pose est de savoir s'il convient d'adopter ces systèmes plutôt que des méthodes où l'homme est au cœur de la prise de décision[24]. Cela devrait conduire les collectivités ou les groupes

---

[22] Voir aussi Brauneis, Robert et Ellen P. Goodman, « Algorithmic Transparency for the Smart City », *Yale Journal of Law & Technology*, vol. 20, nº 103, 2018, p. 126-127 (*« Avec le temps, s'en remettre aveuglément aux algorithmes pourrait affaiblir la capacité de prise de décision des agents de l'État de même que leur sens de l'engagement et de l'action »*).

[23] Lignes directrices, section IV.7.4 *(« Sur la base d'arguments raisonnables, le décideur (personne physique) devrait se voir conférer la liberté de ne pas se baser sur les résultats des recommandations découlant de l'utilisation des mégadonnées »*).

[24] Voir par exemple UIT, 2017, p. 34 (*« Margaret Chan, [désormais ancienne] directrice générale de l'OMS, a observé que les "décisions médicales sont très complexes et reposent sur de nombreux paramètres, parmi lesquels l'empathie et la compassion à porter aux patients. Je doute qu'une machine puisse imiter ces émotions – ou agir avec compassion. Les machines peuvent rationaliser, simplifier, mais l'IA ne peut pas remplacer les médecins et les infirmières dans leurs interactions avec les patients" »*). Voir aussi l'article 5 de la Convention 108 modernisé et le rapport explicatif qui souligne que ce principe [de proportionnalité] *« doit être respecté à toutes*

potentiellement concernés à engager **un débat participatif sur l'adoption de solutions d'IA** en analysant les risques éventuels (voir *infra*, évaluation des risques) et, dans l'hypothèse où ces solutions seraient retenues, à contrôler leur application (voir *infra*, vigilance).

II.2 Transparence

Dans le contexte de l'intelligence artificielle, la transparence[25] peut avoir plusieurs sens différents. Cela peut consister à dévoiler les applications d'IA qui sont utilisées et à donner une description de leur logique ou accès à la structure des algorithmes d'IA et – le cas échéant – aux jeux de données employés pour leur apprentissage. La transparence peut en outre être une exigence *ex ante* ou *ex post* [voir par exemple Binns et al., 2018] pour la prise de décision centrée sur les données.

Même si la transparence est importante pour le contrôle public des modèles de prise de décision automatisée [Reisman et al., 2018, p. 5], une déclaration générique concernant l'usage de l'intelligence artificielle ne contribue guère à lutter contre les risques d'usage déloyal ou illicite des données. Par ailleurs, si accéder à la structure de l'algorithme peut permettre de déceler un biais potentiel, les droits de propriété intellectuelle et les questions potentielles de concurrence restreignent parfois cet accès ; en tout état de cause, même en l'absence de telles barrières, la complexité des modèles adoptés peut représenter un défi majeur pour la cognition humaine [Lipton, 2018, p. 13]. En outre, dans certains cas, la transparence peut empêcher les corps de la fonction publique de s'acquitter de leur mission (systèmes de police prédictive par exemple), ou entrer en conflit avec les obligations en matière de sécurité du responsable du traitement lorsque les données concernées sont celles de personnes autres que celles demandant l'accès [Veale et al., à paraître en 2018][26].

C'est pourquoi une solution axée sur la divulgation de la logique des algorithmes pourrait constituer la meilleure option[27]. Même ainsi, ceci peut être interprété de manière plus ou moins large. Donner des informations sur le type de données de départ et les résultats attendus[28], expliquer les variables et les pondérations opérées ou apporter un éclairage sur l'architecture analytique sont autant de formes de transparence sur la logique qui sous-tend les algorithmes d'IA.

Des processus d'analyse complexe (l'apprentissage profond par exemple) rendent plus ardu l'exercice de transparence – notamment lorsqu'il s'agit d'expliquer la logique des algorithmes [Goodman et Flaxman, 2016] et les décisions fondées sur l'analyse des données[29] ; les systèmes non déterministes ne facilitent pas non plus la tâche si l'on veut communiquer des informations détaillées sur la logique qui préside au traitement des données.

En outre, de nombreux algorithmes sont de nature dynamique, alors que la transparence est de nature statique. Les algorithmes sont constamment mis à jour et modifiés. En revanche, toute divulgation effectuée au nom de la transparence ne concerne que l'algorithme tel qu'il est utilisé à un moment donné.

Enfin, l'accès aux algorithmes d'IA ne suffit pas pour déceler des biais éventuels. Des ressources sont aussi requises, en termes de temps et de compétences, pour effectuer ce type d'analyse

---

les étapes du traitement, y compris au stade initial, c'est-à-dire lorsqu'il est décidé de procéder ou non au traitement des données ».

[25] Voir Convention 108 modernisée, art. 8.

[26] Quoi qu'il en soit, les algorithmes sont parfois plus difficile à lire et à comprendre que le langage mathématique, la notation logique ou le langage naturel ; *« par conséquent, la divulgation du code informatique pourrait s'avérer la démarche la moins utile par rapport à d'autres moyens d'interprétation plus faciles »*, Brauneis, Robert et Ellen P. Goodman, « Algorithmic Transparency for the Smart City », *Yale Journal of Law & Technology*, vol. 20, 2018, p. 103 et 130.

[27] Voir aussi Convention modernisée pour la protection des personnes à l'égard du traitement automatisé des données à caractère personnel, article 9.1.*c*.

[28] Ces informations pourraient être fournies par le biais de modèles « d'apprentissage par l'utilisation », en donnant aux personnes concernées la possibilité de tester les outils d'analyse avec différentes variables de départ. Même alors, cependant, il existe un risque d'identification erronée des variables de départ pertinentes [Diakopoulos, 2013, p. 18].

[29] Voir par exemple l'article 10 de la loi n° 78-17 du 6 janvier 1978 (loi Informatique et Libertés), tel que modifié par la loi n° 2018-493 du 20 juin 2018. Dans certains cas, il peut être impossible d'expliquer les raisons qui justifient la décision suggérée par l'algorithme [Burrell, 2016]. De surcroît, des solutions telles que le droit à l'explication se limitent aux décisions relatives à l'intéressé et ne permettent pas d'aborder les problématiques collectives soulevées par l'usage de l'IA au niveau du groupe.

[Ananny et Crawford, 2016 – *« L'idéal de transparence peut devenir écrasant au sens où cela impose de rechercher des informations concernant un système, d'interpréter ces informations et de déterminer leur signification »*]. En conséquence, l'effet dissuasif de solutions telles que la réalisation d'un audit [Veale et Binns, 2017] ou l'intervention de décideurs humains est compromis[30]. Les travaux de recherche actuels s'efforcent de développer des méthodes de détection des biais qui sont elles-mêmes fondées sur des algorithmes[31] ; cependant, on voit mal comment l'introduction d'un système algorithmique de surveillance des algorithmes peut réduire la complexité de la gouvernance des données.

Aucun de ces points n'affaiblit l'argument en faveur d'une transparence accrue en général [Burrell, 2016], notamment dans le secteur public[32], et de son rôle dans la préservation de l'autodétermination de la personne concernée [Edwards et Vale, 2017 ; Selbst & Powles, 2017 ; Wachter, Mittelstadt et Floridi, 2017 ; Rossi, 2016]. Si la transparence est difficile à atteindre en ce qui concerne l'architecture et la logique des algorithmes, cela pourra néanmoins contribuer à clarifier les raisons sous-tendant la décision d'utiliser un outil aussi complexe [Burt et al., 2018, p. 2].

La transparence n'est qu'une partie de la solution aux enjeux de l'IA et a plusieurs limitations qui doivent être pleinement prises en considération [Ananny et Crawford, 2016]. Il ne faut pas oublier que les algorithmes sont seulement une composante de l'application d'IA, l'autre étant les jeux de données utilisés pour l'apprentissage ou pour l'analyse. Or des jeux de données biaisés produisent automatiquement des résultats biaisés.

Enfin, certaines applications à usage intensif de données privilégient les données décontextualisées, en ignorant l'information contextuelle qui est souvent primordiale pour comprendre et appliquer la solution proposée par l'application d'IA. La **décontextualisation** n'est pas non plus sans danger lors du choix des modèles algorithmiques – lorsque des modèles initialement utilisés pour une finalité précise sont ensuite réutilisés dans un contexte différent et à d'autres fins [Donovan et al. (2018, p. 7) citent le cas de l'algorithme PredPol, inspiré d'un algorithme conçu à l'origine pour prédire les séismes et utilisé par la suite pour établir une cartographie des lieux à risque d'infraction élevé et décider de l'affectation des policiers sur le terrain] – ou lors de l'utilisation de modèles dont l'apprentissage a été fait avec les données historiques d'une population différente [AI Now Institute, 2018].

II.3.1 Évaluation des risques

Étant donné les limites de la transparence et de l'autodétermination de l'individu (voir *supra*, chapitre I.5), la réglementation en matière de protection des données met de plus en plus l'accent sur l'évaluation des risques[33]. L'évaluation des risques par le responsable du traitement et un environnement sécurisé peuvent considérablement renforcer la confiance et la propension à utiliser des applications d'IA. Les préférences des usagers peuvent être déterminées par une analyse des risques effective et des mesures d'atténuation des risques [Autorité norvégienne de protection des données, 2018, p. 4] plutôt que par des campagnes marketing ou la simple image de marque.

---

[30] Ces solutions sont possibles, mais dans de nombreux cas la réalisation de l'audit exige des efforts considérables ; quant à l'intervention humaine, elle se heurte à la complexité du traitement des données.

[31] Voir par exemple Lomas, Natasha, *IBM Launches Cloud Tool to Detect AI Bias and Explain Automated Decisions*, TechCrunch (blog), 19 septembre 2018, consulté le 21 septembre 2018, http://social.techcrunch.com/2018/09/19/ibm-launches-cloud-tool-to-detect-ai-bias-and-explain-automated-decisions/.

[32] Voir par exemple l'article 10.2 de la loi n° 78-17 du 6 janvier 1978 (loi Informatique et Libertés), tel que modifié par la loi n° 2018-493 du 20 juin 2018. Le secteur public accorde une grande attention au principe d'égalité de traitement dans son utilisation des algorithmes et s'attache à garantir la transparence et les droits d'accès dans ses procédures administratives. S'agissant des limitations qui peuvent affecter la transparence algorithmique dans le secteur public, voir Brauneis, Robert et Ellen P. Goodman, « Algorithmic Transparency for the Smart City », *Yale Journal of Law & Technology*, vol. 20, 2018, p. 103-176 *(« … il existe trois grands obstacles qui brouillent la transparence dans l'usage des prédictions fondées sur des données par les pouvoirs publics : 1) l'absence de pratiques appropriées de constitution des dossiers autour des processus algorithmiques ; 2) le fait que les pouvoirs publics n'insistent pas suffisamment sur l'importance de pratiques de divulgation appropriées ; et 3) le secret commercial ou d'autres privilèges confidentiels revendiqués par les prestataires. Dans le présent article, nous étudions chacun de ces obstacles »*).

[33] Voir aussi Convention modernisée pour la protection des personnes à l'égard du traitement automatisé des données à caractère personnel, art. 10.2.

L'utilisation d'algorithmes par les techniques modernes de traitement des données [Comité d'experts sur les intermédiaires d'internet du Conseil de l'Europe (MSI-NET), 2018] ainsi que la tendance à avoir recours à des technologies à usage intensif de données [CEDP, 2018] ont conduit certains acteurs à avoir une **vision plus globale des éventuelles conséquences négatives** du traitement des données [Principes d'Asilomar, 2017 – *« Risques : Les risques posés par les systèmes d'intelligence artificielle, en particulier les risques catastrophiques ou existentiels, requièrent des efforts de préparation et d'atténuation proportionnés à l'impact attendu »*]. Des groupes d'experts et de spécialistes sont allés au-delà de la sphère traditionnelle de la protection des données [Taylor, Floridi et van der Sloot, 2017] pour **examiner l'incidence de l'usage des données sur les droits fondamentaux et les valeurs éthiques et sociales collectives** [Mantelero, 2018 ; Access Now, 2018].

Apprécier le respect des valeurs éthiques et sociales est plus compliqué que de procéder à l'évaluation traditionnelle du dispositif de protection des données. À titre d'illustration, alors que les valeurs (comme l'intégrité des données) qui sous-tendent la sécurité et la gestion des données reposent sur la technologie et peuvent donc être généralisées et appliquées à divers contextes sociaux, il n'en va pas de même avec les valeurs éthiques et sociales. Ces valeurs sont nécessairement liées au contexte et diffèrent d'une communauté à l'autre [Forum économique mondial, 2018, p. 12]. Il est donc plus difficile de déterminer un cadre de référence pour ce type d'évaluation.

Cette question est clairement abordée dans la première partie des lignes directrices sur les mégadonnées [Conseil de l'Europe, 2017], qui exhorte les responsables du traitement et les sous-traitants à « tenir dûment compte de l'impact potentiel du traitement des mégadonnées envisagé et de ses implications éthiques et sociales plus larges », en vue de garantir les droits de l'homme et les libertés fondamentales à la lumière de la Convention 108[34].

Le nouvel élément dans l'évaluation des risques concerne l'éventail des intérêts à sauvegarder et des droits à protéger. L'évaluation porte sur des droits qui dépassent le périmètre traditionnel de la protection des données, comme le droit à la non-discrimination[35] [Barocas et Selbsr, 2016][36] ou le respect des valeurs éthiques et sociales [Comité économique et social européen, 2017 ; AI Now Institute, 2017, p. 34-35 *(« Pour parvenir à des systèmes d'IA éthiques dans lesquels les implications plus larges seraient dûment prises en considération, des changements institutionnels s'imposent afin d'intégrer l'obligation de rendre des comptes »*) ; Access Now, 2018].

Les Lignes directrices reconnaissent le caractère relatif des valeurs éthiques et sociales et insistent sur le fait que le traitement des données ne devrait pas aller à l'encontre des « valeurs éthiques communément acceptées dans la communauté ou les communautés pertinentes, et ne devrait pas porter atteinte à des intérêts, des valeurs et des normes sociétaux »[37]. Tout en admettant que la définition des valeurs à prendre en considération dans le cadre d'une évaluation plus large risque de s'avérer problématique, elles proposent néanmoins quelques mesures concrètes allant dans ce sens. S'inspirant du point de vue de chercheurs spécialistes de la protection de la vie privée [Wright, 2011], elles font observer que « les valeurs éthiques communément reconnues figurent dans les instruments internationaux de protection des droits de l'homme et des libertés fondamentales tels que la Convention européenne des droits de l'homme ».

Dans la mesure où l'évaluation éthique et sociale dépend du contexte et où les textes internationaux ne fournissent que des orientations de haut niveau, les Lignes directrices complètent cette directive

[34] Lignes directrices, section IV, par. 1.1.
[35] Voir aussi Convention modernisée pour la protection des personnes à l'égard du traitement automatisé des données à caractère personnel, art. 6.2.
[36] Concernant l'IA et les voitures autonomes, voir aussi Ministère fédéral des Transports et des Infrastructures numériques, plan d'action du Gouvernement fédéral allemand pour donner suite au rapport de la commission d'éthique sur la conduite automatisée et connectée (Règles éthiques pour les systèmes de conduite automatisée), 2017, http://www.bmvi.de/SharedDocs/EN/publications/action-plan-on-the-report-ethics-commission-acd.pdf?__blob=publicationFile (En cas de dilemme, c.-à-d. une situation où les risques de blessures corporelles ne sauraient être écartés, la commission affirme que toute distinction fondée sur des caractéristiques personnelles [âge, sexe, etc.] est strictement interdite).
[37] Lignes directrice, section IV, par. 1.2.

générale par une option plus adaptée, à savoir la création de « comités d'éthique ad hoc »[38]. Dans l'hypothèse où l'évaluation révèlerait « un fort impact de l'utilisation des mégadonnées sur les valeurs éthiques », les comités, qui dans certains cas existent déjà en pratique, devront identifier les valeurs éthiques spécifiques qu'il convient de protéger dans le cadre de l'utilisation de ces données en donnant des orientations plus détaillées pour l'évaluation des risques, en fonction du contexte[39].

L'« architecture des valeurs » définie par les Lignes directrices comporte trois niveaux. On distingue d'abord un niveau général représenté par les « valeurs éthiques communément reconnues » énoncées dans les instruments internationaux de protection des droits de l'homme et des libertés fondamentales. Le deuxième niveau tient compte des déterminants contextuels de l'évaluation éthique et sociale et met l'accent sur les valeurs et les intérêts sociaux de la communauté considérée. Enfin, le troisième niveau consiste en un ensemble plus spécifique de valeurs éthiques définies par les comités d'éthique au regard de l'usage qui est fait des données.

L'évaluation est d'autant plus complexe que les risques potentiels évoluent constamment, de même que les mesures à prendre pour les maîtriser. À cet égard, les autorités de contrôle en matière de protection des données peuvent jouer un rôle important et appuyer les responsables du traitement en les informant des mesures visant la sécurité des données et en leur fournissant des orientations détaillées sur le processus d'évaluation des risques[40]. Par conséquent, les Lignes directrices ne laissent pas l'évaluation exclusivement aux mains des responsables du traitement. Conformément à l'approche adoptée dans le Règlement (UE) 2016/679, lorsque l'utilisation des mégadonnées est « susceptible d'avoir un impact important » sur les droits et libertés fondamentales des personnes concernées, les responsables du traitement devraient consulter les autorités de contrôle afin de chercher à obtenir des conseils pour réduire les risques mis en évidence dans l'étude d'impact[41].

Les lignes directrices sur les mégadonnées parviennent à une série de conclusions qui peuvent être étendues à l'IA en mettant l'accent sur l'automatisation de la prise de décision, qui est au cœur des enjeux les plus cruciaux de l'intelligence artificielle.

Enfin, l'augmentation de la charge de travail induite par l'élargissement de l'évaluation se justifie non seulement par la nature des droits et libertés susceptibles d'être affectés par les applications de l'intelligence artificielle, mais aussi parce que cela offre une opportunité d'obtenir des avantages comparatifs. En **renforçant la confiance[42]** dans les produits et services d'intelligence artificielle, les entreprises pourront mieux répondre aux préoccupations grandissantes des consommateurs concernant l'IA et l'usage qui est fait des données. De la même façon, une plus forte obligation de rendre des comptes des établissements publics concernant leurs systèmes d'IA accroît la confiance des citoyens dans l'administration publique et empêche des décisions inéquitables. De ce point de vue, un rôle important peut aussi être joué par les certifications [IEEE, 2016, p. 46 – *« Il nous faudra en outre développer un programme de certification des technologies IA/SA garantissant que ces technologies ont fait l'objet d'une évaluation indépendante et été jugées à la fois sûres et éthiques »* ; voir également Brundage et al., 2018, p. 56 et 93], les codes de conduite et les normes. Ces différents outils contribuent à accroître la responsabilité et donnent des orientations relatives à la sécurité des données et à l'intégrité des systèmes[43] qui englobent des procédures assurant une traçabilité du processus de prise de décision et empêchant toute forme de manipulation des résultats générés.

---

[38] Lignes directrices, section IV, par. 1.3 *(« Si l'évaluation de l'impact potentiel d'un traitement de données envisagé, telle que décrite à la section IV.2, révèle un fort impact de l'utilisation des mégadonnées sur les valeurs éthiques, les responsables du traitement des données peuvent établir un comité d'éthique ad hoc, ou s'appuyer sur les existants, afin d'identifier les valeurs éthiques spécifiques qu'il convient de protéger dans le cadre de l'utilisation de ces données »*).

[39] Ce même modèle à deux niveaux, fondé sur des lignes directrices générales et des orientations adaptées fournies par des comités ad hoc, est déjà adopté dans les essais cliniques. Comme pour les mégadonnées, l'application spécifique de la technologie soulève ici des questions liées au contexte qui doivent nécessairement être abordées compte tenu des intérêts concurrents dans chaque cas particulier. Il en résulte une évaluation contextuelle des intérêts concurrents.

[40] Lignes directrices, section IV, par. 2.8.

[41] Lignes directrices, section IV, par. 2.8.

[42] Voir aussi Convention modernisée pour la protection des personnes à l'égard du traitement automatisé des données à caractère personnel, rapport explicatif *(« Ces droits devraient également être respectés lors du développement et de l'utilisation de technologies innovantes. Cela permettra de renforcer la confiance dans l'innovation et les nouvelles technologies et partant, de continuer à favoriser leur développement »*).

[43] Voir aussi Convention 108 modernisée, art. 7.

II.3.2 Comités d'éthique

Eu égard aux applications à usage intensif de données, le comité d'éthique fait l'objet d'une attention grandissante dans les cercles d'IA, bien qu'il n'y ait pas de consensus unanime sur sa nature et sa fonction. Les études théoriques, les documents d'orientation et les initiatives des entreprises proposent autant de solutions différentes à cet égard.

La première différence d'approche qui se dégage a trait au niveau auquel ces comités devraient travailler [Polonetsky, Tene et Jerome, 2015 ; Calo, Ryan, 2013 ; Maison Blanche, 2015 ; IEEE, 2016]. Certaines propositions les décrivent comme des comités nationaux [Villani, 2018] ayant vocation à fournir des orientations générales sur des questions relatives au développement de l'IA[44]. Il ne s'agit pas d'une idée entièrement nouvelle : de telles structures ressemblent aux comités nationaux de bioéthique existants. Cependant, dans le cas des applications d'IA à usage intensif de données qui utilisent des renseignements personnels, les interactions entre ces comités nationaux et les autorités nationales de protection des données doivent être examinées soigneusement [Mantelero, 2016], de même que les interactions avec d'autres organismes nationaux tels que les autorités de concurrence ou les autorités nationales de sécurité. Beaucoup de pays disposent déjà d'organismes indépendants de surveillance qui supervisent des secteurs spécifiques dans lesquels des solutions d'IA sont opérationnelles ou pourraient être déployées. D'un point de vue réglementaire, il est donc important de collaborer avec ces autorités et de reconsidérer leur rôle, voire de renforcer leur coopération mutuelle [CEDP, 2016, p. 3 et 18 ; Conseil national du numérique, 2015, p. 74].

Une autre approche consisterait à mettre en place des comités d'éthique au niveau de l'entreprise, en soutien des responsables du traitement des données pour telle ou telle application, qui se focaliseraient sur les opérations effectuées par ces derniers. Ils pourraient assumer un rôle plus large et agir en tant que comités d'experts non seulement sur les questions éthiques, mais aussi sur un vaste éventail d'enjeux sociétaux liés à l'intelligence artificielle, comme l'application contextuelle des droits fondamentaux [Mantelero, 2018]. Plusieurs entreprises[45] ont déjà créé des comités internes ou externes chargés d'une mission de conseil sur les projets sensibles.

Cette deuxième solution reposant sur des comités d'éthique en entreprise pose moins de difficultés en termes de chevauchement avec les instances de régulation ou de surveillance existantes, mais pourrait nécessiter de définir plus clairement la relation entre ces comités et les autorités de contrôle. Le législateur national pourrait habiliter ces dernières à les soumettre à une surveillance étroite en cas de dysfonctionnements (pouvoirs et moyens insuffisants, faiblesse des décisions, etc.) affectant le traitement des données [Conseil national du numérique, 2015]. Comme pour d'autres types de comités consultatifs, la création de comités d'éthique pour l'intelligence artificielle soulève des questions concernant leur **indépendance**, leur statut, interne ou externe, et les bonnes pratiques à adopter pour éviter tout conflit d'intérêt.

---

[44] Voir aussi la consultation lancée au Royaume-Uni sur le nouveau centre dédié à l'éthique des données et à l'innovation, https://www.gov.uk/government/consultations/consultation-on-the-centre-for-data-ethics-and-innovation/centre-for-data-ethics-and-innovation-consultation.

[45] À cet égard, on ne peut que constater la propension croissante des entreprises de haute technologie ou exploitant des mégadonnées à mettre en place leurs propres comités d'éthique ou comités consultatifs. Voir par exemple Natasha Lomas, « DeepMind now has an AI ethics research unit. We have a few questions for it… », *TechCrunch*, 4 octobre 2017, http://social.techcrunch.com/2017/10/04/deepmind-now-has-an-ai-ethics-research-unit-we-have-a-few-questions-for-it/, consulté le 7 octobre 2017 ; Axon AI Ethics Board, https://it.axon.com/info/ai-ethics, consulté le 9 mai 2018 ; DNA Web Team, « Google drafting ethical guidelines to guide use of tech after employees protest defence project », *DNA India*, 15 avril 2018, http://www.dnaindia.com/technology/report-google-drafting-ethical-guidelines-to-guide-use-of-tech-after-employees-protest-defence-project-2605149, consulté le 7 mai 2018. Voir aussi Nations Unies, *Principes directeurs relatifs aux entreprises et aux droits de l'homme : mise en œuvre du cadre de référence « protéger, respecter et réparer » des Nations Unies*, Conseil des droits de l'homme des Nations Unies (HR/PUB/11/04), 2011.

La composition de ces comités dépendra aussi de la complexité des outils et applications d'IA. Lorsque les enjeux sociétaux sont considérables, une expertise juridique, éthique ou sociologique, outre des connaissances spécifiques au domaine concerné, est essentielle[46].

De tels comités peuvent jouer un rôle encore plus important dans les domaines où la transparence et la participation des acteurs concernés sont difficiles à atteindre, comme la justice prédictive, la détection des infractions ou la police prédictive.

Les comités d'éthique peuvent apporter un concours précieux pour aider les développeurs en intelligence artificielle à concevoir des algorithmes à vocation sociale, fondés sur les droits. En outre, un dialogue entre les développeurs et le comité[47] peut favoriser la création de procédures de traitement des données plus transparentes et permettre de définir plus clairement la logique qui les sous-tend (voir aussi le chapitre II.2).

II.3.3 Évaluation participative

Les experts (les comités d'éthique par exemple) peuvent jouer un rôle majeur en détectant les possibles effets pervers de l'intelligence artificielle et aider les responsables du traitement des données à aborder des questions sensibles. Dans certains cas, cependant, l'analyse est impossible sans la participation des communautés ou des groupes cibles.

Comme dans l'étude d'impact social, il apparaît utile, pour examiner les conséquences de l'IA sur la société, d'inciter le public à participer et de le mobiliser par la responsabilisation individuelle et collective, en garantissant, dans le cadre du processus d'évaluation, la non-discrimination et une égale participation à l'étude. Une approche participative[48] peut aussi contribuer à une meilleure compréhension des différents intérêts en présence et des valeurs éthiques et sociales[49].

---

[46] Lorsque le degré de complexité technique est moindre en termes de conséquences des applications d'IA, le comité pourrait être remplacé par un expert Éthique et Société qui aurait une mission analogue à celle du délégué à la protection des données (DPD). Il devrait aussi y avoir des critères obligatoires concernant la désignation et la qualité des membres du comité d'éthique. Leur nomination devrait être guidée par l'usage qui est fait des données et l'impact potentiel sur les droits fondamentaux, en retenant comme critères principaux les questions éthiques et sociétales. À cet égard, voir IEEE, 2016, p. 41-42 – Il est recommandé de *« créer une fonction aux plus hauts échelons de responsable marketing, éthicien ou juriste qui pourrait travailler pragmatiquement à la mise en place d'une conception intégrant une démarche éthique* (Ethically Aligned Design) *[…] Ce nouveau type de manager a un précédent : la fonction de « Chief Values Officer » imaginée par Kay Firth-Butterfield »* (CSER Cambridge, Kay Firth-Butterfield, Lucid AI's Ethics Advisory Panel, 2016, https://www.youtube.com/watch?v=w3-wYGbNZU4).

[47] Voir aussi Ministère britannique du Numérique, de la Culture, des Médias et des Sports, GOV.UK, Data Ethics Framework, « 3. Use data that is proportionate to the user need », https://www.gov.uk/guidance/3-use-data-that-is-proportionate-to-the-user-need, consulté le 4 juillet 2018.

[48] Le rôle des approches participatives et de la mobilisation des parties prenantes est expressément reconnu dans le contexte des droits fondamentaux [Institut danois des droits de l'homme, 2016, p. 24 ; Paul De Hert, « A Human Rights Perspective on Privacy and Data Protection Impact Assessments », *in* David Wright et Paul De Hert (dir.), *Privacy Impact Assessment*, Springer, Dordrecht, p. 72 – *« Une jurisprudence plus large est requise pour clarifier la portée du devoir d'étudier l'impact de certaines technologies et initiatives, y compris en dehors du champ de la santé environnementale. Quels que soient les termes employés, on peut assurément affirmer que le cadre actuel des droits de l'homme fait obligation aux États d'organiser de solides procédures de prise de décision associant les personnes affectées par les technologies »*].

[49] L'implication des diverses parties prenantes (par exemple participation de la société civile et des entreprises à la définition de lignes directrices sectorielles sur les valeurs) peut être plus efficace que la simple transparence, malgré l'accent mis sur la transparence dans les récents débats sur le traitement de données [Institut danois des droits de l'homme, 2016, p. 10 – *« L'action en direction des titulaires de droits et d'autres parties intéressées est essentielle dans les études d'impact en matière de droits de l'homme […] La mobilisation des parties prenantes a par conséquent été présentée comme une composante transversale essentielle »*]. Voir aussi Walker, 2009, p. 41 (*« La participation n'est pas seulement une fin – un droit – en soi, c'est aussi un moyen de donner aux populations la possibilité d'exercer une influence pour orienter les politiques et les projets qui les concernent ; cela renforce aussi la capacité des décideurs à prendre en compte les droits individuels et collectifs lors de la formulation et de la mise en œuvre de projets et politiques »*). Une forme de participation plus limitée, reposant sur la sensibilisation, a été suggérée par le Comité d'experts sur les intermédiaires d'internet (MSI-NET) du Conseil de l'Europe [MSI-NET, p. 47-48, 2018 – *« La sensibilisation de la population et le discours public sont*

La mobilisation des parties prenantes représente également un **objectif de développement de l'évaluation** [Haut-commissariat des Nations Unies aux droits de l'homme, 2006], dans la mesure où cela réduit le risque de sous-représentation de certains groupes et peut aussi mettre en évidence des aspects critiques qui auraient été sous-estimés ou ignorées par le responsable du traitement des données [Wright et Mordini, 2012, p. 402].

Néanmoins, les décideurs (ici les responsables du traitement) ne doivent pas voir la mobilisation des parties prenantes comme un moyen d'éluder leurs responsabilités en tant qu'intervenants sur l'ensemble du traitement [Palm et Hansson, 2006]. Ils doivent rester déterminés à tout mettre en œuvre pour atténuer autant que possible l'impact négatif du traitement des données sur l'individu et la société.

Enfin, une évaluation participative de l'impact à grande échelle des décisions algorithmiques [CNIL, 2017, p. 30] pourrait aussi pousser les responsables du traitement des données à adopter des **méthodes de conception conjointe** des applications d'IA, en associant activement les groupes potentiellement concernés à leur développement.

II.4 Responsabilité et vigilance

La question de la responsabilité autour des applications de l'intelligence artificielle reste ouverte pour plusieurs raisons. Comme pour le régime de la responsabilité du fait des produits défectueux, dont les principes axés sur la gestion du risque et de l'incertitude peuvent être largement étendus à l'IA, plusieurs modèles réglementaires (responsabilité objective, responsabilité fondée sur la faute, etc.) et stratégies (intervention de l'État, assurance obligatoire, etc.) sont applicables.

Une solution qui paraît intéressante serait d'étendre aux algorithmes la logique de la responsabilité du fait des produits défectueux, en attribuant l'entière responsabilité au fabriquant. Cela semble plus viable que la formule d'un délégué à la protection des données pour les algorithmes [CNIL, 2017, p. 56 – *« en identifiant au sein de chaque entreprise ou administration une équipe responsable du fonctionnement d'un algorithme dès lors que celui-ci traite les données de personnes physiques »*] ; en effet, étant donné l'omniprésence des applications intégrant l'IA, les différentes parties impliquées et le rôle de l'utilisateur, il est difficile de dissocier les différent aspects de la responsabilité du fait de l'IA.

Par ailleurs, une mise en cause de la responsabilité sert en quelque sorte à arrêter le système, ce qui est utile lorsque les diverses solutions *ex ante* (comme la transparence) n'ont rien donné [Principes d'Asilomar, 2017 – *« Transparence en cas de dommages : si une IA cause un dommage, il devrait être possible de savoir pourquoi »*]. Cependant, dans la mesure où la responsabilité civile est normalement réglementée par le législateur, point n'est besoin ici d'examiner les différentes solutions possibles[50].

Il convient toutefois de souligner à quel point gestion du risque, transparence et responsabilité se combinent non seulement lors de la conception des systèmes d'IA, mais encore au stade suivant, lors de l'usage des algorithmes [Access Now, 2018 ; ACM, 2018, 2.5]. Cela pourrait conduire les autorités de contrôle et les responsables du traitement des données à adopter diverses formes de **vigilance sur les algorithmes**, analogues à la pharmacovigilance, afin de réagir rapidement dans l'hypothèse où des effets imprévisibles et dangereux viendraient à se produire (comme les dérapages de Tay, le

---

d'une importance cruciale. Tous les moyens disponibles devraient être mis à profit pour informer et associer le grand public afin que les utilisateurs soient mis en mesure de comprendre de manière critique la logique et le fonctionnement des algorithmes et d'agir en conséquence. Cela peut inclure, sans y être limité, des campagnes d'information et d'éducation aux médias. Il conviendrait d'encourager les institutions ayant recours à des processus algorithmiques à fournir des explications facilement accessibles quant aux procédures suivies par les algorithmes et à la manière dont les décisions sont prises. Les sociétés qui développent les systèmes analytiques utilisés par les processus de prise de décision algorithmique et de collecte de données ont tout particulièrement le devoir de sensibiliser et d'informer les gens, y compris sur les risques de biais qui peuvent être induits par la conception et l'utilisation des algorithmes »].

[50] La responsabilité peut aussi prendre différentes formes dans les différents champs d'application de l'intelligence artificielle (propriété intellectuelle, prise de décision, voitures, etc.). Elle s'inscrit en effet dans un contexte bien spécifique.

« chatbot » de Microsoft[51]) [CNIL-LINC – Laboratoire d'innovation numérique de la Commission nationale de l'informatique et des libertés, 2017].

II. 5 Questions sectorielles

L'IA a des répercussions considérables sur de nombreux secteurs de notre société et de l'économie (police prédictive, justice, médecine de précision, marketing, propagande politique). Ses applications sectorielles soulèvent par conséquent différents enjeux qui ne sauraient être convenablement examinés ici, l'objet du présent rapport étant de donner un aperçu général des principaux aspects de l'interaction entre la protection des données et l'intelligence artificielle. Cette dernière partie se contente donc de mettre en lumière ces problématiques dans deux grands secteurs seulement : le secteur public et l'entreprise.

Les applications de l'intelligence artificielle soulèvent un certain nombre de questions spécifiques lorsqu'elles sont employées dans le secteur public [Reisman et al., 2018], en grande partie en raison du déséquilibre entre les pouvoirs de l'administration et ceux de l'administré et des services essentiels qui sont fournis. De plus, la complexité et l'opacité des solutions d'IA adoptées par l'État et ses agences font que ces organismes éprouvent davantage de difficultés à s'acquitter de leur obligation de rendre compte, et pas uniquement en matière de traitement des données [Reisman et al., 2018].

Pareil état de choses semble justifier l'adoption de garanties renforcées qui dépassent les attributions des comités ad hoc ou le champ du contrôle. Les garanties devraient aussi prévoir un processus d'évaluation qui permette de porter un regard critique sur les solutions d'IA proposées, afin de savoir si elles correspondent à un besoin réel et sont adaptées compte tenu de la nature de la prestation de services réalisée par les établissements publics ou par des sociétés privées chargées d'une mission de service public. Ce processus exige que les applications de l'intelligence artificielle puissent au minimum être examinées, testées et contrôlées par une autorité publique et soumises à des règles régissant la responsabilité [AI Now Institute, 2017].

Pour atteindre cet objectif, **les procédures de passation des marchés publics pourraient imposer des devoirs spécifiques de transparence et d'évaluation préalable aux fournisseurs d'intelligence artificielle**. Elles pourraient en outre aussi aborder des questions relatives au secret des affaires et à la protection de la propriété intellectuelle en introduisant des clauses contractuelles particulières destinées à accroître la transparence et à permettre l'audit de l'intelligence artificielle.

S'agissant des effets de l'IA sur le travail de demain (sans prendre en considération l'impact sur le marché du travail), les solutions de l'intelligence artificielle pourraient avoir une incidence sur les relations dans l'entreprise[52]. D'abord, elles peuvent renforcer le contrôle de l'employeur sur les salariés, dans un contexte souvent marqué par un rapport de force déséquilibré.

Ensuite, l'usage de formes de traitement des données non régulées, effectuées à l'insu de l'intéressé, peuvent transformer l'entreprise en un lieu d'expérimentation sociale *in vivo* qui soulève d'autres questions importantes concernant le rôle de la transparence, des comités d'éthique et de la participation volontaire au traitement des données.

Enfin, les appareils donnés aux salariés par l'employeur peuvent être à double usage. Par exemple, des dispositifs portables de bien-être peuvent être portés dans l'entreprise pour recueillir des données biologiques destinées à protéger la santé des salariés, mais ces derniers peuvent aussi les utiliser dans le cadre de leur pratique sportive pour surveiller leur condition physique. Il conviendrait de dûment examiner les répercussions de ce double usage sous l'angle de la protection des données et de la liberté individuelle. Autrement, cela pourrait brouiller les frontières entre vie professionnelle et vie privée [AI Now Institute, 2017, p. 10] et soulever des questions comme le problème de la surveillance généralisée ou le droit à la déconnexion.

[51] Vincent, James, « Twitter Taught Microsoft's Friendly AI Chatbot to Be a Racist Asshole in Less than a Day », *The Verge*, 24 mars 2016, https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist.
[52] Voir à ce propos CEDH, *Bărbulescu c. Roumanie*, arrêt du 5 septembre 2017, requête nº 61496/08 ; CEDH, *Libert c. France*, arrêt du 22 février 2018, requête nº 588/13.

Partie III – Lignes directrices

Les présentes lignes directrices fournissent un ensemble de mesures de référence que les gouvernements, les développeurs en intelligence artificielle, les fabricants et les prestataires de services devraient appliquer pour garantir la dignité humaine ainsi que les droits de l'homme et les libertés fondamentales de toute personne, notamment en ce qui concerne la protection des données à caractère personnel.

Rien dans les présentes lignes directrices ne saurait être interprété comme excluant ou limitant les dispositions de la Convention européenne des droits de l'homme et de la Convention 108 modifiée (« Convention 108+ »)[53].

I. Orientations générales

1. La responsabilité envers l'individu et la société est le corollaire de tout développement de l'intelligence artificielle, la sauvegarde des droits fondamentaux étant un préalable absolu.
2. Une perspective axée sur les droits fondamentaux doit présider au développement de l'intelligence artificielle et de ses applications, notamment lorsque l'IA est utilisée dans le contexte de processus décisionnels.
3. Un développement de l'IA reposant sur des données à caractère personnel doit être fondé sur les principes figurant dans la Convention 108 +. Les piliers de cette approche sont la proportionnalité du traitement, la responsabilité, la transparence et la gestion du risque.
4. La conscience du risque n'est pas une barrière à l'innovation ; bien au contraire, elle la favorise. Les risques de la « datafication » (mise en données du monde) et les implications potentiellement négatives des solutions fondées sur les données doivent dès lors être pris en considération.
5. Les particuliers et les communautés devraient avoir le droit de décider librement du rôle dévolu à l'IA, qu'il s'agisse d'analyser les comportements collectifs, d'exercer une influence sur les dynamiques sociales ou d'intervenir dans des processus décisionnels touchant des groupes ou des collectivités entières.
6. Conformément aux orientations sur l'évaluation des risques contenues dans les lignes directrices sur les mégadonnées (Big Data)[54], une vision plus large des éventuelles conséquences du traitement des données devrait être adoptée afin d'examiner l'incidence de l'usage des données non seulement sur les droits fondamentaux, mais aussi sur les valeurs éthiques et sociales collectives.
7. Le développement de l'intelligence artificielle et de ses applications ne saurait restreindre ou affecter négativement les droits des personnes concernées consacrés par la Convention 108 +.

II. Orientations à l'intention des développeurs en intelligence artificielle

1. Le Comité de la Convention 108 encourage activement les développeurs en intelligence artificielle à adopter une démarche de conception des produits et services centrée sur les valeurs, conformément à la Convention 108 + et aux autres instruments pertinents du Conseil de l'Europe.
2. Les développeurs en intelligence artificielle doivent évaluer les conséquences négatives des applications d'IA sur les droits et libertés fondamentaux des personnes concernées. Au regard de ces conséquences, il convient d'adopter une approche de précaution fondée sur des politiques de prévention des risques.

---

[53] Protocole d'amendement (STCE n° 123) à la Convention pour la protection des personnes à l'égard du traitement automatisé des données à caractère personnel.

[54] Lignes directrices sur les mégadonnées adoptées par le Comité de la Convention 108 en janvier 2017, disponibles à l'adresse https://rm.coe.int/CoERMPublicCommnSearchServices/DisplayDCTMContent?documentId=09000016806ebe7a.

3. Les développeurs en intelligence artificielle doivent adopter des solutions dès la conception (*by-design*) pour éviter tout biais caché potentiel et non intentionnel, de même que tout risque de discrimination ou d'impact négatif sur les droits et libertés fondamentales des personnes concernées, à tous les stades du traitement des données, lors de la collecte comme de l'analyse.
4. Lors de la conception des applications d'IA, il est important d'adopter un paradigme de conception qui évalue de manière critique la nature et la quantité des données utilisées, en réduisant les données redondantes ou marginales et en commençant par un volume limité de données d'apprentissage, puis en vérifiant l'exactitude du modèle lorsqu'il est alimenté par de nouvelles données. Le recours à des données synthétiques peut être considéré comme l'une des solutions possibles pour minimiser les données personnelles traitées.
5. Les risques inhérents aux données décontextualisées (c.-à-d. ignorant l'information contextuelle propre à la situation spécifique dans laquelle les solutions d'IA proposées devraient être appliquées) et aux modèles algorithmiques décontextualisés (c.-à-d. utilisant des modèles d'IA initialement conçus pour un autre contexte ou pour des finalités différentes) devraient être dûment pris en compte lors du développement d'applications intégrant l'intelligence artificielle.
6. Des comités d'experts issus de différents domaines, ainsi que d'institutions universitaires indépendantes, devraient être associés au développement de l'IA. Ils peuvent apporter un concours précieux pour aider à concevoir des algorithmes à vocation sociale, fondés sur les droits, et contribuer à détecter des biais potentiels. Leur rôle est encore plus important dans les domaines où la transparence et la mobilisation des parties prenantes sont difficiles à atteindre, comme par exemple des outils d'IA destinés à être utilisés dans un contexte judiciaire ou répressif.
7. Des démarches participatives d'évaluation des risques, reposant sur l'engagement actif des groupes potentiellement affectés par les applications de l'intelligence artificielle, devraient être adoptées.
8. Lorsque cela est techniquement faisable, les développeurs en intelligence artificielle devraient concevoir leurs produits et services de manière à préserver la liberté de choix de l'utilisateur concernant l'usage de l'intelligence artificielle, et fournir une alternative à l'usage de services et dispositifs dotés d'une intelligence artificielle.
9. Les personnes concernées ont le droit de savoir quelles applications d'IA sont utilisées et d'avoir connaissance du raisonnement qui sous-tend les opérations de traitement des données, y compris les conséquences de ce raisonnement.

III. Orientations à l'intention des décideurs

1. Les procédures de passation des marchés publics pourraient imposer des devoirs spécifiques de transparence et d'évaluation préalable des systèmes d'IA aux prestataires de service.
2. Un renforcement de la responsabilité des développeurs en intelligence artificielle et l'adoption de procédures d'évaluation des risques pourraient considérablement accroître la confiance du public dans les produits et services intégrant l'IA.
3. Les autorités de contrôle de la protection des données et les responsables du traitement devraient adopter diverses formes de vigilance sur les algorithmes en vue de mieux assurer la conformité avec les obligations en matière de protection des données et les principes relatifs aux droits de l'homme pendant toute la durée de vie des applications de l'intelligence artificielle.
4. Une confiance excessive dans la fiabilité des solutions fournies par les systèmes d'IA, de même que la crainte de voir sa responsabilité mise en cause si la décision prise est autre que celle suggérée par l'IA, risque d'altérer l'autonomie de l'intervention humaine dans la prise de décision. Il est dès lors crucial que le décideur (personne physique) ait la liberté de ne pas se fonder sur le résultat des recommandations découlant de l'utilisation de l'intelligence artificielle.
5. Lorsque des applications d'IA sont susceptibles d'avoir un impact important sur les droits et libertés fondamentales des personnes concernées, les responsables du traitement devraient

consulter les autorités de contrôle afin de chercher à obtenir des conseils visant à réduire ces effets négatifs potentiels.

6. Les pays ayant mis en place des organismes indépendants de surveillance qui supervisent des secteurs spécifiques dans lesquels des solutions d'IA sont opérationnelles ou pourraient être déployées devraient renforcer la coopération mutuelle entre ces organismes et leur coopération avec les autorités de contrôle de la protection des données.
7. Lorsque des comités d'experts sont créés au niveau de l'entreprise, les autorités de contrôle devraient être habilités à les soumettre à une surveillance étroite en cas de dysfonctionnements (manque d'indépendance, pouvoirs et moyens insuffisants, faiblesse des décisions, etc.) affectant le traitement des données.

Références

- Access Now, *The Toronto Declaration: Protecting the rights to equality and non-discrimination in machine learning systems*, 2018, https://www.accessnow.org/the-toronto-declaration-protecting-the-rights-to-equality-and-non-discrimination-inmachine-learning-systems/.
- ACM [Association for Computing Machinary], *ACM Code of Ethics and Professional Conduct*, 2018, https://www.acm.org/code-of-ethics.
- Agence des droits fondamentaux de l'Union européenne (FRA), *#BigData: Discrimination in Data-Supported Decision Making*, 2018, http://fra.europa.eu/en/publication/2018/big-data-discrimination.
- AI Now Institute, *The AI Now Report. The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term*, 2016, https://ainowinstitute.org/AI_Now_2016_Report.pdf.
- AI Now Institute, *AI Now 2017 Report*, 2017, https://assets.contentful.com/8wprhhvnpfc0/1A9c3ZTCZa2KEYM64Wsc2a/8636557c5fb14f2b74b2be64c3ce0c78/_AI_Now_Institute_2017_Report_.pdf, consulté le 26 octobre 2017.
- AI Now Institute, *Litigating Algorithms: Challenging Government Use of Algorithmic Decision Systems*, 2018, https://ainowinstitute.org/litigatingalgorithms.pdf.
- Ananny, M. et Crawford, K., « Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability », *New Media & Society*, 2016, https://doi.org/10.1177/1461444816676645.
- Artificial Intelligence Index, *Annual Report 2017*, http://aiindex.org/2017-report.pdf, consulté le 5 décembre 2017.
- Autorité norvégienne de protection des données, *Artificial Intelligence and Privacy Report*, 2018, https://www.datatilsynet.no/globalassets/global/english/ai-and-privacy.pdf.
- Axon AI Ethics Board, https://it.axon.com/info/ai-ethics.
- Barocas, S. et Nissenbaum, H., « Big Data's End Run around Anonymity and Consent », *in* Lane, J., Stodden, V., Bender, S. et Nissenbaum, H. (dir.), *Privacy, big data, and the public good: frameworks for engagement*, Cambridge University Press, 2015.
- Barocas, S. et Selbstr, A. D., *Big Data's Disparate Impact*, vol. 104, nº 3, *California Law Review*, 2016, p. 671-732.
- Barse, E. L., Kvarnstrom, H. et Jonsson, E., « Synthesizing Test Data for Fraud Detection Systems », in *19th Annual Computer Security Applications Conference, 2003. Proceedings*, 2003, p. 384-394, https://doi.org/10.1109/CSAC.2003.1254343.
- Binns, R. et al., « "It's Reducing a Human Being to a Percentage"; Perceptions of Justice in Algorithmic Decisions », ArXiv:1801.10408 [Cs], 2018, p. 1-14. https://doi.org/10.1145/3173574.3173951.
- Bostrom, N., *Superintelligence paths, dangers, strategies*, Oxford University Press, Oxford, 2016.
- Boyd, D. et Crawford, K., « Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon », *Information, Communication & Society*, vol. 15, nº 5, 2012, p. 662-679.
- Brauneis, R., et Goodman, E.P., « Algorithmic Transparency for the Smart City », *Yale J. L. & Tech.*, vol. 20, 2018, p. 103-176.
- Bray, P. et al., *International differences in ethical standards and in the interpretation of legal frameworks. SATORI Deliverable D3.2*, 2015, http://satoriproject.eu/work_packages/legal-aspects-and-impacts-of-globalization/.
- Brundage, M. et al., *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*, p. 56 et 93, février 2018, https://maliciousaireport.com/.
- Bureau exécutif du président des États-Unis et Conseil national de la science et de la technologie – Comité sur la technologie, *Preparing for the Future of Artificial Intelligence* (Washington D.C.), 2016, https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf.
- Burrell, J., « How the machine 'thinks': Understanding opacity in machine learning algorithms », *Big Data & Society*, vol. 3, nº 1, 2016, https://doi.org/10.1177/2053951715622512.

- Burt, A., Leong, B. et Shirrell, S., *Beyond Explainability: A Practical Guide to Managing Risk in Machine Learning Models*, Future of Privacy Forum, 2018.
- Calo, Ryan, « Consumer Subject Review Boards: A Thought Experiment », Stan. L. Rev. Online, vol. 66, 2013, p. 97, http://www.stanfordlawreview.org/online/privacy-and-big-data/consumer-subject-review-boards, consulté le 23 février 2018.
- Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., Elhadad, N., « Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission », in *Proceedings of the 21st Annual SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, p. 1721-1730.
- CEDP (Le contrôleur européen de la protection des données), Groupe consultatif sur l'éthique, *Towards a digital ethics*, 2018, https://edps.europa.eu/sites/edp/files/publication/18-01-25_eag_report_en.pdf.
- CEDP (Le contrôleur européen de la protection des données), avis 8/2016, *Avis du CEDP sur une application cohérente des droits fondamentaux à l'ère des données massives* (Big Data), 2016.
- Citron, D. K. et Pasquale, F., « The Scored Society: Due Process For Automated Predictions », *Washington Law Review*, vol. 89, 2014, p. 1-33.
- CNIL-LINC, *La plateforme d'une ville. Les données personnelles au cœur de la fabrique de la smart city*, 2017, https://www.cnil.fr/sites/default/files/atoms/files/cnil_cahiers_ip5.pdf.
- CNIL, *Comment permettre à l'homme de garder la main ? Les enjeux éthiques des algorithmes et de l'intelligence artificielle*, synthèse du débat public animé par la CNIL dans le cadre de la mission de réflexion éthique confiée par la loi pour une république numérique, décembre 2017, p. 14, https://www.cnil.fr/sites/default/files/atoms/files/cnil_rapport_garder_la_main_web.pdf.
- CNIL, *Comment permettre à l'homme de garder la main ? Les enjeux éthiques des algorithmes et de l'intelligence artificielle*, synthèse du débat public animé par la CNIL dans le cadre de la mission de réflexion éthique confiée par la loi pour une république numérique, CNIL, 2017, https://www.cnil.fr/sites/default/files/atoms/files/cnil_rapport_garder_la_main_web.pdf.
- Commissariat à la protection de la vie privée du Canada, *L'Internet des objets. Introduction aux enjeux relatifs à la protection de la vie privée dans le commerce de détail et à la maison*, 2016, https://www.priv.gc.ca/media/1809/iot_201602_f.pdf.
- Commission européenne, Groupe européen d'éthique des sciences et des nouvelles technologies, *Déclaration sur L'intelligence artificielle, la robotique et les systèmes « autonomes »*, 2018, https://ec.europa.eu/research/ege/pdf/ege_ai_statement_2018_fr.pdf.
- Commission européenne, *The European Artificial Intelligence Landscape*, 2018, https://ec.europa.eu/digital-single-market/en/news/european-artificial-intelligence-landscape.
- Comité d'experts du Conseil de l'Europe sur les intermédiaires internet (MSI-NET), *Étude sur les dimensions des droits humains dans les techniques de traitement automatisé des données (en particulier les algorithmes) et éventuelles implications réglementaires*, 2018, https://rm.coe.int/algorithms-and-human-rights-en-rev/16807956b5. https://rm.coe.int/etude-sur-les-algorithmes-version-finale/1680770cc2.
- Comité économique et social européen, *L'éthique des mégadonnées (Big Data). Équilibrer les avantages économiques et les questions d'éthique liées aux données massives dans le contexte des politiques européennes*, 2017, https://www.eesc.europa.eu/fr/our-work/publications-other-work/publications/lethique-des-megadonnees-big-data#downloads.
- Conseil de l'Europe, *Lignes directrices sur la protection des personnes à l'égard du traitement des données à caractère personnel à l'ère des mégadonnées*, 2017, https://rm.coe.int/lignes-directrices-sur-la-protection-des-personnes-a-l-egard-du-traite/16806f06d1.
- Conseil national du numérique, *Ambition numérique : Pour une politique française et européenne de la transition numérique*, 2015, http://www.cil.cnrs.fr/CIL/IMG/pdf/CNNum--rapport-ambition-numerique.pdf.
- Cummings, M. L., Roff, H. M., Cukier, K., Parakilas, J. et Bryce H., « Artificial Intelligence and International Affairs Disruption Anticipated », *Chatham House Report*, Chatham House, The Royal Institute of International Affairs, Londres, 2018,

https://www.chathamhouse.org/sites/default/files/publications/research/2018-06-14-artificial-intelligence-international-affairs-cummings-roff-cukier-parakilas-bryce.pdf.
- Diakopoulos, N., *Algorithmic Accountability Reporting: on the Investigation of Black Boxes*, Tow Center for Digital Journalism, 2013.
- DNA Web Team, « Google drafting ethical guidelines to guide use of tech after employees protest defence project' DNA India », 15 avril 2018, http://www.dnaindia.com/technology/report-google-drafting-ethical-guidelines-to-guide-use-of-tech-after-employees-protest-defence-project-2605149.
- Donovan, J., Matthews, J., Caplan, R. et Hanson, L., *Algorithmic Accountability: A Primer*, 2018, https://datasociety.net/output/algorithmic-accountability-a-primer/.
- Edwards, L. et Vale, M., « Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For », *Duke Law and Technology Review*, vol. 16, nº 1, 2017, p. 18-84.
- Forum économique mondial, *How to Prevent Discriminatory Outcomes in Machine Learning*, 2018, http://www3.weforum.org/docs/WEF_40065_White_Paper_How_to_Prevent_Discriminatory_Outcomes_in_Machine_Learning.pdf.
- Gama, J. et al., *A survey on concept drift adaptation*, ACM Computing Surveys, vol. 1, nº 1, 2013, http://www.win.tue.nl/~mpechen/publications/pubs/Gama_ACMCS_AdaptationCD_accepted.pdf.
- Goodman, B. et Flaxman, S., *European Union regulations on algorithmic decision-making and a "right to explanation"*, arXiv:1606.08813 [cs, stat], 2016, http://arxiv.org/abs/1606.08813.
- Haut-Commissariat des Nations Unies aux droits de l'homme, *Frequently asked questions on a human rights-based approach to development cooperation*, Nations Unies, New York et Genève, 2006.
- Hildebrandt, M., *Smart Technologies and the End(s) of Law: Novel Entanglements of Law and Technology*, Edward Elgar Publishing, 2016.
- IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems, *Ethically Aligned Design: A Vision For Prioritizing Wellbeing With Artificial Intelligence And Autonomous Systems*, version 1, IEEE, 2016, https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v1.pdf.
- Information Commissioner's Office (ICO – Bureau du Commissaire à l'information, homologue britannique de la CNIL française), *Big Data, Artificial Intelligence, Machine Learning and Data Protection*, 2017, https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf.
- Institut danois des droits de l'homme, *Human rights impact assessment guidance and toolbox*, Institut danois des droits de l'homme, 2016, https://www.humanrights.dk/business/tools/human-rights-impact-assessment-guidance-and-toolbox.
- Kaye, J. et al., « Dynamic consent: a patient interface for twenty-first century research networks », *European Journal of Human Genetics*, vol. 23, nº 2, 2015, p. 141.
- Kurzweil, R., *The singularity is near : when humans transcend biology*, Duckworth, Londres, 2016.
- Linnet, T., Floridi, L. et van der Sloot, B. (dir.), *Group Privacy: New Challenges of Data Technologies*, Springer International Publishing, 2017.
- Lipton, Z. C., « The Mythos of Model Interpretability. In Machine Learning, the Concept of Interpretability Is Both Important and Slippery », *ACM Queue*, vol. 16, nº 3, 2018, https://queue.acm.org/detail.cfm?id=3241340.
- Lomas, N., 'DeepMind now has an AI ethics research unit. We have a few questions for it…' TechCrunch (4 octobre 2017) http://social.techcrunch.com/2017/10/04/deepmind-now-has-an-ai-ethics-research-unit-we-have-a-few-questions-for-it/, consulté le 3 mai 2018.
- Lycett, M., « Datafication: making sense of (big) data in a complex world », *European Journal of Information Systems*, vol. 22, nº 4, 2013, p. 381-386.

- Maison Blanche, *Administration Discussion Draft: Consumer Privacy Bill of Rights Act of 2015*, sec. 103*c*, 2015, https://www.whitehouse.gov/sites/default/files/omb/legislative/letters/cpbr-act-of-2015-discussion-draft.pdf.
- Mantelero A., « AI and Big Data: A blueprint for a human rights, social and ethical impact assessment », *Computer Law & Security Review*, 2018, https://doi.org/10.1016/j.clsr.2018.05.017.
- Mantelero, A., « The future of consumer data protection in the E.U. Rethinking the "notice and consent" paradigm in the new era of predictive analytics », *Computer Law and Security Review*, vol. 30, nº 6, 2014, p. 643-660.
- Mantelero, A., « Regulating Big Data. The guidelines of the Council of Europe in the Context of the European Data Protection Framework », *Computer Law & Security Review*, vol. 33, nº 5, 2017, p. 584-602.
- Mayer-Schönberger, V. et Cukier, K., *Big Data. A Revolution That Will Transform How We Live, Work and Think*, John Murray, Londres, 2013.
- McCulloch, W.S. et Pitts, W.H., « A Logical Calculus of the Ideas Immanent in Nervous Activity », *Bulletin of Mathematical Biophysics*, vol. 5, p. 115-133, 1943.
- Ministère fédéral des Transports et des Infrastructures numériques (BMVI), Commission de l'éthique, *Automated and Connected Driving*, http://www.bmvi.de/SharedDocs/EN/publications/report-ethics-commission.pdf?__blob=publicationFile, consulté le 17 juillet 2018.
- Nations Unies, *Principes directeurs relatifs aux entreprises et aux droits de l'homme : mise en œuvre du cadre de référence « protéger, respecter et réparer » des Nations Unies*, Conseil des droits de l'homme des Nations Unies (HR/PUB/11/04), 2011.
- Omer, T. et Polonetsky, J., « Privacy in the Age of Big Data. A Time for Big Decisions », *64 Stan. L. Rev. Online*, 2012, p. 63-69.
- O'Neil, C., *Weapons of math destruction*, Penguin Books, Londres, 2017.
- Palm, E. et Hansson, S. O., « The case for ethical technology assessment (eTA) », *Technological Forecasting & Social Change*, vol. 73, nº 5, 2006, p. 543 et p. 550-551.
- Parlement européen, Résolution du Parlement européen du 14 mars 2017 sur les incidences des mégadonnées pour les droits fondamentaux : respect de la vie privée, protection des données, non-discrimination, sécurité et application de la loi (2016/2225(INI)) http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+TA+P8-TA-2017-0076+0+DOC+XML+V0//FR.
- Principes d'Asilomar, 2017, https://futureoflife.org/ai-principles/.
- Reisman, D., Schultz, J., Crawford, K. et Whittaker, M., *Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability*, 2018, https://ainowinstitute.org/aiareport2018.pdf.
- Rossi, F., *Artificial Intelligence: Potential Benefits and Ethical Considerations*, 2016 (Parlement européen, Département thématique C « Droit des citoyens et affaires constitutionnelles », 2016), note d'information PE 571.380, http://www.europarl.europa.eu/RegData/etudes/BRIE/2016/571380/IPOL_BRI(2016)571380_EN.pdf.
- Rouvroy, A., « *Des donnés et des hommes » Droits et libertés fondamentaux dans un monde de données massives,* 2016, https://rm.coe.int/16806b1659.
- Rubinstein, I.S., « Big Data: The End of Privacy or a New Beginning? », *International Data Privacy Law*, vol. 3, nº 2, 2013, p. 74-87.
- Selbst, A. D., « Disparate Impact in Big Data Policing », *Georgia Law Review*, vol. 52, nº 1, 2017, p. 109-195 ; Selbst, Andrew D. et Powles, Julia, « Meaningful Information and the Right to Explanation », *International Data Privacy Law*, vol. 7, nº 4, 2017, p. 233-242.
- Sheehan, M., « Can Broad Consent be Informed Consent? », *Public Health Ethics*, vol. 3, 2011, p. 226-235.
- Speikermann, S., *Ethical IT Innovation. A Value-Based System Design Approach*, CRC Press, Boca Raton, 2016.
- Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R., *Intriguing properties of neural networks*, 2013, https://arxiv.org/abs/1312.6199.

- Turing, A. M., « Computing Machinery and Intelligence », *Mind*, vol. 59, 1950, p. 433–460, Ministère britannique du Numérique, de la Culture, des Medias et des Sports, 'Data Ethics Framework - GOV.UK', https://www.gov.uk/government/publications/data-ethics-framework/data-ethics-framework.
- UIT, rapport de la première édition du Sommet mondial de l'UIT sur l'intelligence artificielle au service du bien social, 2017, https://www.itu.int/en/ITU-T/AI/Documents/Report/AI_for_Good_Global_Summit_Report_2017.pdf.
- Veale M., Binns R., « Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data », *Big Data & Society*, vol. 4, nº 2, 2053951717743530, 2017, https://doi.org/10.1177/2053951717743530.
- Veale, M., Binns, R. et Edwards, L., « Algorithms That Remember: Model Inversion Attacks and Data Protection Law », *Philosophical Transactions of the Royal Society*, à paraître en 2018, https://doi.org/10.1098/rsta.2018.0083.
- Villani, C., *Donner un sens à l'intelligence artificielle. Pour une stratégie nationale et européenne*,https://www.aiforhumanity.fr/pdfs/9782111457089_Rapport_Villani_accessible.pdf, 2018.
- Wachter, S., Mittelstadt, B. et Floridi, L., « Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation », *International Data Privacy Law*, vol. 7, nº 2, 2017, p. 76-99.
- Walker, S.M., *The Future of Human Rights Impact Assessments of Trade Agreements*, G. J. Wiarda Institute for Legal Research, Utrecht, 2009, https://dspace.library.uu.nl/bitstream/handle/1874/36620/walker.pdf?sequence=2.
- Wright, D. et Mordini, E., « Privacy and Ethical Impact Assessment », *in* Wright, D. et De Hert, P. (dir.), *Privacy Impact Assessment*, Springer, Dordrecht, 2012, p. 397-418.
- Wright, D. et De Hert, P. (dir.), *Privacy Impact Assessment*, Springer, Dordrecht, 2012.
- Wright, D., 2011, « A framework for the ethical impact assessment of information technology », *Ethics and Information Technology*, vol. 13, nº 199, p. 201-202.