Responsabilité et IA

Étude du Conseil de l'Europe
DGI(2019)05

Rapporteur: Karen Yeung

Préparée par le Comité d'experts sur les dimensions des droits de l'homme dans le traitement des données et les différentes formes d'intelligence artificielle (MSI-AUT)

1949.2019
COUNCIL OF EUROPE
CONSEIL DE L'EUROPE

DGI(2019)05

Étude sur les incidences des technologies numériques avancées (dont l'intelligence artificielle) sur la notion de responsabilité, sous l'angle des droits humains

Préparée par le Comité d'experts sur les dimensions des droits de l'homme dans le traitement automatisé des données et les différentes formes d'intelligence artificielle (MSI-AUT)

Rapporteuse : Karen Yeung

Edition anglaise :
Responsibility and AI

Couverture : Service de la production des documents et des publications (SPDP), Conseil de l'Europe
Photo: Shutterstock

Cette publication n'a pas fait l'objet d'une relecture typographique et grammaticale de l'Unité éditoriale du SPDP.

Imprimé dans les ateliers du Conseil de l'Europe

TABLE DES MATIÈRES

Introduction 4

Résumé 6

Chapitre 1. Introduction 16

1.1 Champ de l'étude 16

1.2 Structure de l'étude 17

1.3 Incidences de l'IA sur la notion de responsabilité 18

1.4 Incidences envisagées sous l'angle des droits de l'homme 26

Chapitre 2. Risques potentiels et avérés, dommages et autres effets négatifs des technologies numériques avancées 28

2.1 La montée des systèmes de prise de décision algorithmique (ADM) 28

2.1.1 Quels sont les droits compromis par les systèmes d'ADM ? 29

2.1.2 Problèmes sociétaux associés au profilage par les données 33

2.2 Menaces et risques sociétaux collectifs générés par d'autres technologies d'IA 38

2.2.1 Attaques malveillantes, conception non éthique du système ou défaillance involontaire du système 39

2.2.2 Perte d'un rapport humain authentique, réel et significatif 39

2.2.3 Effet dissuasif de la réutilisation des données 40

2.2.4 Exercice irresponsable du pouvoir conféré par le numérique 40

2.2.5 La privatisation cachée de décisions relatives aux intérêts public 41

2.2.6 Exploitation de main-d'œuvre humaine pour entraîner les algorithmes 42

2.3 Asymétrie des pouvoirs et menaces pour les fondements socio-techniques de la communauté morale et démocratique 42

2.4 Résumé 44

Chapitre 3. Qui est responsable des menaces, des risques, des préjudices et des torts causés par les technologies numériques avancées ? 45

3.1 Qu'est-ce que la responsabilité et en quoi est-elle importante ? 46

3.2 Les différents aspects de la responsabilité 49

3.3 Liens entre technologies numériques avancées (dont l'IA) et conceptions actuelles de la responsabilité 50

3.3.1 Responsabilité prospective : codes d'éthique volontaires et « robotique responsable » 52

3.3.2 Machines autonomes et « problème de contrôle » 54

3.4 Modèles d'attribution de la responsabilité 57

3.4.1 Modèles fondés sur l'intention/la culpabilité 59

3.4.2 Modèles fondés sur le risque/la négligence 60
3.4.3 Responsabilité absolue 62
3.4.4 Régime d'assurance obligatoire 63
3.5 Défis liés à la complexité des systèmes sociotechniques 64
3.5.1 Le problème des acteurs multiples 65
3.5.2 L'interaction humain-machine 66
3.5.3 Interactions imprévisibles et changeantes entre systèmes sociotechniques complexes 69
3.6 Obligation de l'État d'assurer une protection effective des droits de l'homme 70
3.7 Rôle des mécanismes extrajudiciaires 71
3.7.1 Dispositifs de protection techniques 71
3.7.2 Instruments et techniques de régulation 72
3.7.3 Définition, suivi et application de normes 74
3.8 Renouveler le discours sur les droits de l'homme à l'heure des réseaux numériques 75
3.9 Résumé 78
Chapitre 4. Conclusion 80
Annexe A 83
Références 87

Introduction

Dans le cadre de son mandat biennal 2018-2019 du Comité directeur sur les médias et la société de l'information (CDMSI), le Comité des Ministres du Conseil de l'Europe a demandé au CDMSI de réaliser « une étude sur le développement et l'utilisation des nouveaux services et technologies numériques, comme les différentes formes d'intelligence artificielle, dans la mesure où ils peuvent affecter la jouissance des droits et des libertés fondamentales à l'époque numérique en vue d'offrir des orientations pour un futur instrument normatif dans le domaine » ; et ce par le biais du Comité d'experts sur les dimensions des droits de l'homme dans le traitement automatisé des données et les différentes formes d'intelligence artificielle (MSI-AUT), structure subordonnée nommée par le Comité des Ministres pour faciliter le travail du CDMSI.
Lors de sa première réunion, les 6 et 7 mars 2018, le comité d'experts a décidé de concentrer son étude sur les incidences des systèmes de prise de décision basé sur l'IA sur la notion de responsabilité, à travers le prisme des droits humains. Ils ont alors designé Prof. Karen Yeung comme rapporteuse pour la préparation de l'étude.

Composition du comité d'experts MSI-AUT

Abraham BERNSTEIN, Professeur d'informatique, Université de Zürich

Jorge CANCIO, Spécialiste en relations internationales, Office fédéral de la communication, Suisse

Luciano FLORIDI, Professeur de philosophie et d'éthique de l'information, Université d'Oxford

Seda GÜRSES, Professeure Assistante, Université Technique Delft

Gabrielle GUILLEMIN, Juriste principale, ARTICLE 19

Natali HELBERGER, Professeure de droit de l'information, Université d'Amsterdam

Luukas ILVES (Président), Directeur adjoint et maître de recherche, Conseil de Lisbonne

Tanja KERŠEVAN SMOKVINA, Secrétaire d'État, Ministère de la Culture, Slovénie

Joe MCNAMEE, Consultant Indépendant

Evgenios NASTOS, Chef de l'Unité de l'information, Ministère de la politique numérique, des télécommunications & médias, Grèce

Pierluigi PERRI, Professeur de droit de l'informatique, Université de Milan

Wolfgang SCHULZ (Vice-Président), Professeur de droit, Université de Hamburg

Karen YEUNG, Professeure agrégée interdisciplinaire en droit, éthique et informatique, Université de Birmingham

Résumé

Les technologies et services numériques avancés, dont l'intelligence artificielle (« IA ») appliquée à des tâches spécifiques, recèlent un potentiel extraordinaire. Ils ont déjà suscité des progrès remarquables, notamment en rendant de nombreux services numériques plus efficaces, précis, rapides et faciles à utiliser.

Cependant, l'émergence de ces technologies s'est accompagnée d'une montée des inquiétudes sur leurs éventuels effets indésirables pour les individus, les populations fragiles et l'ensemble de la société. Pour que ces technologies marquent un progrès, et non un recul, et favorisent l'épanouissement des individus et de la société, nous devons absolument chercher à mieux comprendre ces inquiétudes. Cela suppose non seulement de mieux appréhender l'impact des technologies sur l'exercice des droits de l'homme et des libertés fondamentales, mais aussi d'étudier de près la question des responsabilités en cas de conséquences néfastes.

La présente étude repose sur une **prémisse** : dans les régimes constitutionnels et démocratiques d'aujourd'hui, la définition du concept de responsabilité et la manière dont les sociétés l'appliquent et l'incarnent dans des institutions revêtent une importance cruciale. C'est la condition nécessaire pour que les individus et les organisations répondent des dommages qu'ils causent à autrui, et pour créer et entretenir au sein de la société les bases d'une coopération fiable et pacifique.

De ce fait, l'**objectif** de cette étude est d'examiner dans un premier temps les incidences des technologies numériques avancées (dont l'IA) sur la notion de responsabilité, en particulier là où ces technologies pourraient entraver l'exercice des droits de l'homme et des libertés fondamentales protégés par la CEDH, et dans un second temps les modalités d'attribution des responsabilités associées à leurs risques et à leurs conséquences.

L'**approche méthodologique** adoptée est interdisciplinaire, puisant dans des concepts et des connaissances issus du droit, de la philosophie, des sciences sociales et – dans une moindre mesure – de l'informatique. L'étude conclut qu'à notre époque numérique et interconnectée, si nous tenons aux droits de l'homme, nous ne pouvons laisser les systèmes numériques avancés et leurs développeurs acquérir de plus en plus de pouvoir en l'absence de toute responsabilité. Or, le devoir de protéger les droits de l'homme appartient en premier lieu aux États. Ils doivent donc s'assurer que ceux qui œuvrent à la conception, au développement et au déploiement de ces technologies et en tirent des bénéfices répondent aussi de leurs impacts négatifs. Pour cela, les États doivent, entre autres, assurer l'existence de mécanismes institutionnels effectifs et légitimes destinés à *prévenir et à stopper* les atteintes aux droits de l'homme que ces technologies risquent d'entraîner, et plus généralement préserver la *santé de l'environnement sociotechnique* collectif et partagé qui sert d'assise aux droits de l'homme et à l'État de droit. Les grandes lignes de l'étude sont brièvement résumées ci-dessous.

Chapitre 1 – Introduction

Le chapitre 1 explique **ce qu'est l'intelligence artificielle (« IA ») et avec quel degré de spécialisation les technologies d'IA fonctionnent**. Il présente l'IA comme un ensemble de technologies avancées à usage général, utilisant des techniques issues de la statistique, de l'informatique et de la psychologie cognitive pour permettre à des machines d'exécuter efficacement des tâches très complexes. Ces technologies visent soit à reproduire, soit à surpasser des aptitudes qui supposeraient de l'« intelligence » chez les êtres humains : raisonnement, autonomie, créativité, etc. Les technologies d'IA utilisent l'apprentissage

automatique, qui consiste à alimenter des systèmes informatiques en exemples, en données et en expérience jusqu'à les rendre capables d'accomplir intelligemment des tâches spécifiques. Or, les technologies d'apprentissage automatique soulèvent des questions de responsabilité, puisqu'elles rendent possible l'automatisation des tâches et permettent aux machines de prendre des décisions et d'accomplir des actions en s'affranchissant dans une certaine mesure de leurs développeurs.

Le chapitre attire l'attention sur la capacité des systèmes d'apprentissage automatique à apprendre et à évoluer au fil du temps, en se fixant leurs propres sous-objectifs et en s'adaptant aux conditions locales grâce aux informations transmises par leurs capteurs ou en assimilant des données mises à jour. Les concepteurs de ces systèmes ne déterminent et règlent que les paramètres initiaux et l'objectif global que le système doit atteindre de manière optimale. Les systèmes d'apprentissage automatique prennent ensuite des décisions de manière indépendante, en optant pour la meilleure alternative selon des modalités non programmées à l'avance et sans aucune intervention humaine. En continu et par itération, ces systèmes tirent des enseignements de leur environnement, lui-même souvent changeant et désordonné – ce qui peut rendre leur fonctionnement instable et imprévisible. En particulier, ces systèmes sont susceptibles d'évoluer de manière inattendue (chapitre 1.3).

Le chapitre 1 explique ensuite que dans le contexte de notre infrastructure mondiale de données, les technologies d'IA présentent une série d'autres propriétés qui se répercutent directement sur le concept de responsabilité. Ces technologies sont notamment :

- opaques et impénétrables
- complexes et évolutives
- fondées sur l'apport d'êtres humains, leur libre arbitre et les interactions avec eux
- de nature généraliste
- interconnectées, applicables et déployables dans le monde entier
- fondées sur de grands ensembles de données
- fonctionnant automatiquement et en continu, souvent en temps réel
- capables d'extraire des connaissances « cachées » à partir des masses de données fusionnées entre elles
- capables d'imiter avec fidélité des traits humains
- associées à des logiciels de plus en plus complexes (notamment exposés aux attaques et vulnérables aux dysfonctionnements)
- capables de « personnaliser » et de configurer un environnement adapté à chaque usager
- capables d'orienter les choix sociaux de manière à répartir les risques et les avantages pour optimiser un but donné (chapitre 1.3).

Le chapitre explique également l'**approche interdisciplinaire « par les droits de l'homme »** adoptée dans la présente étude, qui se fonde sur les droits de l'homme et les libertés fondamentales protégés par la CEDH pour :

- comprendre la nature des risques et des effets négatifs générés par les technologies numériques avancées ;
- contribuer à établir comment les responsabilités devraient être définies et attribuées ;
- nourrir le débat sur les types de mécanismes institutionnels qui pourraient être nécessaires pour assurer la protection effective des droits de l'homme.

Enfin, le chapitre 1 attire l'attention sur les travaux existants concernant les effets négatifs des technologies d'IA sur les droits de l'homme et les libertés fondamentales, travaux qui serviront de base au chapitre 2.

Chapitre 2 – Menaces, risques, préjudices et torts associés aux technologies numériques avancées

Le chapitre 2 examine une série de conséquences néfastes attribuables à l'usage des technologies numériques avancées. Il revient d'abord sur le contexte sociohistorique de cette innovation : les progrès actuels des technologies numériques en réseau pourraient susciter dans la vie économique et sociale des changements d'une ampleur aussi vaste et déstabilisante que la Première révolution industrielle. La « Nouvelle révolution industrielle » que l'on voit poindre pourrait ressembler à la Première, c'est-à-dire apporter d'innombrables avantages, mais aussi des effets indésirables qui ne se révéleront qu'avec le temps. Il est, de fait, extrêmement difficile de formuler des prévisions fiables sur les effets cumulés à long terme de la révolution des réseaux numériques que nous vivons actuellement.

Le chapitre 2 examine ensuite comment l'utilisation de systèmes algorithmiques pour la prise de décision («ADM»), reposant sur des techniques de profilage fondées sur les données, pourrait menacer plusieurs droits de l'homme (chapitre 2.1), dont les suivants :

- **le droit à un procès équitable et à une procédure régulière** (article 6 CEDH), en particulier lorsque les systèmes d'ADM servent à automatiser des décisions qui affectent de manière significative les individus tout en leur interdisant d'y participer, de les contester ou de mettre en question les éléments ayant abouti à ces décisions ou la décision en elle-même. Certains systèmes sont incapables d'expliquer leur logique sous-jacente en des termes intelligibles pour la personne concernée ;

- **le droit à la liberté d'expression et d'information** (article 10 CEDH), en particulier compte tenu de la puissante influence exercée aujourd'hui par les plateformes numériques mondiales sur les environnements informationnels des individus et des sociétés : des algorithmes décident automatiquement comment traiter, hiérarchiser, diffuser et supprimer les contenus en ligne, y compris en période de campagnes politiques ou électorales. Même bien intentionnés, les efforts des plateformes pour repérer et retirer les contenus « extrémistes » risquent sérieusement de contrevenir aux exigences de l'article 10.2, c'est-à-dire que ces activités ne répondent pas aux éxigences de légalité, légitimité et proportionnalité des ingérences autorisées à la liberté d'expression ;

- **le droit à la vie privée et à la protection des données** (article 8 CEDH), du fait de l'usage de technologies de profilage fondées sur les données. Les traces très personnelles que chacun de nous laisse en ligne étant collectées et traitées au niveau de toute une population, l'application de ces techniques de profilage ne peut qu'affecter le droit à la vie privée et familiale, énoncé à l'article 8. Bien que les régimes contemporains de protection des données (comme la Convention n° 108 modernisée) jouent un rôle important dans la préservation des droits et des intérêts des personnes concernées, la protection qu'ils offrent peut ne pas s'avérer effective et complète dans la pratique ;
- **le droit à la protection contre la discrimination dans l'exercice des droits et des libertés** (article 14 CEDH), en raison des risques importants de discrimination découlant du recours aux algorithmes d'apprentissage automatique. Il peut y avoir des partis pris chez les développeurs de l'algorithme eux-mêmes, au cœur du modèle qui sous-tend les systèmes, dans les ensembles de données utilisés pour entraîner les modèles ; les systèmes peuvent aussi devenir biaisés au cours de leur mise en œuvre dans le monde réel. Ces partis pris peuvent non seulement violer le droit à la protection contre la discrimination dans l'exercice des droits et libertés protégés, énoncé à l'article 14, mais aussi renforcer les

préjugés contre des populations traditionnellement défavorisées, exacerbant ainsi les discriminations et les désavantages structurels.

Le chapitre traite ensuite de la possibilité du profilage fondé sur les données, , de nuire aux valeurs et aux intérêts collectifs en étant appliqué à grande échelle, car il rend ainsi possible des pratiques de surveillance constante, de personnalisation et de manipulation au niveau de toute une population, risquant d'ébranler la dignité et l'autonomie humaines, par exemple en traitant systématiquement les individus comme des objets et non comme des sujets (chapitre 2.1.2).

Sont ensuite examinées les conséquences sociales négatives qui peuvent accompagner le développement et l'usage des technologies d'IA en général, même lorsqu'elles ne reposent pas sur le profilage des individus (chapitre 2.2). Ces conséquences sont les suivantes :

- risques de dommages à grande échelle en cas d'attaque contre un système
- conception des systèmes contraire à l'éthique ou dysfonctionnement imprévu des systèmes
- perte de contacts humains authentiques et concrets
- effet tétanisant de la réutilisation des données
- exercice irresponsable du pouvoir conféré par le numérique
- privatisation cachée de décisions d'intérêt public (dont la justice distributive)
- exploitation de main-d'œuvre humaine pour entraîner les algorithmes.

Pour finir, les réflexions soulignent le déséquilibre de pouvoir entre ceux qui développent et appliquent les technologies d'IA et ceux qui interagissent avec elles et y sont soumis (chapitre 2.3). Tandis que les fournisseurs de services numériques (et les autres acteurs concernés) qui appliquent des systèmes d'IA peuvent observer à la loupe les données de leurs usagers et en extraire des prédictions très détaillées sur leur personnalité et leurs goûts, les usagers ne saisissent généralement pas toute la complexité des technologies numériques qu'ils utilisent. Ils n'ont pas non plus accès, en retour, à des informations détaillées sur les organisations et entreprises dont ils utilisent les services. Non seulement cette opacité et ce déséquilibre renforcent la possibilité d'une éventuelle exploitation, mais ils peuvent aussi menacer des valeurs et intérêts collectifs que le discours actuel sur les droits de l'homme n'énonce pas clairement, comme les fondements sociotechniques des communautés démocratiques et morales. Ces menaces et risques collectifs sont aggravés par la vitesse et l'échelle sans précédent de ces technologies, qui génèrent des défis auxquelles les sociétés contemporaines ne s'étaient encore jamais confrontées. Parallèlement, ces technologies risquent aussi de compliquer les actions collectives : bien que leurs effets négatifs cumulés soient très importants à grande échelle, ils peuvent s'avérer relativement mineurs à l'échelle de l'individu, qui ne songera pas nécessairement à protester.

Chapitre 3 – Qui est responsable ?

Le chapitre 3 s'interroge sur la responsabilité associée aux effets néfastes des technologies numériques avancées. Pour commencer, il précise ce qu'on entend par responsabilité et pourquoi cette question doit être abordée. Il souligne le rôle crucial de nos institutions et de nos pratiques en matière de responsabilité pour que ceux qui ont porté préjudice à autrui et aux intérêts et valeurs collectifs soient tenus de rendre des comptes. Ces institutions et pratiques ont pour rôle inestimable d'assurer et de favoriser une coopération fiable et pacifique au sein de la société et de donner corps à la notion d'État de droit. Bien que le terme

de responsabilité connaisse de nombreuses acceptions différentes, le chapitre 3 insiste sur la distinction entre :

- **responsabilité historique (ou rétrospective)**, tournée vers le passé, cherchant à établir les responsabilités pour des comportements et événements qui se sont déjà produits, et
- **responsabilité prospective**, tournée vers l'avenir, définissant les obligations associées aux différents rôles et tâches afin de favoriser les résultats désirables et prévenir les indésirables. La responsabilité prospective joue un important rôle d'orientation, puisqu'elle nous renseigne sur nos droits et obligations vis-à-vis d'autrui et sur la manière dont nous devrions nous comporter avec autrui.

Pour nous, la responsabilité des effets néfastes des technologies d'IA doit être envisagée à la fois sous l'angle prospectif et sous l'angle historique (chapitre 3.2). Seule cette approche garantira que des mesures soient prises pour éviter les préjudices liés au développement et à la mise en œuvre de ces technologies (et le cas échéant, pour y mettre un terme). Nos sociétés doivent donc veiller à disposer des structures et de mécanismes institutionnels nécessaires pour que les parties lésées obtiennent réparation et pour éviter que les technologies d'IA ne causent des torts supplémentaires.

Le chapitre 3 examine ensuite les liens entre technologies numériques avancées (dont l'IA) et conceptions actuelles de la responsabilité (chapitre 3.3). À cette fin, il souligne les différences entre la notion de responsabilité morale et celle de responsabilité juridique. Contrairement à la morale, le droit dispose d'un système très développé d'institutionnalisation et de mise en œuvre de la responsabilité (y compris par l'application de sanctions), puisqu'il doit trancher des différends dans le monde réel. Il faut aussi garder à l'esprit que les systèmes d'IA peuvent avoir deux grands types d'effets négatifs, distincts bien qu'ils se recoupent parfois :

- les violations des droits de l'homme (dont les droits protégés par la CEDH), et
- les dommages matériels pour la santé humaine, les biens ou l'environnement.

Cette étude analyse au premier chef la responsabilité pour les violations des droits de l'homme, en se concentrant sur les personnes qui créent, développent, déploient et supervisent les systèmes d'IA et leurs environnements et sur l'obligation des États de veiller à ce que les droits de l'homme soient correctement protégés.

Parce que les systèmes d'IA ont une envergure géographique et temporelle sans précédent, ils peuvent remettre en question notre conception préexistante de la responsabilité. Le chapitre 3 étudie deux grands thèmes issus du débat contemporain sur les effets négatifs des technologies d'IA. Premièrement, la promulgation par le secteur des technologies de « normes d'éthique » qu'il s'engage à respecter. À nos yeux, quoique bienvenus sous de nombreux aspects, les codes et normes en question sont généralement dépourvus de tout mécanisme de mise en œuvre et de sanction et ne peuvent donc servir de base à une véritable protection (chapitre 3.3.1). Deuxièmement, le « problème de contrôle » qu'engendrerait l'aptitude des systèmes pilotés par IA à fonctionner plus ou moins indépendamment de leurs créateurs, problème qui créerait un « vide de responsabilité » puisqu'il serait injuste d'imputer aux développeurs les résultats produits par ces systèmes. Le chapitre 3 démontre que ce « problème de contrôle » allégué repose sur une théorie morale de la responsabilité très singulière, qui insiste indûment sur le comportement de l'agent et néglige les intérêts des victimes et la sécurité des personnes et des biens (chapitre 3.3.2).

Le chapitre 3 poursuit en identifiant et en exposant brièvement **plusieurs « modèles de responsabilité »** pouvant servir à attribuer et à répartir les responsabilités pour divers types

d'effets négatifs provoqués par les systèmes d'IA (chapitre 3.4). Ces modèles peuvent reposer sur :

- l'intention / la culpabilité (chapitre 3.4.1)
- le risque / la négligence (chapitre 3.4.2)
- la responsabilité absolue (chapitre 3.4.3)
- les régimes d'assurance obligatoires (chapitre 3.4.4).

Pour identifier le modèle le plus adapté à l'attribution de la responsabilité historique, l'analyse souligne l'importance de distinguer les violations des droits de l'homme, d'une part, et les dommages matériels pour la santé, les biens ou l'environnement, d'autre part (bien qu'un seul événement puisse entraîner à la fois des dommages matériels et une violation des droits de l'homme). En cas de violation de droits de tous types, y compris les droits de l'homme, la responsabilité est largement envisagée comme « absolue » : dès lors qu'une violation d'un droit de l'homme est établie, il n'est pas nécessaire qu'une faute soit prouvée. En revanche, les obligations de réparation en cas d'atteinte concrète à la santé ou aux biens peuvent être attribuées juridiquement, conformément à divers modèles historiques de responsabilité. Chaque modèle ménage un équilibre différent entre notre intérêt à pouvoir agir librement et notre intérêt, en tant que victimes, à préserver notre sécurité et celle de nos biens. Aucun de ces modèles n'est considéré comme évidemment « correct » ou « préférable » pour déterminer les responsables des menaces, risques et préjudices associés au fonctionnement des technologies numériques avancées. La détermination du modèle le plus approprié (s'il y en a) relève d'un *choix politique engageant toute la société.*

Le chapitre 3 attire ensuite l'attention sur plusieurs défis, de taille, soulevés par la détermination des responsabilités lorsque des systèmes sociotechniques complexes et en interaction entraînent des risques et des conséquences négatives (chapitre 3.5) :

a) **le problème de la multiplicité des acteurs**. Le développement et le fonctionnement de systèmes d'IA font entrer en jeu de nombreux individus, organisations, circuits et composantes, logiciels, algorithmes et usagers humains, souvent dans des environnements complexes et évolutifs. Le problème de la multiplicité des acteurs n'est pas nouveau et repose largement, en philosophie morale, sur la « théorie du choix ». Les systèmes juridiques contemporains ont développé des principes et procédures relativement élaborés pour déterminer les responsabilités lorsque plusieurs accusés peuvent être considérés comme ayant contribué à un événement indésirable. L'aptitude du droit à offrir des réponses pratiques et effectives au problème de la multiplicité des acteurs s'explique en partie par l'accent qu'il met sur les intérêts légitimes des victimes (et des victimes potentielles) à préserver leur sécurité. À cet égard, la réponse apportée par le droit diffère des théories de la responsabilité fondées sur le choix en philosophie morale, presque exclusivement centrées sur les actions de l'être moral. En outre, s'agissant des atteintes aux droits de l'homme entraînées par des systèmes d'IA, le chapitre souligne l'importance de mécanismes destinés à prévenir et à stopper les violations des droits de l'homme générées par l'application de technologies numériques avancées. Il est particulièrement important d'assurer une véritable prévention, car les effets cumulés de ces technologies pourraient sérieusement menacer les bases collectives indispensables à l'exercice concret des droits de l'homme et des libertés fondamentales. Ces menaces pointent la nécessité de renforcer et de raviver, à l'heure

du triomphe de la « data », la protection des droits de l'homme et le discours à leur sujet (chapitre 3.5.1) ;

b) **l'interaction humain-ordinateur.** Il est extrêmement difficile d'attribuer et de répartir correctement les responsabilités entre humains et ordinateurs, en particulier lorsqu'il y a « un humain dans la boucle ». Préoccupation récurrente : pour veiller à ce que les systèmes sociotechniques complexes intégrant l'IA restent au service de l'humanité, ils devraient toujours être conçus de manière à ce qu'un opérateur puisse les désactiver. Pourtant, et on peut le comprendre, les individus chargés de superviser le fonctionnement de ces systèmes hésitent à intervenir. Cela risque de transformer les humains placés dans la boucle en « amortisseurs de chocs moraux », en totems dont le rôle central deviendrait de prendre la faute sur eux, alors qu'ils ne contrôlent que partiellement le système, et que les développeurs et organisations concernés pourraient facilement ériger en boucs émissaires pour fuir leurs responsabilités en cas de problème (chapitre 3.5.2) ;

c) **l'interaction entre les systèmes algorithmiques.** La situation se complique encore lorsqu'il s'agit d'identifier, d'anticiper et de prévenir les effets néfastes des interactions entre des systèmes sociotechniques complexes et guidés par des algorithmes, fonctionnant à une vitesse et à une échelle qui n'étaient tout simplement pas possibles avant l'ère du numérique et des réseaux (citons par exemple le « krach éclair » de 2010). La nature imprévisible des interactions entre de multiples systèmes algorithmiques engendre des risques nouveaux et potentiellement catastrophiques que nous commençons tout juste à cerner, sans parler de les anticiper et d'y mettre un terme (chapitre 3.5.3).

Tous ces problèmes demandent à être suivis et étudiés de près à l'avenir.

Bien que le chapitre 3 insiste sur la responsabilité des concepteurs et développeurs des technologies et de ceux qui détiennent et appliquent les systèmes fondés sur ces technologies, le chapitre 3.6 rappelle qu'il appartient en premier lieu aux États de veiller à la protection effective des droits de l'homme. Il attire l'attention sur un enjeu que les systèmes d'IA sont susceptibles de soulever à notre époque d'interconnexion mondiale : celui des actions collectives, soulignant l'importance vitale a) d'une législation nationale garantissant la protection des droits de l'homme, b) de services répressifs nationaux dotés des ressources et des pouvoirs nécessaires et c) de mécanismes accessibles et adaptés de réclamation collective, s'ajoutant aux voies de recours individuelles, pour assurer une protection effective des droits de l'homme.

Le chapitre évoque ensuite une série de mécanismes extrajudiciaires qui pourraient aider à établir la responsabilité prospective et historique pour les effets négatifs des systèmes d'IA, dont les évaluations des impacts, les méthodes d'audit et les mécanismes techniques de protection, ces derniers s'avérant particulièrement prometteurs (chapitre 3.7). Ces mécanismes demandent à être inscrits dans un cadre de gouvernance permettant de définir les normes techniques pertinentes de façon transparente et participative, et d'assurer une supervision externe indépendante et un examen de leur fonctionnement.

Le chapitre 3.8 aborde brièvement la question de savoir si notre vision actuelle des droits de l'homme, et les mécanismes par lesquels ils sont protégés et appliqués, sont adaptés à notre époque d'interconnexion numérique mondiale. Il avance que le pouvoir des technologies numériques en réseau apparues ces dernières années rend possibles des actes et des

pratiques qui ne l'étaient pas auparavant et créent par là des menaces, des risques et des formes de préjudice nouveaux. Par conséquent, l'heure est peut-être venue de *renouveler notre discours sur les droits de l'homme*, afin de préserver et d'entretenir les bases sociotechniques nécessaires à la liberté d'action et à la responsabilité humaines, sans lesquelles les droits de l'homme et les libertés ne peuvent s'exercer véritablement. Une conception renforcée et renouvelée des droits de l'homme pourrait conduire au développement de nouveaux mécanismes institutionnels, mieux placés pour nous protéger des effets négatifs des nouvelles technologies numériques dans un monde régi par les données.

Les constats énoncés au chapitre 3 sont synthétisés en fin de chapitre (3.9).

Chapitre 4 – Conclusion

Le chapitre 4 conclut en résumant les arguments avancés au cours de l'étude. Il met en lumière quatre grands constats.

Premièrement, il est crucial que nous disposions de mécanismes effectifs et légitimes destinés à prévenir les atteintes aux droits de l'homme et à y mettre un terme, compte tenu de la rapidité et de l'échelle auxquelles de nombreux systèmes numériques avancés fonctionnent, faisant peser de fortes menaces sur les droits de l'homme sans nécessairement générer d'importants risques de dommages matériels. Il est particulièrement important d'adopter une approche préventive, car ces menaces pourraient gravement saper les bases sociales de nos ordres moraux et démocratiques, préalables essentiels à l'exercice de la liberté individuelle, de l'autonomie et des droits de l'homme. Une telle prévention suppose à la fois, entre autres, de mettre en place des mécanismes de réclamation collective pour assurer une véritable protection des droits et de renforcer et renouveler notre vision et notre acception actuelles des droits de l'homme.

Deuxièmement, le modèle de responsabilité juridique s'appliquant aux violations des droits de l'homme est largement considéré comme un modèle de « responsabilité absolue », qui s'applique même sans qu'une faute n'ait été prouvée. En revanche, les préjudices matériels peuvent entraîner des obligations juridiques de réparation, attribuées selon différents modèles de responsabilité dont chacun ménage un équilibre différent entre nos intérêts en tant que personnes libres de nos actions et nos intérêts en tant que victimes d'atteintes à notre sécurité et à celle de nos biens. Décider lequel de ces modèles (s'il y en a) est à même de prévenir les différents risques et menaces associés au fonctionnement des technologies numériques avancées ne va pas de soi : il s'agit en fait d'un *choix politique engageant toute la société*. Dans des sociétés démocratiques constitutionnelles œuvrant à protéger et à respecter les droits de l'homme, l'État a une responsabilité cruciale, celle de s'assurer que ces choix sont opérés de manière transparente et démocratique et de manière à ce que les politiques adoptées à terme protègent effectivement les droits de l'homme.

Troisièmement, nous devrions encourager et soutenir des recherches techniques sur l'établissement de la responsabilité prospective et historique en vue de faire respecter les valeurs qui sous-tendent la protection des droits de l'homme, recherches susceptibles de faciliter le développement de mécanismes techniques de protection et d'un « audit des algorithmes ». Puisque le sujet concerne à la fois les milieux techniques et ceux du droit, de la philosophie et des sciences sociales, ces recherches devraient être interdisciplinaires, en vue de mieux traduire la sauvegarde des droits de l'homme en mécanismes de protection intégrés

aux systèmes d'IA et de comprendre en quoi une approche par les droits de l'homme peut résoudre les problèmes de conflits de valeurs.

Quatrièmement, la protection effective des droits de l'homme à notre époque de connexion mondiale exige que nous nous dotions de mécanismes, instruments et institutions de gouvernance efficaces et légitimes pour surveiller, encadrer et superviser la conception, le développement, la mise en œuvre et le fonctionnement responsables de nos systèmes sociotechniques. Cela suppose, au minimum, que les normes pertinentes soient définies à travers une démocratie participative et que des autorités indépendantes, dotées des ressources et des compétences nécessaires, rassemblent méthodiquement les informations, enquêtent sur les cas de non-conformité et sanctionnent les violations. Ces autorités devraient notamment pouvoir examiner les systèmes concernés pour vérifier qu'ils respectent effectivement les normes et les valeurs des droits de l'homme.

L'étude conclut qu'à l'heure de la mondialisation et de l'interconnexion numérique, si nous voulons vraiment protéger et promouvoir les droits de l'homme, nous ne pouvons tolérer que les technologies et systèmes numériques avancés et ceux qui les développent et les appliquent exercent un pouvoir de plus en plus grand sans aucune responsabilité. Un principe fondamental s'applique ici, celui de la réciprocité : ceux qui offrent des services (dont ils tirent des bénéfices) en profitant des avantages de ces technologies numériques avancées, y compris l'IA, doivent assumer la responsabilité de leurs conséquences négatives. Il est par conséquent crucial que les États, tenus de protéger les droits de l'homme, s'engagent aussi à faire en sorte que ceux qui exercent le pouvoir numérique (dont le pouvoir tiré de l'accumulation massive de données) aient à répondre des conséquences de leurs actes. L'engagement des États à protéger les droits de l'homme les oblige aussi à faire en sorte qu'il existe, en droit national, des structures de gouvernance et des mécanismes de mise en œuvre assurant dûment l'attribution de la responsabilité prospective et historique pour les risques, les préjudices et les torts engendrés par les technologies numériques avancées.

Incidences des technologies numériques avancées (dont l'intelligence artificielle) sur la notion de responsabilité, sous l'angle des droits humains

par Karen Yeung*

« Un grand défi mondial se pose à tous ceux qui œuvrent à promouvoir les droits de l'homme et l'État de droit : comment les États, les entreprises et la société civile peuvent-ils faire en sorte que les techniques d'intelligence artificielle respectent et renforcent les droits de l'homme plutôt que de les fragiliser et de les menacer ? »

David Kaye, Rapporteur spécial sur la promotion et la protection du droit à la liberté d'opinion et d'expression, Assemblée générale des Nations Unies (2018)

* Avec les contributions de mes collègues Ganna Pogrebna et Andrew Howes. Assistants de recherche : Charlotte Elves et Helen Ryland, Université de Birmingham. Je remercie Imogen Goold pour ses conseils sur la teneur et la structure du droit anglo-américain de la responsabilité délictuelle.

Chapitre 1. Introduction

1.1 Champ de l'étude

Cette étude examine les incidences des « nouveaux services et technologies numériques – y compris l'intelligence artificielle » sur la notion de responsabilité, sous l'angle des droits de l'homme. Elle se concentre sur les technologies dites d'« intelligence artificielle » (IA). La difficulté à définir l'intelligence artificielle est notoire ; aucune définition ne semble faire l'unanimité, même parmi les spécialistes. Aux fins de la présente étude, nous adopterons la définition proposée par la Commission européenne dans sa Communication sur l'intelligence artificielle[1]. Elle est rédigée comme suit :

> L'intelligence artificielle (IA) désigne les systèmes qui font preuve d'un comportement intelligent en analysant leur environnement et en prenant des mesures – avec un certain degré d'autonomie – pour atteindre des objectifs spécifiques. Les systèmes dotés d'IA peuvent être purement logiciels, agissant dans le monde virtuel (assistants vocaux, logiciels d'analyse d'images, moteurs de recherche ou systèmes de reconnaissance vocale et faciale, par exemple), mais l'IA peut aussi être intégrée dans des dispositifs matériels (robots évolués, voitures autonomes, drones ou applications de l'internet des objets, par exemple). […] De nombreuses technologies de l'IA ont besoin de données pour gagner en efficacité. Une fois performantes, elles peuvent contribuer à améliorer et à automatiser le processus décisionnel dans le même domaine.

Ainsi, notre étude appelle « IA » une série de technologies généralistes avancées qui permettent à des machines d'accomplir efficacement des tâches d'une grande complexité, en puisant dans une série de techniques complémentaires issues de la statistique, de l'informatique et de la psychologie cognitive[2]. Ces technologies visent à reproduire ou à surpasser (via l'informatique) des aptitudes qui nécessiteraient de l'« intelligence » chez les êtres humains : aptitude à apprendre et à s'adapter, compréhension et interactions sensorielles, raisonnement et planification, optimisation des procédures et des paramètres, autonomie, créativité, extraction de savoirs et de prédictions à partir d'une grande diversité de données numériques, etc.[3]. Notre étude se limite aux technologies d'IA actuellement disponibles (au moins dans le cadre de recherches ou de démonstrations de prototypes) ou envisageables dans les cinq ans à venir, en insistant sur celles qui utilisent l'apprentissage automatique, et part de l'idée que l'IA appliquée à des tâches spécifiques va davantage progresser que l'IA « généraliste[4] ». Elle ne s'intéresse qu'à l'emploi de l'IA comme technologie, c'est-à-dire au service de tâches concrètes, et non à l'IA comme outil de recherche scientifique à l'usage des universitaires et des chercheurs[5].

Les technologies d'IA présentent indéniablement d'énormes avantages. En particulier, elles ont amélioré l'efficacité, la précision, la rapidité et la commodité de nombreux services. Beaucoup de ces applications peuvent être vues comme élargissant la portée des services en

---

1 Commission européenne 2018a. Définition approfondie dans Groupe d'experts de haut niveau de l'UE sur l'intelligence artificielle, 2019b.

2 EPSRC ; Hall et Pesenti 2017.

3 EPSRC.

4 Bostrom 2014.

5 L'utilisation de l'apprentissage automatique dans la recherche privée et scientifique ne va pas sans difficultés. Voir par exemple Leonelli 2018 ; Metcalfe et Crawford 2016.

question et, par là, l'exercice des libertés et des droits de l'homme. Par exemple, sans moteurs de recherche utilisant l'IA, la masse d'informations aujourd'hui disponible sur internet ne serait ni utile ni accessible en pratique ; l'IA favorise ici le droit à la liberté d'information (protégé par l'article 10 de la Convention européenne de sauvegarde des droits de l'homme et des libertés fondamentales, ci-après : « CEDH »). Dans le monde entier, de nombreux gouvernements nationaux et organisations régionales consacrent des sommes considérables à l'élaboration de stratégies en faveur de l'innovation et du développement des technologies d'IA, sur la base d'une conviction largement partagée : ces technologies peuvent, et vont, améliorer significativement l'efficacité, la productivité et la qualité des services[6]. Pourtant, les premiers hauts faits des technologies numériques avancées, qui ont suscité un « boom de l'IA » et ouvert une véritable « course à l'IA[7] », se sont accompagnés d'inquiétudes grandissantes face aux possibles effets négatifs de ces technologies pour les individus et pour toute la société[8]. Ces inquiétudes, en soulignant les risques potentiels et avérés et les impacts négatifs, ont mis en lumière la question de la responsabilité : il faut veiller, dans le cadre de nos régimes démocratiques constitutionnels, à ce que les individus et les organisations répondent des effets négatifs de leurs actes sur autrui[9]. De ce fait, l'objectif de cette étude est d'examiner les incidences des technologies numériques avancées (dont l'IA) sur la notion de responsabilité, en particulier là où ces technologies pourraient entraver l'exercice des droits de l'homme et des libertés fondamentales. À cette fin, elle examine les effets, à la fois volontaires[10] et involontaires[11], du développement et de l'usage de l'IA qui peuvent être considérés comme nuisant directement à l'exercice des libertés et des droits de l'homme. Cependant, notre étude n'aborde ni les effets négatifs *indirects* de l'IA, comme le risque de chômage de masse, ni les autres effets indirects plus ou moins éloignés, ni les aspects militaires (dont les systèmes d'armement autonomes). Non que ces risques soient négligeables ; mais ils soulèvent des préoccupations particulières qui n'entrent pas dans le champ de nos travaux.

1.2 Structure de l'étude

Notre étude vise à cerner les responsabilités à l'égard des risques potentiels et avérés et des conséquences néfastes, au niveau individuel et social, associés à l'état actuel et prévisible des technologies numériques avancées, qui ne cessent de gagner en puissance et en complexité. Elle adopte ce qu'on pourrait appeler « le point de vue des droits de l'homme », dans la mesure où les droits de l'homme et les libertés fondamentales protégés par la CEDH peuvent aider à la fois a) à comprendre la nature de ces risques et de ces conséquences, b) à

---

6 La Commission européenne a prévu de consacrer « au moins 20 milliards d'euros » aux technologies d'IA d'ici 2020 (White 2018), tandis que le Royaume-Uni a récemment engagé un milliard de livres (UK Department for Digital, Culture, Media and Sport 2018 ; UK Department for Business, Energy and Industrial Strategy 2018).

7 Voir Financial Times 2018. Sur la rivalité entre la Chine, les États-Unis et l'UE, voir Centre européen de stratégie politique 2018.

8 Voir les sources citées à la note n° 50, ci-dessous.

9 Concernant le concept de responsabilité et son importance, voir le chapitre 3.1. Bien que les effets positifs et bénéfiques de l'IA sur des tiers puissent aussi soulever des questions de responsabilité, étant donné que notre étude s'intéresse aux impacts négatifs des technologies numériques avancées *sur les droits de l'homme*, nous nous concentrerons sur les responsabilités à l'égard des impacts *négatifs* associés à ces technologies.

10 L'IA connaît des « usages malveillants », consistant à attaquer des tiers (Brundage et al. 2018). D'autres effets négatifs peuvent être volontaires mais non nécessairement malveillants. Voir les exemples étudiés par Sandvig et al. 2014.

11 O'Neil 2016.

déterminer comment les responsabilités les concernant devraient être attribuées et réparties, et c) à étudier les types de mécanismes institutionnels qui pourraient être nécessaires pour que les droits de l'homme soient effectivement protégés et les responsabilités en matière de protection de ces droits dûment attribuées[12]. À cette fin, l'étude s'appuie sur des concepts et des connaissances issues du droit, de la philosophie et des sciences sociales, dont la philosophie morale, juridique et politique et l'économie politique, ainsi que de l'informatique, plutôt que de se concentrer sur la jurisprudence de la Cour européenne des droits de l'homme. Elle se divise en quatre chapitres.

Le **chapitre 1** décrit dans ses grandes lignes le fonctionnement des technologies d'IA avant d'identifier leurs propriétés et celles de leurs applications, actuelles et prévisibles, qui s'avèrent pertinentes pour la question de la responsabilité.
Le **chapitre 2** examine les conséquences individuelles et collectives négatives que pourrait avoir l'application des technologies numériques avancées. Il aborde en premier lieu le profilage par les données, qui pourrait faire peser une menace constante sur certains droits et sur des valeurs et intérêts collectifs. Il étudie ensuite les risques potentiels et avérés posés par d'autres technologies d'IA et par leurs applications, contemporaines ou prévisibles. Le chapitre 2 conclut en soulignant le déséquilibre de pouvoir croissant entre, d'une part, ceux qui ont la capacité et les ressources nécessaires pour développer et appliquer les technologies d'IA et, d'autre part, les individus, groupes et populations directement affectés par leur usage.

Le **chapitre 3** s'interroge sur les responsabilités en cas d'impacts négatifs, en particulier s'ils dégénèrent en atteintes aux droits et/ou en dommages, y compris contre des valeurs et intérêts collectifs, et menacent les bases sociotechniques de la liberté démocratique et des droits de l'homme. Il examine plusieurs « modèles juridiques de responsabilité » pouvant servir à définir les responsables de ces risques et de ces conséquences. Il évoque également les difficultés soulevées par l'attribution des responsabilités en présence de systèmes sociotechniques extrêmement complexes, auxquels participent de multiples organisations, individus et composantes matérielles et logicielles en interaction constante. Il pointe ensuite une série de mécanismes qui pourraient aider à lever certaines de ces difficultés afin d'assurer une protection effective et légitime des droits de l'homme.

Le **chapitre 4** conclut.

1.3 Incidences de l'IA sur la notion de responsabilité

Pour pouvoir examiner les incidences de l'IA sur la notion de responsabilité sous l'angle des droits de l'homme, il faut connaître dans leurs grandes lignes les modalités de développement et de fonctionnement de ces technologies.

---

12 Comme observé par la Commission australienne des droits de l'homme, l'approche par les droits de l'homme offre « un mécanisme plus substantiel » que celui de l'« éthique des technologies » pour « identifier, prévenir et atténuer les risques » en « transformant les concepts de droits et de libertés en politiques et pratiques effectives et en réalités concrètes. Les principes internationaux des droits de l'homme incarnent ces valeurs fondamentales, et l'approche par les droits de l'homme fournit les mécanismes et outils nécessaires pour les concrétiser en les appliquant et en amenant les personnes concernées à rendre des comptes ». Australian Human Rights Commission 2018 : 17.

(a) Machines intelligentes et machines apprenantes

L'effervescence autour de l'IA et de son potentiel – faire progresser de nombreux pans de la société, de la production industrielle à la sécurité alimentaire en passant par la santé, la médecine ou la gestion de l'environnement – repose pour beaucoup sur la puissance de l'apprentissage automatique[13]. On appelle « apprentissage automatique » la technologie qui permet à un système informatique d'exécuter avec intelligence des tâches spécifiques en apprenant à partir d'exemples, de données et de l'expérience[14]. Si elles existaient déjà depuis un certain temps, les techniques d'apprentissage automatique ont connu un grand essor ces dernières années du fait des évolutions technologiques, de l'apparition d'ordinateurs de plus en plus puissants et de la croissance fulgurante du volume de données numériques disponibles. Ces avancées ont permis la mise au point de machines qui dépassent aujourd'hui les humains sur des tâches spécifiques (comme le traitement des langues, l'analyse des données, la traduction ou la reconnaissance des images), alors qu'elles peinaient à atteindre des résultats corrects il y a quelques années encore[15]. Ces technologies ont aujourd'hui envahi le quotidien des sociétés très industrialisées. Dans ces sociétés, les habitants interagissent désormais régulièrement avec des systèmes d'apprentissage automatique qui permettent à des services numériques (moteurs de recherche, systèmes de navigation ou de recommandation de produits, etc.) de répondre de manière précise, efficace et en temps réel aux demandes des utilisateurs tout en apprenant de leurs erreurs pour s'améliorer en continu[16].

(b) Propriétés de l'IA et responsabilité

Pour cerner en quoi les technologies numériques avancées (dont l'IA) remettent en question nos conceptions actuelles de la responsabilité sur le plan juridique, moral et social, il faut identifier leurs propriétés « pertinentes pour la responsabilité », c'est-à-dire susceptibles d'affecter l'impact de ces technologies sur autrui.

Automatisation des tâches

À cet égard, l'une des propriétés les plus importantes de ces technologies réside dans leur capacité à réaliser des tâches (dont beaucoup nécessitaient auparavant des opérateurs humains) « automatiquement », c'est-à-dire sans intervention humaine directe[17].

Autonomie des machines

Les progrès des techniques d'apprentissage automatique ont entraîné le développement et l'utilisation croissante de systèmes non seulement automatisés, mais présentant aussi une certaine autonomie. Bien que le terme d'« autonomie » soit couramment employé pour décrire des applications de l'IA dans le débat public et politique, il ne semble pas y avoir de consensus au sein des milieux techniques sur le sens exact à lui donner et sur les conditions qu'une entité non humaine doit remplir pour être qualifiée d'« autonome ». Cependant, dans les documents politiques, le terme d'« autonomie » désigne souvent la capacité fonctionnelle

---

13 Russell et Norvig 2016.

14 Royal Society 2017 : 16.

15 Royal Society 2017 : 16. Il peut arriver, par exemple, qu'un algorithme analyse une image mieux qu'un radiologue (The Economist 2018a) ou que l'IA fasse mieux que des juristes sur certaines fonctions précises (Mangan 2017).

16 Pour un exemple de la co-évolution des comportements humains en réaction aux systèmes de navigation avec apprentissage automatique, voir Girardin et Blat 2010.

17 Liu 2016.

d'agents informatiques à accomplir des tâches de manière indépendante, ce qui suppose que ces agents « décident » de leur propre comportement sans intervention directe d'opérateurs et en l'absence de contrôle humain. Les agents informatiques de ce type fonctionnent en percevant leur environnement et en adaptant leur comportement en fonction des retours sur leur exécution des tâches. Leurs décisions et leurs actions sont conçues d'emblée comme « non entièrement déterminées » (et donc non entièrement prévisibles), en raison de la diversité presque infinie de contextes et d'environnements dans lesquels ces agents peuvent opérer[18]. De ce point de vue, il n'y a pas de « tout ou rien » : l'autonomie peut être plus ou moins présente en fonction du degré de supervision et d'intervention humaines nécessaire au fonctionnement du système[19].

Encadré 1 : Autonomie des machines et sensibilité au contexte

Comparaison : aspirateur autonome, voiture autonome

- Fondamentalement, l'architecture technique est la même : les concepteurs du système définissent l'objectif global, mais pour l'atteindre, les deux machines sont capables de déterminer leurs propres sous-objectifs.
- Dans les deux cas, le comportement des machines ne peut être totalement déterminé à l'avance.
- Chacune est capable de percevoir son environnement et d'adapter ses décisions et ses actions en conséquence.
- Pourtant, elles opèrent dans des contextes très différents (environnement fermé relativement stable pour l'aspirateur, circulation routière complexe et évolutive pour la voiture).

Plus l'environnement ou le contexte dans lequel ces systèmes opèrent est stable et prévisible, plus il est possible d'anticiper leurs réactions et leurs performances. Il sera donc plus facile de prévoir et d'anticiper le comportement d'un aspirateur autonome que celui d'une voiture autonome.

Certains systèmes d'apprentissage automatique se distinguent par leur capacité à apprendre et à évoluer avec le temps, en se fixant leurs propres sous-objectifs et en s'adaptant aux

18 Groupe européen d'éthique des sciences et des nouvelles technologies (GEE) 2018. Le GEE observe une tendance à rechercher des niveaux d'automatisation et d'« autonomie » toujours plus élevés dans les domaines de la robotique, de l'IA et de la mécatronique (combinaison d'IA et d'apprentissage profond, de science des données, de technologie des capteurs, d'internet des objets et d'ingénierie mécanique et électrique), mais aussi « une évolution vers une interaction toujours plus étroite entre l'homme et la machine », notant que des équipes bien préparées de systèmes d'IA et de professionnels obtiennent de meilleures performances dans certains domaines que les hommes et les machines séparément.

19 Pour décrire le degré de contrôle ou de participation des opérateurs dans un système, la Royal Academy of Engineering distingue quatre niveaux différents : a) systèmes contrôlés : l'humain exerce un contrôle total ou partiel (exemple : une voiture ordinaire) ; b) systèmes supervisés : le système suit les instructions d'un opérateur (exemple : tour ou autre mécanisme industriel programmé) ; c) systèmes automatiques : accomplissent sans intervention d'un opérateur des fonctions définies (exemple : un ascenseur), et d) systèmes autonomes, qui s'adaptent, apprennent et peuvent prendre des « décisions ». Royal Academy of Engineering 2009 : 2. SAE International a élaboré la norme J3016_201806 : *Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems* (SAE International 2018), utilisée par exemple par le ministère étasunien des Transports dans le cadre de sa Politique fédérale sur les véhicules automatiques : US Department of Transportation 2017.

conditions locales grâce aux informations transmises par leurs capteurs ou en recevant des données mises à jour[20]. Les concepteurs du système définissent et règlent son état et ses paramètres initiaux, dont l'objectif global qu'il est censé atteindre de manière optimale ; mais une fois le système déployé, son fonctionnement et ses réalisations évoluent en fonction de son environnement. En particulier, ces systèmes informatiques sont conçus pour prendre des décisions de manière indépendante, en optant pour la meilleure alternative selon des modalités non programmées à l'avance, et pour le faire sans intervention humaine. Les systèmes d'IA actuels ne peuvent pas choisir l'objectif global que le système est censé atteindre (fixé par les développeurs du système), mais ils sont capables de déterminer eux-mêmes leurs sous-objectifs ou buts intermédiaires.

Pour savoir qui sont les responsables des résultats et des effets de ces systèmes, leur caractère *stable* et *prévisible* est particulièrement important (voir l'encadré 1). Parce que ces systèmes tirent en continu, par itération, des enseignements de leur environnement (lui-même souvent mouvant et désordonné), ces technologies peuvent présenter des évolutions inattendues. En pratique, cela signifie que leurs réalisations deviennent parfois opaques et imprévisibles (nous y reviendrons plus loin), avec des conséquences directes sur la possibilité d'appliquer le concept de responsabilité à leurs décisions, à leurs actions et aux conséquences qui en découlent.

Outre qu'elles peuvent fonctionner en l'absence de supervision et de contrôle directs par un être humain, ces technologies présentent plusieurs autres caractéristiques pertinentes pour la responsabilité. Elles sont notamment :

a. opaques et impénétrables. Les inquiétudes autour de l'opacité des technologies[21] revêtent trois aspects, distincts mais liés entre eux[22]. Premièrement, contrairement aux premières formes d'IA – des « systèmes experts » qui raisonnaient par règles : « si… alors… » –, les systèmes d'apprentissage automatique contemporains créent et utilisent des modèles très complexes, dont il peut être difficile de déceler la logique pour savoir pourquoi et comment ils ont produit tel ou tel résultat. Bien que certaines formes de systèmes d'apprentissage (comme ceux qui utilisent des arborescences de décisions) permettent de retracer et de comprendre la logique sous-jacente, tel n'est pas le cas pour d'autres (notamment ceux qui utilisent les réseaux neuronaux et la rétropropagation[23]). Deuxièmement, même lorsqu'ils utilisent des algorithmes dont le fonctionnement et la logique sous-jacente sont compréhensibles et explicables en termes humains, les systèmes développés par des entités commerciales ne sont pas toujours examinables, car soumis à des droits de propriété intellectuelle qui autorisent les détenteurs des droits à garder leurs algorithmes secrets[24]. Troisièmement, même lorsque les informations relatives au système sont fournies (technique utilisée pour entraîner l'algorithme, règles suivies par le système informatique, etc.), ceux qui n'ont pas les connaissances techniques nécessaires restent incapables de comprendre ces informations ou d'en saisir tous les aspects, ce qui réduit en

20 Michalski et al. (2013).

21 Voir Wagner 2017 : 36-37.

22 Burrell 2016.

23 On constate une montée des recherches techniques sur l'« IA explicable », visant à identifier des méthodes par lesquelles rendre ces systèmes intelligibles pour les êtres humains. Cette opacité pointe nos limites lorsqu'il s'agit de comprendre totalement, ou d'expliquer, le fonctionnement de systèmes complexes : nous raisonnons différemment des machines (Zalnieruite 2019). Voir plus loin, chapitre 3.7.1.

24 Voir par exemple *State vs Loomis* 881 N.W. 2d 749 (Wis. 2016). Noto La Diega (2018).

pratique la transparence du système[25]. Ce caractère opaque et en même temps impénétrable amène à qualifier les algorithmes de « boîtes noires[26] » et se répercute directement sur la transparence des applications qui les utilisent, la possibilité de les expliquer et la capacité de leurs concepteurs à rendre des comptes[27] ;

b. **complexes et évolutives.** Les applications qui utilisent l'IA à des fins sociales précises peuvent être considérées comme des systèmes sociotechniques complexes : leurs mécanismes sous-jacents, tout comme leurs interactions évolutives et continues avec les environnements dans lesquels elles opèrent, obéissent à une logique opérationnelle complexe, générant des résultats souvent difficiles à prévoir, en particulier lorsque ces applications recourent à des algorithmes d'apprentissage automatique[28]. Par conséquent, il peut s'avérer extrêmement difficile de comprendre et d'anticiper leur fonctionnement en contexte réel, même pour les personnes dotées des connaissances techniques voulues, qui par ailleurs s'étendent souvent à des domaines multiples ;

c. **fondées sur l'apport d'êtres humains, leur libre arbitre et les interactions avec eux.** Bien que les progrès de l'IA soient fortement liés à l'avènement de l'« ère de l'informatique », il ne faut pas oublier que chaque étape du développement et de la mise en œuvre des technologies d'IA associe des êtres humains : idées initiales, propositions de développement, conception, modélisation, collecte et analyse des données, tests, mise en œuvre, fonctionnement et évaluation[29]. En outre, ces systèmes sont souvent conçus pour fonctionner dans des environnements réels et pour interagir avec des êtres humains, souvent à grande échelle de surcroît (comme le système News Feed de Facebook). En particulier, de nombreuses applications utilisant l'IA sont pensées, sur la forme, pour préserver le libre arbitre de leurs utilisateurs : au lieu de décisions ou de fonctions automatiques prédéfinies, le système présente des « suggestions[30] ». Ainsi, par exemple, les moteurs de recommandation de produits suggèrent des produits aux utilisateurs, mais la décision appartient toujours à ces derniers, ce qui peut avoir des incidences notables en matière de responsabilité[31] ;

d. **de nature généraliste.** Les technologies d'IA peuvent être comprises comme « généralistes » car on peut leur imaginer des applications dans presque tous les domaines de la société. Cette polyvalence s'accompagne d'un classique « revers de la médaille » : comme d'autres technologies, elles peuvent être utilisées par des acteurs bienveillants, mais aussi malveillants ou ne poursuivant que leurs propres intérêts[32] ;

e. **interconnectées, applicables et déployables dans le monde entier.** Il est important de reconnaître que l'interconnexion et la portée mondiale d'internet (et des technologies connectées) ont permis un déploiement rapide des technologies d'IA à très grande échelle, encore renforcé par l'essor généralisé des appareils connectés « intelligents ». Beaucoup

---

25 Burrell 2016 : 4.

26 Pasquale 2015.

27 Burrell 2016 ; Datta et al. 2016. Weller 2017 ; Yeung et Weller 2019.

28 Schut et Wooldridge 2000.

29 Bryson et Theodorou 2018.

30 Su et Taghi 2009.

31 Voir plus loin la question des « humains dans la boucle », chapitre 3.5.2.

32 Sur la conception de systèmes algorithmiques pour servir certains intérêts, voir le passage consacré au système de réservation pour le transport aérien SABRE, dans Sandvig et al. 2014. Sur les applications malveillantes de l'IA, voir Brundage et al. (2018).

d'applications de l'IA utilisées quotidiennement par les habitants du monde industrialisé sont devenues omniprésentes. Comme elles permettent de gérer facilement et efficacement les tâches courantes des sociétés contemporaines, il devient de plus en plus difficile de concevoir la vie moderne sans elles[33]. Cependant, l'étendue et le taux de pénétration des infrastructures de données connectées et la place des appareils connectés intelligents restent faibles et limitées dans le Sud par rapport au Nord, si bien que les habitants du Sud n'ont pas autant accès aux services et aux gains en efficacité et en commodité que les habitants des pays plus riches et plus industrialisés[34] ;

f. **fonctionnant automatiquement, en continu et en temps réel.** L'efficacité et la commodité de nombreuses applications d'IA s'expliquent, dans une large mesure, par leur capacité à fonctionner automatiquement et en temps réel[35]. Par exemple, les systèmes de navigation par IA offrent une aide précieuse à ceux qui doivent gagner une destination totalement inconnue, puisqu'ils indiquent en temps réel quelle direction prendre tout en donnant une estimation du temps de trajet et en conseillant des itinéraires alternatifs[36]. Ces applications ne sont possibles que parce que les technologies d'IA peuvent collecter des données numériques à partir de capteurs intégrés à des appareils connectés, ce qui leur permet de suivre les activités et les déplacements des individus de très près et, souvent, sans que ces derniers en aient conscience. Ces possibilités technologiques ont des incidences directes sur la notion de responsabilité, qui peuvent affecter l'exercice des droits de l'homme et des libertés de trois manières au moins. Premièrement, l'avènement d'internet – et des appareils connectés – a permis à beaucoup de ces technologies de fonctionner en réseau, et donc à grande échelle et en temps réel. Par conséquent, il peut y avoir une distance considérable, dans le temps comme dans l'espace, entre la conception et la mise en œuvre de ces systèmes et le lieu/moment où leurs décisions et conséquences surviennent et se font directement sentir. Deuxièmement, il est extrêmement difficile de superviser et de contrôler des systèmes qui fonctionnent en temps réel et à grande échelle, défi sur lequel nous reviendrons plus loin. Troisièmement, l'offre de conseils très personnalisés et tenant compte en temps réel des tendances environnantes (comme les embouteillages) nécessite une surveillance continue des individus à l'échelle de toute une population, et donc une collecte et un traitement constants de données personnelles qui touchent forcément à deux droits de l'homme et valeurs collectives : la vie privée et la protection des données[37] ;

g. **fondées sur de grands ensembles de données.** Tandis que le fonctionnement d'un algorithme informatique dépend du modèle sur lequel il repose, les systèmes d'apprentissage automatique ne peuvent fonctionner correctement qu'en étant alimentés par des ensembles de données[38]. Privés d'accès à des ensembles de données pertinents, les algorithmes d'apprentissage automatique ne seraient que des coquilles vides. De ce fait, la disponibilité, la taille et la qualité des ensembles de données servant à entraîner, tester et valider les algorithmes jouent un rôle crucial pour leurs performances et pour l'exactitude et la légitimité de leurs résultats, tout comme la disponibilité et la qualité des données sur lesquelles ces systèmes reposent en cours de fonctionnement ;

---

33 Zuboff 2015 ; Royal Society 2017.

34 McSherry 2018.

35 Pour des exemples d'applications d'IA en temps réel, voir Narula (2018).

36 Swan 2015.

37 Voir les réflexions au chapitre 2, ci-dessous.

38 Kitchin 2014 ; Prainsack 2019.

h. **capables d'extraires des connaissances suite à la fusion d'ensembles de données.** L'effervescence entourant les technologies d'IA s'explique en grande partie par leur capacité à extraire de nouvelles connaissances suite à la fusion d'ensembles de données, ensuite utilisables pour déterminer et orienter les prises de décisions. Certains ensembles de données ne contiennent que des informations personnelles relativement anodines et sans intérêt. Cependant, la fusion et l'exploitation de plusieurs ensembles livre parfois des renseignements qui peuvent permettre de déduire des informations personnelles très intimes, avec un degré d'exactitude très élevé[39]. Par conséquent, les modalités de gouvernance de la collecte et du traitement des données numériques ont des répercussions très profondes sur les droits de l'homme et sur la notion de responsabilité, rendues plus importantes encore par la facilité et le coût pratiquement négligeable du transfert et de la copie de données numérique et par la complexité de l'écosystème de données de notre monde contemporain ;

i. **capables d'imiter des traits humains.** Ces dernières années, l'imitation de traits humains par les technologies d'IA – voix de synthèse, représentations visuelles de comportements humains et robots capables d'interagir avec des êtres humains en paraissant sensibles aux émotions – est devenue de si grande qualité qu'il peut s'avérer très difficile, pour un individu lambda, de comprendre que ces traits sont générés artificiellement. Certains redoutent par conséquent que ces technologies ne trompent les êtres humains (notamment en produisant des trucages ultra-réalistes) et soient employées à des fins contraires à l'éthique ou dans d'autres buts malveillants[40] ;

j. **associées à des logiciels de plus en plus complexes.** L'apprentissage automatique et les systèmes d'apprentissage profond sont en train de gagner en complexité, du fait non seulement de la disponibilité des données, mais aussi de la complexité croissante des programmes. Ces systèmes présentent donc trois types de vulnérabilités : premièrement, la grande complexité des programmes augmente la propension des systèmes à générer des éléments aléatoires (c'est-à-dire à faire des erreurs[41]) ; deuxièmement, cette complexité ouvre la voie à un large éventail d'attaques[42] ; et troisièmement, l'imprévisibilité des résultats peut susciter de très lourds effets indésirables sur les tiers (« externalités ») ;

k. **capables de « personnaliser » et de configurer un environnement adapté à chaque usager.** La « personnalisation » des services offerts explique en partie pourquoi les systèmes d'IA ont pu rendre un large éventail de processus et d'opérations plus efficaces et plus précis. Par exemple, le recours à des techniques de profilage permet aux sites de commerce en ligne (comme Amazon) de recommander à chaque client des produits « personnalisés », sur la base de prédictions fondées sur les données (issues de la collecte et de l'analyse en continu des traces numériques de ce client comparées à celles des autres[43]). Bien que cette personnalisation des services et des offres numériques bénéficie

---

39 Kosinski et al. 2013.

40 The Economist 2017 ; Chesney et Citron 2019. Voir les réflexions au chapitre 2.2.2, ci-dessous.

41 De récentes recherches sur la reconnaissance des images montrent que les technologies ne savent pas démêler des données comportant beaucoup de bruit. Confronté à des photos de chihuahuas mélangées à des photos de muffins, un algorithme d'IA s'est montré incapable de les distinguer (Yao 2017).

42 Les cybercriminels peuvent exploiter sans difficulté – et exploitent effectivement – les points faibles des systèmes d'IA à leurs propres fins. Ils peuvent par exemple falsifier la reconnaissance vocale et les systèmes CAPTCHA pour pénétrer dans des comptes personnels et commerciaux (Polyakov 2018).

43 Yeung 2016.

aux utilisateurs, puisqu'elle réduit le volume de propositions non pertinentes, elle produit aussi une segmentation : chacun ne voit que son environnement informationnel personnalisé, qui peut être très différent de ce que voient les autres. Lorsque la personnalisation par IA prend de l'ampleur et devient courante, elle risque de favoriser la fragmentation sociale[44] et d'éroder la cohésion sociale et la solidarité[45] ;

l. **capables de répartir les risques, les avantages et les inconvénients entre les groupes et les individus, via des systèmes d'optimisation par IA, pour reconfigurer les choix et les environnements sociaux.** Les systèmes d'IA peuvent fonctionner en temps réel et à grande échelle, à travers l'architecture mondiale du réseau internet. Ces systèmes peuvent donc être configurés de manière à optimiser la réalisation de l'objectif global assigné par leurs développeurs sur une échelle qui était impossible avant internet [46]. La personnalisation de l'environnement informationnel dans lequel chaque individu opère ses choix est particulièrement puissante lorsqu'elle s'applique à grande échelle. Elle permet de concevoir et de déployer des systèmes de répartition des contenus destinés à influencer et à orienter le comportement de toute une population d'usagers, au lieu d'un usager isolé, conformément au type d'optimisation choisi par les développeurs. Inévitablement, ces systèmes font passer certaines valeurs avant d'autres, et par là configurent et façonnent les environnements sociaux et informationnels d'une manière qui peut bénéficier à certains individus et groupes au détriment des autres. Par exemple, un système de navigation par IA peut viser à permettre à chaque conducteur de gagner sa destination le plus rapidement possible compte tenu de la densité et de l'emplacement du trafic à tel ou tel moment. Cumulés, les trajets identifiés par le système et recommandés aux conducteurs ont des effets de répartition : les habitants des quartiers vers lesquels le trafic est orienté subissent des niveaux plus élevés de bruit, d'émissions et d'embouteillages que ceux des quartiers vers lesquels le trafic n'est pas orienté. Des questions se posent donc quant aux responsabilités à l'égard des effets de répartition de ces systèmes d'optimisation, d'autant que les personnes, groupes et populations affectés ne sont pas consultés sur la redistribution des risques et des avantages qui en résulte : elle s'impose en l'absence de débat[47] ;

m. **capables de compliquer les actions collectives.** Comme ils peuvent fonctionner de manière très ciblée et personnalisée, et à l'échelle de toute une population d'utilisateurs, les systèmes d'optimisation par IA peuvent avoir à la fois des effets relativement mineurs au niveau individuel et un impact grave et significatif au niveau collectif et/ou sociétal. Il n'est pas difficile d'imaginer dans quelles circonstances un système d'optimisation par IA pourrait faire obstacle à l'action collective. Tous les individus gagneraient à coopérer, mais aucun ne le fait car l'impact sur chaque individu est trop faible pour justifier les efforts et les ressources nécessaires pour agir[48]. Prenons par exemple le problème du microciblage politique et de l'offre d'informations politiques trompeuses, inexactes ou douteuses à chaque électeur dans l'intention de l'inciter à voter pour tel ou tel candidat. Même si une personne est poussée à voter pour un candidat qu'elle n'aurait peut-être pas soutenu autrement, il est peu probable en pratique qu'elle soit suffisamment motivée pour porter plainte ou entamer une autre procédure judiciaire contre ceux qui ont diffusé les

44 Pariser 2012.
45 Yeung 2018a.
46 Yeung 2016.
47 Yeung 2017a.
48 Olsen 1965.

informations. Pourtant, les effets se font sentir au niveau collectif/d'une population, et représentent donc un danger réel et potentiellement grave pour l'intégrité des élections démocratiques et pour les processus démocratiques en général[49]. En d'autres termes, l'un des défis nouveaux et singuliers posés par les systèmes d'IA tient à leur capacité à fonctionner de manière très ciblée et personnalisée, mais aussi en temps réel et à l'échelle d'une population, fonctionnement qui peut faire peser de sérieux dangers sur la société mais contre lequel chaque personne concernée, prise individuellement, est très peu susceptible de lutter.

1.4 Incidences envisagées sous l'angle des droits de l'homme

Signe de leur importance, les incidences de l'IA sur les droits de l'homme sont abordées dans différents rapports et enquêtes commandés et produits par un nombre croissant d'organisations de la société civile, et intéressent de plus en plus les universitaires travaillant sur l'« éthique de l'IA[50] ». Cela englobe les travaux du Conseil de l'Europe, dont son étude sur les dimensions des droits humains dans les techniques de traitement automatisé des données et leurs éventuelles implications réglementaires (MSI-NET) (ci-après : « étude Wagner[51] »). L'étude Wagner pointe des exemples de systèmes de prise de décision algorithmique, actuellement appliqués, qui peuvent violer ou entraver l'exercice des « droits qui sont les plus manifestement concernés et appartiennent déjà à des degrés divers au débat public[52] ». Ce sont notamment :

- le droit à un procès équitable et les garanties d'une procédure régulière (article 6[53])
- la vie privée et la protection des données (article 8[54])
- la liberté d'expression (article 10)
- la liberté d'association (article 11[55])
- le droit à un recours effectif (article 13[56])

---

49 UK Information Commissioner's Office 2018. UK House of Commons Digital Culture Media and Sports Committee 2019.

50 Voir par exemple Amnesty International 2017 ; Access Now 2018 ; Australian Human Rights Commission 2018 ; Cath 2017 ; Hildebrandt 2015 ; Bureau exécutif du président des États-Unis 2016 ; The Montreal Declaration for Responsible AI 2017 ; The Toronto Declaration: Protecting the Rights to Equality and Non-Discrimination in Machine Learning Systems 2018 ; Latonero 2019 ; Mantelero 2018 ; Raso et al. 2018 ; Risse 2018 ; Rouvroy 2016 ; Assemblée générale des Nations Unies 2018 ; Mantalero 2019 ; Nuffield Foundation and Leverhulme Centre for the Future of Intelligence 2019 ; Groupe d'experts de haut niveau de l'UE sur l'intelligence artificielle 2019a.

51 L'étude Wagner s'intéresse avant tout aux incidences sur les droits de l'homme des systèmes de prise de décision algorithmique touchant le grand public. Elle identifie plusieurs problèmes de droits de l'homme suscités par le rôle croissant des algorithmes dans les décisions, et observe que ces problèmes ne peuvent que s'aggraver, les systèmes associés étant de plus en plus complexes et interagissant entre eux « d'une manière de plus en plus impénétrable pour l'esprit humain ». Étude Wagner 2017 : 5.

52 Étude Wagner 2017 : 32.

53 Voir le chapitre 2.1.1(a).

54 Voir le chapitre 2.1.1(b).

55 Bien que les sites internet et de réseaux sociaux aient accru la capacité des individus à exercer leur liberté d'association (article 11 CEDH), le tri et le profilage automatiques des protestataires en ligne peuvent aussi nuire à cette liberté. Étude Wagner 2017 : 23-24.

56 L'article 13 CEDH impose aux États de s'assurer que les personnes ont accès à des procédures judiciaires ou autres à même de statuer en toute impartialité sur leurs allégations de violations de droits de l'homme, y compris en ligne, dont des mécanismes non judiciaires effectifs, et de veiller à ce que les acteurs du secteur privé respectent ces droits en mettant en place des mécanismes de réclamation effectifs qui remédient rapidement aux griefs des personnes. Or, l'opacité des processus décisionnels automatisés limite la capacité

- l'interdiction de la discrimination (article 14[57])
- le droit à des élections libres (Protocole n° 1, article 3[58]).

Cependant, comme le conclut le rapport du Rathenau Instituut commandé par l'Assemblée parlementaire du Conseil de l'Europe,

> Malgré […] l'ampleur de l'impact des technologies numériques sur les droits de l'homme, ce thème crucial n'a jusqu'à présent suscité que peu d'attention et n'a pratiquement pas fait l'objet d'un débat politique et public sur le fond. Aussi assistons-nous actuellement à une grave dégradation de ces droits. Il faut donc que le débat sur les droits de l'homme, qui accuse un retard considérable par rapport à l'évolution technologique, s'intensifie au plus vite[59].

La présente étude prolonge l'étude Wagner en examinant les incidences possibles des technologies numériques avancées sur la notion de responsabilité. Le chapitre 2 commence par identifier et examiner les risques de l'IA pour les individus et la société. Il le fait sous l'angle des droits de l'homme, c'est-à-dire en se concentrant sur la manière dont ces technologies pourraient saper en pratique et *structurellement*, à notre époque d'omniprésence des technologies d'IA, la possibilité d'exercer certains droits de l'homme et libertés fondamentales – sans entreprendre d'analyse détaillée de certaines applications d'IA pouvant nuire à certains droits et libertés. Ces impacts structurels sont examinés sous deux aspects : premièrement, les menaces que les systèmes de prise de décision algorithmique font peser sur une série de droits[60] ; deuxièmement, les impacts collectifs plus larges des technologies d'IA (y compris celles intégrées aux systèmes de prise de décision algorithmique, mais pas uniquement), dont quelques-uns seulement peuvent être aisément décrits à l'aide du vocabulaire existant des droits de l'homme. Sur la durée, ces effets négatifs plus larges pourraient menacer les bases sociotechniques que la notion même de droits de l'homme suppose, et qui lui servent d'ancrage.

---

des individus à exercer leur droit à un recours effectif, et l'utilisation croissante de processus de traitement automatisé des réclamations est « extrêmement préoccupante », car il n'est pas certain que ces processus constituent un recours effectif (étude Wagner 2017, 27).

57 Voir le chapitre 2.1.1 d).

58 L'article 3 du Protocole n° 1 à la CEDH impose aux États d'assurer à chacun la libre expression de son opinion en organisant à intervalles raisonnables des élections libres, au scrutin secret. Cependant, la montée des réseaux sociaux et le recours à des systèmes de recommandations automatiques peuvent servir des fins de manipulation politique et menacer le droit à des élections libres (étude Wagner 2017, 33-36).

59 Van Est et Gerritsen 2017 : 46.

60 Plusieurs de ces droits sont examinés dans l'étude Wagner 2017.

Chapitre 2. Risques potentiels et avérés, dommages et autres effets négatifs des technologies numériques avancées

D'après de nombreux observateurs, les avancées des technologies numériques en réseau, dont ce qu'on appelle actuellement l'intelligence artificielle (« IA »), sont en train de susciter une « Nouvelle révolution industrielle » qui va modifier en profondeur tous les aspects de la vie sociale, avec une ampleur aussi perturbatrice et déstabilisante que celle de la Première[61]. Avant d'examiner les risques potentiels et avérés associés à ces technologies émergentes, on gagnera à dépeindre brièvement le contexte économique et sociopolitique dans lequel s'inscrivent leur développement, leur mise en œuvre et leur adoption, ainsi que le contexte historique et notre expérience moderne de l'innovation scientifique et technologique.

On peut pour cela dresser un parallèle entre les effets sociétaux de la Première révolution industrielle et les effets prévisibles de la « Nouvelle révolution industrielle » que nous voyons poindre. La Révolution industrielle du XIX^{e} siècle, par exemple, a apporté d'innombrables avantages aux individus comme à la société, ainsi qu'une nette amélioration du niveau de vie et du bien-être individuel et collectif dans de nombreux pays ; cependant, elle a aussi eu des effets négatifs imprévus. D'une part, les premières formes de production industrielle ont directement nui à la santé et à la sécurité des personnes ; d'autre part, la combustion de carburants fossiles nécessaire à l'activité industrielle a entraîné un grave problème de changement climatique à l'échelle planétaire, que nous ne sommes pas encore parvenus à contenir. Les effets négatifs sur le climat des technologies à l'origine de la Révolution industrielle ne se sont fait sentir qu'au bout de plus d'un siècle, alors qu'il était trop tard pour y apporter des solutions efficaces. Il se peut que nos sociétés se trouvent aujourd'hui face au même dilemme. Identifier et anticiper les effets sociétaux négatifs des innovations technologiques suppose de relever deux défis : prédire leur succès et les applications qu'elles vont avoir, et anticiper leurs effets cumulés dans l'espace et dans le temps.

2.1 La montée des systèmes de prise de décision algorithmique (ADM)

Avec l'essor rapide et généralisé des appareils « intelligents », les systèmes informatiques utilisant des algorithmes d'apprentissage automatique appliquent désormais des prises de décision algorithmiques, destinées à exploiter (et souvent à monnayer) les données glanées en collectant systématiquement les traces numériques que les utilisateurs laissent en ligne. Au moyen de technologies numériques avancées (dont l'IA), ces systèmes produisent de nouvelles connaissances qui servent à orienter les décisions dans le monde réel. Beaucoup de ces systèmes reposent sur des techniques de profilage par les données, qui passent par la collecte systématique et massive de données auprès d'une population d'individus en vue de repérer des tendances et donc de prédire les goûts, les centres d'intérêt et les comportements d'individus et de groupes, souvent avec une très grande exactitude. Ces profils servent ensuite à trier les individus pour identifier les « candidats intéressants » dans le but de produire des « connaissances utiles », qui pourront servir aux auteurs du profilage (ou à leurs clients) pour orienter et automatiser les décisions au sujet des différents individus[62]. Ces systèmes sont largement utilisés par les commerçants, pour présenter leurs produits aux individus identifiés comme plus susceptibles de les acheter[63] ; par les acteurs et organisations politiques, pour

61 Boyd et Crawford 2013 ; Skilton et Hovsepian 2017.

62 Mayer-Schonenberg et Cukier 2013.

63 Draper et Turrow 2017 ; Gandy 1993.

adresser des messages ciblés et sur mesure aux individus identifiés comme plus susceptibles de se laisser convaincre[64] ; et de plus en plus par les autorités pénales, pour mesurer au moyen d'algorithmes le « risque » que certains individus représenteraient pour la sûreté publique afin de prendre des décisions sur leur privation de liberté (qu'il s'agisse de suspects ou de condamnés pour des infractions pénales[65]).

Dans ce contexte socio-économique sont nées des inquiétudes quant aux *effets sociétaux* des technologies numériques avancées (dont l'IA), et notamment du recours croissant au profilage par les données. L'attention s'est récemment portée sur l'utilisation des technologies de profilage par les réseaux sociaux et les autres sites de diffusion de contenus, et sur ses profondes répercussions sur le droit à la liberté d'expression et d'information protégé par l'article 10, en particulier depuis le scandale Cambridge Analytica – les données de millions de profils Facebook auraient été collectées illégalement dans le but d'adresser à des individus des messages politiques très finement ciblés et d'influencer leur vote[66]. Les réflexions qui suivent, cependant, ne s'attardent pas sur les applications des prises de décisions algorithmique dans des domaines spécifiques, mais s'interrogent plus généralement sur les menaces que ces systèmes font peser sur certains droits de l'homme.

2.1.1 Quels sont les droits compromis par les systèmes d'ADM ?

Le recours à la prise de décision algorithmique (ci-après : « ADM », pour *algorithmic decision-making*) menace intrinsèquement plusieurs droits, dont les suivants.

(a) Droit à un procès équitable et à une procédure régulière (article 6)

Beaucoup de systèmes d'ADM, dans un large éventail de contextes, utilisent le profilage par les données pour créer les profils numériques d'individus et de groupes, les trier et les classer par catégories afin d'aider à la prise de décision. Ce profilage, lorsqu'il sert à automatiser et à orienter des décisions qui ont un fort impact sur des intérêts significatifs et sur les droits individuels, peut avoir des conséquences graves. Et l'individu concerné n'a quasiment aucune possibilité pratique d'influencer, de contester ou de remettre en cause la décision et/ou le raisonnement sur lequel elle se fonde, ou encore la qualité et l'intégrité des données utilisées[67]. Le droit à un procès équitable (protégé par l'article 6) englobe une série de droits procéduraux spécifiques[68], dont le droit de chacun à connaître les motifs des décisions qui lui sont défavorables ; or, de par leurs capacités et leur configuration, les systèmes d'ADM utilisés pour orienter la prise de décision ne peuvent pas toujours livrer d'explications valables en termes intelligibles pour la personne affectée, ni même pour les développeurs des algorithmes eux-mêmes (dans le cas de réseaux neuronaux utilisant la rétropropagation[69]). Ces inquiétudes sont encore renforcées par l'opacité de ces systèmes, due à leur complexité technique, à la difficulté d'estimer la qualité et la provenance des données ayant servi à entraîner le modèle[70] ou encore au secret commercial, qui fait de l'algorithme une propriété intellectuelle non

64 Gorton 2016.
65 Oswald et al. 2018 ; Ferguson 2016.
66 UK House of Commons, Digital Culture Media and Sport 2019.
67 Hildebrandt 2015 ; Hildebrandt et Gutwirth 2008.
68 Galligan 1997.
69 Weller 2017 ; Matthias 2004 ; Burrell 2016.
70 Lohr et al. 2019.

publiable[71] – secret typiquement défendu par les propriétaires car il empêcherait les utilisateurs de « jouer avec le système[72] ». Par conséquent, de tels systèmes risquent de violer les garanties de procédure régulière protégées par l'article 6 (dont la présomption d'innocence), en particulier lorsque les conséquences pour l'individu affecté sont graves et limitent son horizon[73]. Point particulièrement préoccupant, les systèmes d'IA sont de plus en plus utilisés dans des contextes pénaux, pour orienter les décisions sur les peines et les privations de liberté, avant tout aux États-Unis mais aussi dans d'autres pays (dont le Royaume-Uni[74]). De plus, comme l'observe Hildebrandt, nous avons développé une résistance à l'égard de l'idée que les résultats d'un outil d'IA concernant des suspects potentiels puissent être incorrects, incomplets ou même sans intérêt[75].

(b) Droit à la liberté d'expression (article 10)

Le profilage algorithmique pourrait fortement affecter le droit à la liberté d'expression, protégé par l'article 10 et qui comprend le droit de recevoir et de diffuser des informations, vu la puissante influence qu'exercent aujourd'hui les plates-formes numériques mondiales sur notre environnement informationnel au niveau individuel comme sociétal. Par exemple, les moteurs de recherche jouent un rôle de filtre crucial pour les personnes qui souhaitent rechercher, recevoir ou partager des informations, puisque les contenus non indexés ou n'apparaissant pas en tête des résultats ont moins de chances de toucher un large public – à supposer que quelqu'un les consulte. Or, les algorithmes de recherche sont sciemment conçus pour servir les intérêts commerciaux de leurs propriétaires, et favorisent donc inévitablement certains types de contenus ou de prestataires. Ce sont le plus souvent des algorithmes, et non des êtres humains, qui décident comment traiter, hiérarchiser, diffuser et supprimer les contenus de tiers sur les plates-formes en ligne, y compris en période de campagne politique ou électorale. De telles pratiques touchent non seulement le droit individuel à la liberté d'expression, mais aussi l'objet même de l'article 10 : créer un environnement favorable à un débat public pluraliste et accessible à tous[76].

Par ailleurs, la pression s'est accrue sur les plates-formes en ligne pour qu'elles luttent activement contre le discours de haine en détectant et en supprimant automatiquement les contenus illégaux, en particulier depuis la diffusion en direct sur certains réseaux sociaux de l'agression de civils par un terroriste à Christchurch, début 2019. Aux termes de l'article 10.2, toute ingérence dans la liberté d'expression, englobant donc les systèmes algorithmiques qui

---

71 Pasquale 2015.

72 Bennett-Moses et de Koker 2017.

73 Davidow 2016.

74 Ces applications pèsent non seulement sur les droits de l'article 6, mais aussi sur ceux des articles 5 (droit à la liberté et à la sûreté) et 14 (interdiction de discrimination).

75 Hildebrandt 2016. Pour Hildebrandt, le principe de l'« égalité des armes » inscrit à l'article 6 demande à être repensé dès lors que les procureurs, les juges ou les avocats sont incapables de vérifier comment l'outil d'IA de la police est parvenu à ses conclusions ; pour permettre un véritable examen, ces outils d'IA devraient obligatoirement garder trace de leurs activités, de leurs résultats, de leurs objectifs et de la façon dont ils ont atteint leurs résultats. Le Rathenau Instituut partage cette opinion, et a suggéré que le Conseil de l'Europe envisage d'établir un cadre normatif minimal dont les « tribunaux » (désignant ici toutes les instances décisionnaires du système judiciaire, en particulier celles que décident de la privation de liberté des individus dans le cadre de la justice pénale) devraient tenir compte lorsqu'ils recourent à l'IA, afin de prévenir dans la mesure du possible l'élaboration de cadres par les États eux-mêmes, qui présenterait le risque d'offrir des niveaux variables de protection au sens de l'article 6 CEDH (Van Est et Gerritsen, 2017 : 42-43).

76 Voir Assemblée générale des Nations Unies 2018.

bloquent l'accès à des contenus en les filtrant ou en les supprimant, doit être prévue par la loi, poursuivre l'un des buts légitimes énoncés à l'article 10.2 et être nécessaire dans une société démocratique[77]. Par conséquent, l'usage généralisé d'algorithmes pour filtrer et supprimer des contenus, notamment sur les sites de réseaux sociaux, pose aussi problème pour l'État de droit puisqu'il n'est pas certain que cet usage soit légal, légitime et proportionné, étant donné notamment que les plates-formes en ligne sont souvent confrontées à un cadre législatif ambigu qui les incite à prendre les devants et à retirer des contenus sans base juridique claire. Bien que leurs intentions soient louables, il existe un manque de transparence sur le processus et sur les critères adoptés pour décider quels contenus sont « extrémistes » ou « clairement illégaux[78] ». Ces mesures, qui risquent de constituer une ingérence excessive dans le droit à la liberté d'expression, laissent penser que les États « sous-traitent » des missions de maintien de l'ordre à des entreprises privées. Certains régimes juridiques nationaux imposent aux intermédiaires du numérique de limiter l'accès à des contenus sur la base de notions vagues, comme l'« extrémisme » ; ils les obligeant donc à surveiller toutes les communications en ligne pour y repérer les contenus illégaux, violant ainsi le principe établi selon lequel aucun intermédiaire ne devrait être tenu d'appliquer une surveillance générale sous peine de créer un « effet tétanisant » sur la liberté d'expression[79]. En outre, la capacité des plates-formes à décider elles-mêmes ce qu'il convient de supprimer pour « extrémisme » soulève des inquiétudes quant au processus : le choix des outils et des mesures appartient à des opérateurs privés qui risquent, en l'absence d'une supervision éclairée et effective de la part de l'État, d'outrepasser les limites de la loi et de la constitution en violation des normes de l'État de droit[80].

Bien que l'impératif d'agir résolument contre la propagation des messages de haine et l'incitation à la haine raciale soit indiscutable, de telles pratiques soulèvent de vives inquiétudes quant à la légalité des ingérences dans la liberté d'expression. Les contenus extrémistes et les incitations à la violence sont souvent difficiles à identifier, même pour une personne dûment formée, en raison de la complexité des facteurs qui entrent en jeu – dont le contexte culturel et l'humour. Les algorithmes ne savent aujourd'hui détecter ni l'ironie, ni l'analyse critique. Ainsi, les filtres algorithmiques destinés à éliminer les contenus nocifs risquent fort d'avoir les mailles trop fines et de bloquer des propos non seulement inoffensifs,

---

77 Conformément à la jurisprudence de la Cour européenne des droits de l'homme, toute restriction à la liberté d'expression doit obéir à un « besoin social impérieux » et être proportionnée au(x) but(s) légitime(s) poursuivi(s). Voir *Yildirim c. Turquie*, 18 mars 2013, n° 3111/10.

78 Voir Menn et Volz 2017.

79 Ce principe est inscrit dans le droit de l'UE et dans les lignes directrices pertinentes du Conseil de l'Europe, dont la récente Recommandation CM/Rec(2018)2. Voir aussi Assemblée générale des Nations Unies (2018). Plusieurs États ont adopté des lois ou engagé des réformes législatives pour contrer la propagation des contenus nocifs en ligne. L'Allemagne, par exemple, a adopté en 2017 une loi sur les réseaux (« NetzDG »). Elle oblige les plates-formes en ligne comptant plus de deux millions d'utilisateurs enregistrés en Allemagne à retirer les contenus « manifestement illégaux », dans les 24 heures suivant une notification ou une plainte pour les contenus contraires à certains articles du Code pénal allemand (concernant par exemple l'holocauste ou le discours de haine) et dans les sept jours pour tous les autres contenus « illégaux ». Les contrevenants encourent jusqu'à 50 millions d'euros d'amende. Cette loi cherche aussi à rendre les sites plus responsables en imposant une plus grande transparence et d'importantes obligations de signalement. Ses conséquences restrictives sur la liberté d'expression ont suscité de vives critiques ; voir par exemple Access Now 2018 : 22. Le Royaume-Uni a récemment publié un Livre blanc sur les préjudices en ligne, *Online Harms White Paper*. Il crée une obligation de diligence en ligne, pour que les entreprises assument davantage de responsabilités à l'égard de la sécurité de leurs usagers et luttent contre les préjudices causés par les contenus ou activités utilisant leurs services – obligation qui devrait être mise en œuvre par une instance de régulation indépendante (UK Government 2019).

80 Voir Étude Wagner 2017, 22.

mais aussi susceptibles de contribuer positivement au débat public. Par ailleurs, la capacité des plates-formes médiatiques à diffuser des messages en temps réel et dans le monde entier multiplie la portée, l'étendue et donc l'impact des propos nocifs. Le recours au filtrage automatique des contenus en ligne met en lumière les considérables problèmes de responsabilité suscités par la présence accrue des systèmes algorithmiques dans notre vie quotidienne. Comme les approches automatiques offrent une rapidité, une couverture géographique et une efficacité hors de portée des humains, les plates-formes numériques affirment que le contrôle par des êtres humains serait voué à l'échec, générant un « vide de responsabilité » qu'on ne saurait leur demander de combler[81].

(c) Droit à la vie privée et à la protection des données (article 8)

Le droit au respect de la vie privée et familiale et celui à la protection des données personnelles, protégés par l'article 8, sont sous pression comme jamais auparavant depuis que les algorithmes facilitent la collecte et le réemploi à d'autres fins de vastes volumes de données, dont des données personnelles issues du pistage des utilisateurs et susceptibles de générer de nouveaux savoirs, avec des résultats totalement imprévisibles pour la personne concernée[82]. Comme l'observe l'étude Wagner, l'utilisation des données personnelles pour établir le profil de chaque individu, et le réemploi ultérieur de ces données, menacent le droit de chacun à « l'autodétermination de son information[83] », d'autant plus que (comme relevé au chapitre 2.1) même des données anodines et triviales, issues des traces numériques des utilisateurs, peuvent être fusionnées avec d'autres ensembles de données et exploitées de manière à livrer des connaissances dont on peut déduire, avec une très faible marge d'erreur, des informations personnelles très intimes[84].

Bien que les régimes actuels de protection des données (dont la Convention 108 modernisée) constituent une garantie importante, offrant aux intéressés une série de « droits à la protection des données[85] » destinés à les protéger des collectes et traitements de données inutiles et illicites, ils risquent de ne pas s'avérer assez complets et effectifs dans la pratique contre les applications intrusives du profilage.

(d) Interdiction de discrimination dans l'exercice des droits et des libertés (article 14)

Le potentiel de partis pris et de discrimination associé aux techniques d'apprentissage automatique a suscité une attention considérable, de la part des responsables politiques comme des chercheurs en IA. Les traitements injustes ou illégaux redoutés ont un lien direct avec l'article 14 CEDH, qui dispose que la jouissance des droits et libertés reconnus dans la Convention doit être assurée « sans distinction aucune, fondée notamment sur le sexe, la race,

---

81 Voir les réflexions sur le prétendu « problème de contrôle », chapitre 3.2.2 ci-dessous.

82 Voir par exemple les tensions entre la concurrence entre services en ligne d'une part, la vie privée des consommateurs d'autre part (Oxera 2018).

83 Étude Wagner 2017 :14.

84 Kosminski et al. 2015.

85 Les nouveaux droits instaurés par la Convention 108 récemment modernisée sont entre autres : le droit de chacun de ne pas être soumis à une décision l'affectant de manière significative qui serait prise uniquement sur le fondement d'un traitement automatisé de données, sans prise en compte de son point de vue ; le droit d'obtenir, à sa demande, connaissance du raisonnement qui sous-tend le traitement de données lorsque les résultats de ce traitement lui sont appliqués ; et le droit de s'opposer, pour des raisons tenant à sa situation, à ce que des données à caractère personnel le concernant fassent l'objet d'un traitement, à moins que le responsable du traitement ne démontre que des motifs légitimes justifient le traitement et prévalent sur ses intérêts, ou sur ses droits et libertés fondamentales. Convention 108 modernisée, article 9.

la couleur, la langue, la religion, les opinions politiques ou toutes autres opinions, l'origine nationale ou sociale, l'appartenance à une minorité nationale, la fortune, la naissance ou toute autre situation[86] ». En effet, les techniques d'apprentissage automatique peuvent livrer des résultats biaisés pour de nombreuses raisons, puisque peuvent s'y glisser les partis pris des développeurs des algorithmes, ceux intégrés au modèle sur lequel reposent les systèmes, ceux inhérents aux ensembles de données utilisés pour entraîner les modèles, ou encore les partis pris qui se créent une fois que ces systèmes fonctionnent dans des environnements réels[87]. Ce phénomène peut non seulement entraîner des discriminations et des décisions erronées, mais aussi créer d'importants préjudices, en aboutissant à des décisions systématiquement défavorables à des groupes sociaux déjà traditionnellement défavorisés (et aux personnes membres de ces groupes), avec pour effet de renforcer et d'aggraver les discriminations et les désavantages structurels même si les concepteurs du système n'en avaient pas l'intention[88]. Ces préoccupations sont particulièrement vives face à l'utilisation, dans le système pénal étasunien, de techniques d'apprentissage automatique pour orienter les décisions relatives aux peines et à la privation de liberté, car ces techniques défavoriseraient nettement la minorité noire et les autres minorités raciales[89]. En réaction, de plus en plus de travaux s'intéressent actuellement aux approches techniques permettant de contrer ce phénomène discriminatoire[90].

2.1.2 Problèmes sociétaux associés au profilage par les données

Les applications contemporaines des technologies de profilage par les données peuvent aussi saper d'importants intérêts et valeurs collectifs, dont seuls certains entrent actuellement dans le champ de la protection des droits de l'homme. L'intérêt de ces technologies réside pour une bonne part dans leur capacité à trier les individus et les groupes, à automatiser les prises de décisions et à permettre des interventions personnalisées et prédictives à l'échelle de toute une population. Les pratiques ci-dessous peuvent créer des problèmes sociétaux significatifs, et pourtant souvent ignorés dans le débat public et académique.

a. Surveillance très fine à l'échelle d'une population

Comme le profilage par les données, qui consiste à établir le profil des individus et des groupes pour en déduire leurs goûts et leurs centres d'intérêt, requiert la collecte de données très fines auprès des membres de toute une population (c'est-à-dire à grande échelle[91]), il suppose le recours à une surveillance de masse, souvent à la fois très intrusive et pratiquement invisible. On voit aisément quelles menaces de telles pratiques représentent pour la vie privée et pour le

86 Le Protocole n° 12 à la CEDH, article 1, dispose que « la jouissance de tout droit prévu par la loi doit être assurée, sans discrimination aucune, fondée notamment sur le sexe, la race, la couleur, la langue, la religion, les opinions politiques ou toutes autres opinions, l'origine nationale ou sociale, l'appartenance à une minorité nationale, la fortune, la naissance ou toute autre situation ». Voir aussi la Charte des droits fondamentaux de l'Union européenne, article 21.

87 Veale et Binns 2017.

88 Barocas et Selbst 2016 ; étude Wagner 2017 : 27-28.

89 Angwin et al. 2016. Mais voir aussi Dieterich et al. 2016.

90 Voir plus loin, chapitre 3.7.1. Pour citer David Kaye, Rapporteur spécial des Nations Unies, « L'élimination des problèmes de discrimination dans les systèmes d'IA est un défi existentiel pour les entreprises comme pour les pouvoirs publics ; faute de supprimer les éléments discriminatoires et leurs effets, ces outils sont non seulement inefficaces, mais dangereux ». Assemblée générale des Nations Unies 2018, 21.

91 Comme l'affirme la Commission de la culture, de la science, de l'éducation et des médias de l'Assemblée parlementaire du Conseil de l'Europe, « le modèle économique d'internet repose avant tout sur la surveillance de masse ». Conseil de l'Europe 2017, par. 18.

droit à la protection des données au niveau individuel (voir plus haut) ; cependant, elles compromettent aussi sérieusement la valeur de la vie privée *au niveau collectif*, sapant les conditions sociales qui rendent la vie privée individuelle possible et sans lesquelles elle ne peut exister. Comme l'observe l'Assemblée parlementaire du Conseil de l'Europe[92] :

> Étant donné qu'aujourd'hui, de nombreuses technologies sont capables de fonctionner à distance, nous sommes une grande majorité à ignorer jusqu'à l'existence de cette surveillance de masse et sommes donc quelque peu démunis face à ce phénomène, car les possibilités d'y échapper sont rares. Cette évolution silencieuse et son impact sur la société et sur les droits de l'homme n'ont suscité jusqu'à présent que peu d'attention dans le débat politique et public. […] L'effet cumulatif de la surveillance de masse a suscité peu de débats. Des applications et des incidents spécifiques ont donné lieu à des « mini-débats », l'issue de chacun d'eux étant un exercice d'équilibriste qui favorise principalement la sécurité et les intérêts économiques nationaux. Il ressort toutefois de ces débats que le droit des personnes au respect de la vie privée et à l'anonymat disparaît progressivement mais régulièrement[93].

Ces risques se sont amplifiés depuis les récentes avancées de l'IA, qui ont permis l'émergence de puissantes applications biométriques pouvant être utilisées à des fins d'identification, menaçant sérieusement plusieurs droits de l'homme, dont ceux protégés par l'article 8. En Chine par exemple, des technologies de reconnaissance faciale intégrant l'IA ont été installées dans le métro de Beijing pour permettre d'identifier les traits des voyageurs et de pister leurs déplacements. De telles technologies ont déjà été déployées dans des gares ferroviaires, lors d'un concert de pop pour localiser un suspect en fuite, et même dans des établissements scolaires pour repérer les élèves distraits et avertir automatiquement l'enseignant[94]. Il n'est guère difficile d'imaginer comment un régime répressif pourrait employer les technologies de lecture sur les lèvres intégrant l'IA récemment développées par DeepMind (et présentées comme plus performantes que les professionnels de cette technique[95]), ce qui fait craindre pour le simple droit à être laissé tranquille, et l'effet tétanisant que de telles technologies pourrait avoir sur la liberté d'expression, l'épanouissement personnel et la liberté démocratique, en particulier si des États les appliquent pour repérer et arrêter des dissidents politiques[96]. De tels outils, associés à des technologies de profilage par les données permettant de fusionner et d'exploiter des données à première vue anodines et triviales pour révéler des

---

92 Conseil de l'Europe 2017, par. 60-61.

93 L'étude Wagner attire aussi l'attention sur les risques suscités par l'agrégation des données et la génération de données nouvelles, qui « peuvent ensuite être exploitées au moyen d'algorithmes, ce qui crée un risque de surveillance à grande échelle (« dataveillance ») aussi bien par des entités privées que par des gouvernements » ; point de vue repris par le Conseil des droits de l'homme des Nations Unies (22 mars 2017). Étude Wagner 2017 : 17. Comme l'observe le Rathenau Instituut, « de nos jours, la surveillance qu'exercent des États ou des entreprises par le biais d'internet ou de l'internet des objets implique par définition le traitement de données à caractère personnel. Les chercheurs essaient encore d'appréhender toute l'étendue des conséquences préjudiciables de cette surveillance sur la vie des personnes. Et ses effets connus ne sont pas rassurants. Non seulement cette surveillance a sur la parole un effet dissuasif […] mais elle entraîne également des modifications comportementales. Par exemple, les individus se sentant surveillés se conforment aux normes qu'ils perçoivent comme étant celles du groupe. Cet effet conformiste peut même se produire à leur insu (Kaminski & Witnov, 2015). Les États et les entreprises se renforcent mutuellement dans leurs activités de surveillance, sous l'influence du complexe *surveillance-innovation* (Cohen 2016) ». Van Est et Gerritsen, 2017 : 20.

94 Cowley 2018.

95 Hutson 2018.

96 Donahoe 2016.

caractéristiques très intimes (comme l'orientation sexuelle[97]), peuvent conférer un pouvoir sans précédent aux gouvernements, qu'ils soient libéraux ou répressifs, et donc faire peser une terrible menace sur l'exercice de l'ensemble des droits de l'homme et des libertés fondamentales.

b. Personnalisation à l'échelle d'une population

Pour ceux qui cherchent à établir des profils, les attraits des technologies de profilage sont assez évidents : elles permettent de trier et de cibler automatiquement les individus intéressants afin de *personnaliser* la façon dont ils sont traités. Ces techniques peuvent s'appliquer à grande échelle, tout en permettant de réajuster et de reconfigurer les offres personnalisées en temps réel en réaction au comportement de l'usager[98]. L'application, par les prestataires numériques, d'une personnalisation à l'échelle de toute une population peut affecter en profondeur la solidarité sociale et le sens de la communauté. Prenons, par exemple, la pratique de la « tarification personnalisée » rendue possible par le profilage par les données et la montée de la vente en ligne. Sous le capitalisme industriel, les biens étaient produits en masse puis vendus à des détaillants, et généralement mis à la disposition des consommateurs en divers emplacements géographiques, dans des conditions et à des prix qui s'appliquaient à tous les clients entrant dans le magasin à un moment donné. À l'inverse, le profilage par les données permet aujourd'hui d'offrir aux clients potentiels des biens et services à des prix « personnalisés » – car chaque client ne voit que sa propre « vitrine numérique » individualisée, sans accès aux prix et aux offres proposés aux autres – dont le montant peut être fixé en fonction du « prix maximum acceptable » pour chaque individu, optimisant ainsi les revenus du détaillant[99]. Bien que cette discrimination volontaire ne soit pas nécessairement illégale, car elle pourrait ne pas constituer une discrimination directe ou indirecte fondée sur des motifs couverts par le droit contemporain en matière d'égalité, de telles pratiques s'écartent nettement des modes de tarification qui prévalaient avant le règne du numérique et des données et pourraient, si elles se généralisent, saper sérieusement la cohésion et la solidarité sociales[100].

c. Manipulation de toute une population

La personnalisation des environnements informationnels au moyen du profilage par les données crée de nouvelles manières de manipuler les individus, subtiles mais extrêmement efficaces[101]. Au niveau individuel, la manipulation peut menacer l'autonomie personnelle et le droit émergent à la souveraineté cognitive[102] ; mais comme l'a si bien illustré le récent

97 Kosinski et al., 2013.

98 Yeung 2016.

99 Townley et al. 2017 ; Miller 2014.

100 Yeung 2018a. La Commission européenne a étudié la prévalence des pratiques de personnalisation en ligne (Commission européenne 2018b). Au Royaume-Uni, la Competition and Markets Authority (CMA) a commandé à des économistes une étude sur l'usage des algorithmes de tarification et les problèmes de concurrence qui pourraient en découler, dont l'entente sur les prix et les prix personnalisés (UK Competition and Markets Authority 2018).

101 Yeung 2016. Par exemple, le Norwegian Consumer Council a analysé dans une récente étude un échantillon de paramètres sur Facebook, Google et Windows 10. Il montre que les paramètres par défaut et les structures, techniques et caractéristiques dissimulés de l'interface, conçus pour manipuler les utilisateurs, servent à orienter ces derniers vers des options qui constituent une intrusion dans leur vie privée (ForbrukerRadet 2018).

102 Certains chercheurs plaident pour la reconnaissance d'un nouveau droit individuel à la « souveraineté cognitive », afin d'offrir une protection fondée sur les droits contre les formes de tromperie et de

scandale Cambridge Analytica – avant les élections présidentielles étasuniennes en 2016 et le référendum sur le Brexit –, le microciblage politique visant à manipuler les électeurs, qui passe parfois par des robots fonctionnant automatiquement sur les réseaux sociaux, peut menacer le droit à la liberté d'expression et d'information (article 10) et ébranler fortement les fondations mêmes des ordres démocratiques en pervertissant le droit à des élections libres (Protocole n° 1 à la CEDH, article 3[103]). Les manipulations rendues possibles par les technologies numériques dites « persuasives » peuvent être vues comme des ingérences dans les droits protégés par les articles 8 et 10. En effet, elles peuvent être automatiquement configurées (et reconfigurées en continu) pour ajuster l'environnement informationnel proposé aux individus, à l'aide d'un profilage par les données qui permet de prédire à grande échelle (et souvent très précisément) le comportement de chaque personne, ses centres d'intérêt, ses goûts et ses fragilités. De telles applications peuvent être employées pour manipuler et tromper les individus, constituant une ingérence attentatoire à leur vie privée dans leurs informations et dans leurs décisions[104].

Les possibilités de pratiques manipulatrices ont été exacerbées par la récente émergence de puissantes applications d'IA simulant des traits humains (dont la voix de synthèse, les représentations visuelles du comportement humain et les robots capables d'interagir avec des êtres humains en paraissant sensibles aux émotions), de manière si fidèle et précise qu'il peut s'avérer très difficile pour les intéressés de comprendre que ces traits sont artificiels. Ces technologies devraient intéresser ceux qui cherchent à tromper et à manipuler autrui. Par exemple, certains chercheurs prédisent déjà que les voix de synthèse perfectionnées seront utilisées pour recueillir des informations par téléphone à des fins frauduleuses. Si de telles attaques deviennent courantes et si les individus ciblés ne peuvent les détecter facilement, elles risquent de menacer sérieusement le droit à la liberté et à la sûreté (article 5) ainsi que la sécurité collective et le respect de l'État de droit sur lesquels reposent la liberté et la sûreté de tous et de chacun. Il est aussi possible que ces technologies servent à saper l'intégrité du processus juridique. Comme l'observent Brundage et al. dans leur rapport sur l'IA malveillante,

> Aujourd'hui, les technologies d'enregistrement et d'authentification conservent encore une longueur d'avance sur les technologies de contrefaçon. La vidéo d'un crime peut constituer une preuve de poids, même si elle provient d'une source peu fiable. À l'avenir cependant, des faux de grande qualité utilisant l'IA pourraient bousculer l'axiome « voir c'est croire » traditionnellement associé aux preuves vidéo (et audio). Ils pourraient aussi encourager les accusés à nier les charges pesant sur eux, compte tenu de la facilité avec laquelle les preuves pourraient avoir été fabriquées. Outre la diffusion croissante d'informations trompeuses, l'écriture et la publication de fausses actualités pourraient être automatisées, comme le sont aujourd'hui couramment les actualités financières et sportives. Plus les coûts de production et de diffusion de faux de grande qualité diminueront, plus la part des contenus multimédias de synthèse risque d'augmenter dans l'écosystème des médias et de l'information[105].

---

manipulation que le progrès des technologies numériques facilite de plus en plus et de garantir aux individus un niveau minimal de souveraineté sur leurs propres esprits (voir Bublitz 2013). Ce droit pourrait être isolé, mais aussi envisagé comme relevant de l'article 9.1 CEDH, qui affirme le droit à la liberté de pensée, de conscience et de religion.

103 Gorton 2016 ; Étude Wagner 2017 : 17. UK House of Commons, Digital Culture Media and Sport 2019.

104 Yeung 2016 ; Lanzing 2018 ; Conseil de l'Europe 2017.

105 Brundage et al. 2018 : 46.

d. Traitement des individus comme des objets et non comme des sujets

Bien qu'aux dires des entreprises de réseaux sociaux, la personnalisation des environnements informationnels aide à proposer des contenus « plus intéressants », le système socio-technologique sur lequel ces pratiques reposent présente deux caractéristiques qui tendent à traiter les individus comme des objets plutôt que comme des sujets. Premièrement, les individus ne sont pas triés selon une théorie rationnelle, mais sur la seule base des corrélations présentes dans les ensembles de données. Par conséquent, ces systèmes sont incapables d'expliquer *pourquoi* ils ont traité tel individu de telle manière. Deuxièmement, leur logique sous-jacente et leurs opérations de traitement sont aussi obscures que complexes, ce qui les rend incompréhensibles dans la pratique et parfois même sur le plan technique (comme nous l'avons vu plus haut). Autrement dit, comme de nombreux systèmes d'apprentissage automatique sont aujourd'hui conçus pour suivre et analyser les traces numériques de nos comportements quotidiens afin de capter, monétiser et optimiser de la valeur dans l'intérêt du propriétaire du système, comprendre *pourquoi* les individus adoptent tel ou tel comportement n'est pas leur première préoccupation. Rieder affirme que les applications commerciales de ces techniques de « big data » font de la réalité une lecture « intéressée[106] », opposée à la quête désintéressée du savoir qui caractérise la recherche scientifique dans le cadre universitaire[107]. De telles applications ont pour effet direct de traiter de plus en plus les êtres humains non comme des sujets, mais comme des objets à trier, filtrer, noter et évaluer à l'aide de systèmes technologiques, selon des modalités qui contrastent fortement avec le droit fondamental de chacun à la dignité et au respect – droit à l'origine de l'ensemble des droits de l'homme et des libertés fondamentales[108]. Comme l'explique le Groupe européen d'éthique (2018) :

> L'« optimisation » des processus sociaux entraînée par l'IA et basée sur des systèmes d'évaluation sociale que certains pays expérimentent, viole l'idée fondamentale d'égalité et de liberté de la même manière que les systèmes de castes, parce qu'elle crée « différentes catégories d'individus » là où il n'y a en réalité que différentes « caractéristiques » des individus. Comment empêcher ces attaques contre les systèmes démocratiques et cette utilisation de systèmes d'évaluation comme fondement de la domination de ceux qui ont accès à ces puissantes technologies ? […] La dignité humaine, en tant que fondement des droits de l'homme, implique qu'une intervention et une participation humaines significatives doivent être possibles pour ce qui concerne les hommes et leur environnement. Par conséquent, contrairement à l'automatisation de la production, il n'est pas approprié de gérer le sort des hommes et d'en décider de la même manière que nous gérons et décidons de ce qu'il advient des objets ou des données, même si c'est techniquement concevable. Une telle gestion

106 Rieder 2016.

107 Merton 1942.

108 L'utilisation de l'IA pour appliquer un profilage individuel dans le cadre du système pénal est particulièrement inquiétante. Comme l'observe l'institut AI Now, Axon a acquis deux entreprises de vision automatique et propose désormais des caméras-piétons gratuites à tous les services de police des États-Unis. « Le tournant d'Axon vers les méthodes de police prédictive – inspiré de l'usage de l'apprentissage profond par Wal-Mart et Google pour faire monter les ventes – soulève de nouvelles inquiétudes pour les libertés publiques. Au lieu des habitudes d'achat, ces systèmes scruteront des phénomènes beaucoup plus vagues et dépendants du contexte, comme les « activités suspectes ». Sous leur apparente neutralité technique, ils reposent sur des présupposés très subjectifs concernant les personnes ou les comportements à considérer comme suspects » (AI Now, 2017 : 25). Des individus deviennent ainsi « objets de soupçons » sur la base d'une analyse de données dépourvue de tout lien causal démontrable.

> « autonome » des êtres humains serait contraire à l'éthique et nuirait aux valeurs européennes fondamentales et profondément enracinées[109].

Parallèlement, les applications commerciales de l'IA à des fins de profilage se sont accompagnées d'expérimentations sur les individus, à l'échelle de populations entières, via la méthode des tests comparatifs A/B, sans qu'aucune institution ne surveille l'éthique de ces recherches conformément à la Déclaration d'Helsinki. Cette dernière énonce les principes éthiques applicables aux recherches sur des sujets humains[110]. L'usage courant et généralisé de telles pratiques reflète encore une fois l'idée que les utilisateurs ne seraient que des objets se prêtant bien aux expérimentations, si bien qu'il n'y aurait pas lieu d'appliquer les normes fondamentales et les mécanismes de contrôle institutionnels conçus pour préserver et protéger la dignité et les droits des individus. Comme l'écrit Julie Cohen, « nous, les citoyens, nous trouvons réduits au rang de matière première, extraite, troquée et exploitée dans ce curieux « espace commun privatisé » fait de données et de surveillance[111] ».

e. Résumé des menaces associées aux technologies de profilage par les données

Pris ensemble, les effets cumulés des pratiques ci-dessus justifient les préoccupations concernant le profilage exprimées en termes très forts par Korff dans son rapport pour le Conseil de l'Europe, « Utilisation d'internet et des services afférents, vie privée et protection des données : tendances, menaces et implications ». En effet, les systèmes de profilage paraissent infaillibles, objectifs, fiables et précis alors qu'ils génèrent inévitablement des erreurs (faux positifs ou faux négatifs) ou ont sur certains groupes des effets discriminatoires[112] que les individus ne peuvent pratiquement pas contester ; ce qui amène Korff à conclure :

> Le profilage risque sérieusement de nous entraîner dans un monde kafkaïen, dans lequel de puissantes entreprises et des organismes d'État prennent des décisions qui affectent significativement leurs clients et citoyens sans avoir la capacité, ou la volonté, d'expliquer les motifs de ces décisions, et dans lequel les intéressés n'ont accès à aucun recours effectif, qu'il soit individuel ou collectif. C'est pourquoi le problème du profilage est grave : il menace de saper les principes les plus élémentaires de l'État de droit et des relations entre la population et ceux qui ont le pouvoir dans une société démocratique[113].

Ces remarques nous alertent sur les impacts collectifs et cumulés des applications contemporaines des technologies fondées sur les données, qui pourraient à terme, si leur usage se généralise, éroder et déstabiliser gravement les bases sociales et morales nécessaires à l'épanouissement de sociétés démocratiques dans lesquelles les individus peuvent exercer leurs droits et leurs libertés.

2.2 Menaces et risques sociétaux collectifs générés par d'autres technologies d'IA

Bien que les préoccupations énumérées ci-dessus aient beaucoup à voir avec le profilage par les données, d'autres menaces sur les intérêts et valeurs collectifs ne proviennent pas du profilage individuel. Elles sont présentées ci-dessous.

---

109 Groupe européen d'éthique des sciences et des nouvelles technologies 2018 : 9-10.

110 Kramer et al. 2015 ; Tufecki 2015.

111 Powles 2015.

112 Korff et Browne 2013 : 6.

113 Korff et Browne 2013 : 21.

2.2.1 Attaques malveillantes, conception non éthique du système ou défaillance involontaire du système

Des craintes compréhensibles et fondées ont commencé à s'exprimer face aux problèmes de sécurité liés aux technologies d'IA, et notamment face aux effets catastrophiques que pourraient avoir des attaques contre des systèmes d'IA (empoisonnement de données, détournement malveillant de l'apprentissage automatique…) si elles parviennent à toucher des éléments essentiels à la sécurité. Même en l'absence d'intention malveillante, beaucoup craignent aussi que des pannes sur certaines technologies utilisant l'IA (comme les véhicules autonomes) ne nuisent sérieusement à la sécurité publique[114]. Pire encore, ces systèmes pourraient être conçus pour donner la priorité à certaines catégories de personnes par rapport à d'autres, ce que beaucoup considèrent comme contraire à l'éthique, voire illégal. Plus nos sociétés deviendront dépendantes des appareils connectés à internet et, plus généralement, des systèmes cyber-physiques (dont beaucoup comportent des éléments essentiels pour la sécurité), plus il sera crucial d'assurer la sécurité et l'innocuité de ces systèmes[115]. D'autant que les possibilités d'attaques se multiplient et se diversifient, non seulement contre les systèmes eux-mêmes, mais aussi au moyen de stratégies visant à exploiter les effets de réseau pour cibler et contacter des individus à grande échelle mais dans un relatif anonymat[116].

2.2.2 Perte d'un rapport humain authentique, réel et significatif

Outre les craintes évoquées ci-dessus concernant le recours aux technologies d'IA pour imiter les comportements humains, on constate une anxiété diffuse, mais très prégnante, devant le spectre d'une vie en commun de plus en plus « déshumanisée » et l'automatisation de tâches auparavant assurées par des êtres humains. Beaucoup redoutent que des valeurs et des qualités qui nous sont chères, comme les échanges humains réels, la véritable empathie, la compassion, le souci de l'autre, ne soient remplacées par l'implacable efficacité des services pilotés par intelligence artificielle. Ces craintes s'accentuent encore quand les technologies d'IA sont utilisées dans des milieux de soins (robots infirmiers, gardes d'enfants et autres assistants robotisés) ou lorsqu'elles risquent, comme les robots sexuels par exemple, de dépouiller nos sociétés de valeurs et de traits inhérents aux rapports humains authentiques et concrets, bien sûr imparfaits et risqués, mais sans lesquels l'expérience humaine n'aurait ni saveur ni sens[117]. De telles applications ont suscité des appels à s'assurer que leur conception et leur fonctionnement respectent la dignité des personnes concernées. Elles pourraient relever de la « protection de la vie privée et familiale » (article 8) et ont amené certains à plaider pour un « droit aux rapports humains authentiques[118] ».

---

114 « Toutes les technologies peuvent tomber en panne, et les systèmes autonomes n'y échapperont pas (considération pertinente pour décider s'il convient de créer des systèmes autonomes sur lesquels on ne peut reprendre la main) » (Royal Academy of Engineering 2009 : 3).

115 Thomas 2017a.

116 Brundage et al. 2018 ; ForbrukerRadet 2018.

117 Yearsley 2017.

118 Les préoccupations de ce type ont poussé l'Assemblée parlementaire du Conseil de l'Europe (APCE) à avancer que lorsque l'interaction et les rapports humains jouent un rôle central, comme dans l'éducation des enfants et les soins aux personnes âgées ou aux personnes handicapées, le « droit à de véritables rapports humains » pourrait avoir sa place. Voir Assemblée parlementaire du Conseil de l'Europe 2017, par. 65.

2.2.3 Effet dissuasif de la réutilisation des données

Autre source d'inquiétude, certaines personnes pourraient refuser des interventions susceptibles d'améliorer leurs conditions de vie (un traitement contre le cancer, par exemple) de peur que leurs données personnelles, prélevées dans des contextes très sensibles, ne soient utilisées par des systèmes d'IA ou dans d'autres contextes d'une manière contraire à leurs intérêts[119]. Cet « effet dissuasif », provoqué par la facilité avec laquelle les données obtenues dans un certain but peuvent être réemployées dans un but différent, souligne l'importance d'honorer et de conserver le principe de spécification des finalités affirmé dans de nombreux instruments contemporains de protection des données. Si l'autonomie et la liberté individuelle comprennent la capacité à naviguer entre des rôles et des identités multiples, en les mélangeant ou, au contraire, en les cloisonnant à notre gré, alors l'usage systématique du profilage et des décisions fondés sur les données personnelles menace cet aspect de notre personnalité[120].

2.2.4 Exercice irresponsable du pouvoir conféré par le numérique

Le fait que les systèmes d'IA traitent les personnes comme des objets plutôt que comme des sujets s'inscrit dans un ensemble de préoccupations plus larges, concernant l'exploitation des individus au service des « géants du numérique ». Ces préoccupations sont de plusieurs sortes. Premièrement, les technologies d'IA (le News Feed de Facebook, par exemple) peuvent fonctionner instantanément à l'échelle de toute une population et, en pratique, il n'est guère possible d'appliquer à des systèmes de ce type une véritable « supervision humaine ».[121] Or, laisser l'IA s'appliquer automatiquement sans supervision humaine complète risque de créer un sérieux vide de responsabilité – qui permet précisément aux géants du numérique de récolter les bénéfices de ces plates-formes sans en assumer les inconvénients[122].

Une telle situation viole les normes élémentaires de réciprocité sociale, et donc « dépossède » sans justification des citoyens et des communautés entières ; mais il s'agit aussi, tout simplement, d'un exercice irresponsable du pouvoir. En d'autres termes, le « vide de responsabilité » qui serait né, selon Matthias[123], de l'apparition de systèmes informatiques capables d'apprendre[124] a pris récemment une nouvelle tournure, du moins sur les sites de réseaux sociaux, où des systèmes automatiques peuvent être conçus pour supprimer ou

---

119 Cet « effet tétanisant » est avéré aux États-Unis, où des personnes ont refusé de se soumettre à des tests génétiques dans des circonstances qui auraient facilité leurs soins de peur que des tiers n'utilisent leurs résultats d'une manière contraire à leurs intérêts, en particulier dans le contexte de l'emploi et de l'assurance vie (Farr 2016).

120 À la lumière de la conception de l'autonomie de Joseph Raz, qui suppose que chacun dispose d'un éventail de choix suffisant, le réemploi généralisé des données pour aider des organisations à prendre des décisions sur les individus pourrait réduire le nombre de choix à notre disposition, et par là notre autonomie. D'après Raz, « pour être auteur de sa propre existence, une personne doit avoir la capacité mentale de concevoir des intentions suffisamment complexes et d'en planifier la réalisation. Cela suppose un minimum de rationalité, l'aptitude à envisager les moyens nécessaires aux fins visées, les facultés mentales nécessaires pour planifier des actions, etc. Pour mener une existence autonome, nous devons user de ces facultés pour déterminer le cours de notre vie. Autrement dit, nous devons avoir à notre disposition plusieurs choix appropriés. Enfin, nous devons être indépendants, c'est-à-dire choisir sans contrainte ni manipulation de la part d'autrui » (Raz 1986 : 373).

121 Voir les réflexions ci-dessus, chapitre 2.1.1(b).

122 Keen 2018.

123 Matthias 2004.

124 Abordé plus loin, chapitre 3.3.2.

diffuser des contenus à une vitesse et à une échelle telles qu'aucun modérateur humain ne peut plus suivre le rythme et que les entreprises de réseaux sociaux elles-mêmes déclinent toute responsabilité[125]. Par ailleurs, les géants du numérique ont jusqu'ici réussi à se prémunir des réglementations extérieures en déclarant obéir à des « codes d'éthique », y compris en affirmant utiliser des solutions technologiques (abordées au chapitre 3.7.1) destinées à intégrer des valeurs normatives à la conception et au fonctionnement de systèmes technologiques mais qui, faute de supervision et de sanctions extérieures, ont peu de chances d'offrir une véritable protection[126].

2.2.5 La privatisation cachée de décisions relatives aux intérêts public

Les technologies d'IA visent à reproduire ou à améliorer des performances sur certaines tâches qui demanderaient de l'« intelligence » si elles étaient accomplies par des êtres humains. Pourtant, l'affirmation selon laquelle ces technologies « surpasseraient » l'être humain repose sur une définition très étroite de l'objectif global, réduit à l'exécution d'une tâche très précise (comme identifier une tumeur sur une radiographie). Or, l'insertion de cette IA spécialisée dans des systèmes sociotechniques complexes, destinés à offrir des services en contexte réel, fait invariablement entrer en jeu des valeurs qui ne se limitent pas à la précision et à l'efficacité dans l'exécution des tâches.

Les systèmes d'IA reflètent les valeurs et les priorités de leur modèle sous-jacent et de ses développeurs, qui ne coïncident pas toujours avec les valeurs partagées par la population ou avec les valeurs démocratiques et constitutionnelles que les droits de l'homme défendent. Pourtant, même s'agissant de systèmes d'IA en contact direct avec le public, les citoyens et les autres groupes et organisations affectés n'ont pratiquement pas leur mot à dire sur la configuration des systèmes au niveau des valeurs et des dilemmes à résoudre[127]. Le recours à l'apprentissage automatique dans les analyses de risque utilisées pour mesurer le « risque de récidive » d'auteurs d'infractions qui demandent une remise en liberté en offre un exemple criant : bien que le système pénal des démocraties contemporaines repose sur plusieurs valeurs importantes, qu'il est censé concrétiser, ces systèmes de notation ne sont à ce jour conçus que pour optimiser une seule valeur : la protection du public[128]. Étant donné que les technologies d'IA servent de plus en plus à optimiser des phénomènes de coordination sociale (systèmes de navigation intelligents ou gestion intelligente des infrastructures, par exemple), il est inévitable que leurs décisions placent certaines valeurs avant d'autres et aient un impact direct sur les individus et les groupes, dont certains en bénéficieront et d'autres non. Comme Sheila Jasanoff[129] et d'autres spécialistes des STS l'ont souligné à maintes reprises, les systèmes technologiques reflètent des valeurs normatives. Étant donné l'ampleur des effets de ces systèmes, ces valeurs devraient faire l'objet d'une participation et de délibérations démocratiques au lieu d'être principalement déterminées en privé, par des prestataires privés poursuivant avant tout leurs intérêts commerciaux.

---

125 Voir cependant les réflexions note 79, ci-dessus.

126 Voir le chapitre 3.3.4, plus loin. Ces stratégies technologiques peuvent être interprétées comme une « sous-traitance » des questions de droits de l'homme aux entreprises de technologie, leur permettant de définir (souvent étroitement) l'étendue et la teneur des droits des usagers et d'exercer seules le pouvoir de faire respecter les règles.

127 Korff et Browne 2013.

128 Zweig et al. 2018.

129 Jasanoff 2016.

2.2.6 Exploitation de main-d'œuvre humaine pour entraîner les algorithmes

On entend souvent dire que les systèmes d'IA et d'apprentissage automatique « surpassent l'être humain » parce que les algorithmes sont entraînés par un très grand nombre de personnes. Un algorithme d'apprentissage automatique destiné à répondre à des recherches en ligne est évalué à l'aune de toute une armée de travailleurs cachés, qui agissent comme l'algorithme jusqu'à ce que les réponses de ce dernier surpassent les leurs. Même une fois l'algorithme entraîné, son application automatique peut générer des effets secondaires indésirables qui demandent à être identifiés et supprimés par des êtres humains. Sur les réseaux sociaux, des modérateurs sont chargés de supprimer les contenus inappropriés. L'entraînement des algorithmes, tout comme le travail conséquent accompli par des êtres humains pour supprimer les effets secondaires des modèles, sont souvent dissimulés afin de préserver le mythe de l'automatisation complète[130]. Les personnes qui entraînent les modèles d'apprentissage automatique vivent souvent dans les régions pauvres, souvent dans l'hémisphère sud, et leurs conditions de travail sont le plus souvent très précaires[131]. En outre, rien n'est généralement prévu pour les aider à surmonter le fardeau psychologique qui peuvent découler de leur travail de « nettoyage ». Par ailleurs, étant donné que beaucoup d'algorithmes apprennent en continu à partir du comportement des utilisateurs, certains estiment que les propriétaires des systèmes « surfent » gratuitement sur le travail des internautes, alimentant un mode de production de l'IA qui contribue à banaliser et à légitimer le travail non rémunéré, les travailleurs humains n'ayant ni droits ni reconnaissance[132].

2.3 Asymétrie des pouvoirs et menaces pour les fondements socio-techniques de la communauté morale et démocratique

Les effets négatifs évoqués ci-dessus, nés de la puissance et de la sophistication croissantes des technologies numériques nouvelles et émergentes, sont exacerbés par le radical déséquilibre de pouvoir entre ceux qui développent et déploient les systèmes algorithmiques et les utilisateurs qui leur sont soumis. Ce déséquilibre tient pour une grande part à la surveillance omniprésente et en temps réel que seuls les premiers sont capables de pratiquer, en collectant et en consultant en continu d'énormes ensembles de données issues de nos comportements en ligne. Par ce biais, les individus et des populations entières sont soumis à une évaluation algorithmique, classés et notés[133], et les propriétaires des plates-formes peuvent communiquer directement avec les utilisateurs – communication qui s'effectue à sens unique, automatiquement et à grande échelle. Pour les individus, dans la pratique, les possibilités de comprendre et d'explorer la complexité des écosystèmes de données dans lesquels ils évoluent sont très restreintes, tout comme la capacité de chacun à savoir si les informations et les autres services numériques lui sont fournis dans les mêmes conditions qu'aux autres usagers[134].

Ce déséquilibre de pouvoir pointe la nécessité, du moins compte tenu des structures institutionnelles actuelles, de réexaminer la capacité des droits et des mécanismes de

130 Irani 2015.

131 Voir par exemple Chen 2014.

132 Ekbia et Nardi 2014.

133 Ferraris et al. 2013.

134 Voir les conclusions réunies dans Which? 2018. Mireille Hildebrandt évoque l'« inconscient numérique », saturé de données, au contraire des informations que les individus peuvent appréhender (Hildebrandt 2015 : 196).

supervision et de mise en œuvre *existants* à traiter de façon complète les risques associés à la montée en puissance des technologies numériques. Comme l'observe l'étude Wagner :

> De fait, le recours croissant à l'automatisation et aux algorithmes décisionnels dans toutes les sphères de la vie publique et privée constitue une menace potentielle pour le concept même de droits de l'homme considérés comme remparts contre l'ingérence des États, dans la mesure où l'on passe progressivement de l'asymétrie traditionnelle du pouvoir et de l'information existant entre les structures d'État et les êtres humains à une asymétrie entre opérateurs d'algorithmes (publics ou privés) et celles et ceux qui sont influencés et gouvernés[135].

En particulier, les institutions existantes de défense des droits de l'homme risquent d'avoir de grandes difficultés à offrir une protection effective, pour au moins trois raisons.

Premièrement, ces technologies sont si opaques et complexes qu'il est très difficile pour les individus, en pratique, de déterminer si leurs droits ont été violés et, le cas échéant, de quelle manière. Les utilisateurs ignorent souvent que ces technologies sont utilisées pour les évaluer. Même quand ils veulent faire valoir leurs droits face à des décisions prises automatiquement, par exemple, les recours à leur disposition n'aboutissent pas toujours au résultat souhaité. Par exemple, les intéressés ne chercheraient pas à obtenir des explications sur le traitement défavorable qui leur a été réservé, mais plutôt à insister sur leur droit à l'égalité de traitement[136].

Deuxièmement, même lorsque les individus soupçonnent un système d'IA d'avoir porté atteinte à leurs droits, il est peu probable qu'ils cherchent à y remédier en pratique si, à leurs yeux, l'atteinte n'est pas assez grave pour justifier qu'ils consacrent du temps, de l'argent et de l'énergie au dépôt et au maintien d'une plainte. Il existe donc un obstacle à l'action collective, qui fait que les effets négatifs cumulés de ces systèmes vont probablement perdurer, du moins en l'absence de mécanismes de réclamation collective ou d'un organisme officiel ayant les compétences, les ressources et le champ d'action nécessaires pour imposer l'application de mesures de protection des droits de l'homme.

Troisièmement, beaucoup des enjeux sociétaux préoccupants ne sont pas faciles à exprimer en termes de droits de l'homme car ils concernent des valeurs et des intérêts *collectifs*, dont certains ont de surcroît des contours assez flous, comme la culture et les contextes moraux et sociopolitiques dans lesquels les technologies numériques avancées opèrent. Parallèlement, la vitesse et l'échelle de ces technologies suscitent des risques potentiels et avérés et des défis auxquels nos sociétés contemporaines n'avaient encore jamais été confrontées. Or, sous de nombreux aspects, l'effet cumulé et collectif de ces systèmes pourrait à terme s'avérer fatal pour les bases sociales et techniques indispensables à l'exercice des droits de l'homme et des libertés fondamentales. Parce qu'elles sont fortement axées sur l'individu[137], les approches actuelles de l'interprétation et de la mise en œuvre des droits de l'homme auront beaucoup de mal à répondre aux risques et aux préjudices *collectifs et cumulés* que ces technologies peuvent générer. Autrement dit, les approches fondées sur les droits et le discours sur les droits tels qu'ils existent aujourd'hui tendent à ignorer les préoccupations structurelles et

135 Étude Wagner 2017 : 33.

136 Edwards et Veale 2017.

137 Yeung 2011.

sociétales profondes, dont les menaces qui pèsent sur le tissu démocratique et moral dans lequel les droits individuels sont ancrés et sans lequel ils n'auraient pas de sens[138].

2.4 Résumé

Ce chapitre a passé en revue les problèmes individuels et collectifs que les technologies numériques avancées peuvent poser à la société. Il a montré en quoi l'usage répandu et croissant des technologies numériques avancées (dont l'IA), en particulier celles fondées sur le profilage par les données, pouvait menacer systématiquement l'exercice des droits de l'homme et, plus généralement, des valeurs et intérêts collectifs non couverts par le champ actuel de la protection des droits de l'homme. Il a également étudié les risques posés par d'autres technologies d'IA et par leurs applications, contemporaines ou prévisibles. Ce sont notamment les applications hostiles ou malveillantes des systèmes utilisant l'IA ou leur conception peu sûre ou contraire à l'éthique, la diminution des occasions de rapports humains authentiques et concrets, l'effet dissuasif de la réutilisation des données, l'exercice irresponsable du pouvoir par les propriétaires des plates-formes numériques et des autres entités utilisant l'IA, la privatisation rampante et cachée des décisions relatives aux intérêts public et l'exploitation de main-d'œuvre humaine pour entraîner les algorithmes. Enfin, le chapitre a pointé le déséquilibre de pouvoir grandissant entre ceux qui ont les capacités et les ressources nécessaires pour développer et appliquer les technologies d'IA et les utilisateurs, groupes et populations directement affectés, déséquilibre qui pourrait réduire considérablement leur aptitude à détecter les atteintes et à déposer des recours devant les institutions existantes de protection des droits. Parce qu'ils sont de grande ampleur et potentiellement graves, les risques individuels et collectifs associés aux technologies numériques avancées soulèvent inévitablement d'importantes questions sur la répartition des responsabilités : qui doit les éviter, les prévenir et les atténuer ? Par ailleurs, si ces risques dégénèrent en préjudices et/ou portent atteinte aux droits de l'homme, qui doit être tenu pour responsable de ces conséquences et sur quels mécanismes institutionnels pouvons-nous compter pour offrir une mise en œuvre et des recours adéquats, en particulier compte tenu des obstacles à l'action collective rencontrés par les détenteurs des droits ? Le chapitre 3 cherche à répondre à ces questions. Il commence par examiner la notion de responsabilité, puis son importance, avant d'analyser en quoi les technologies d'IA remettent en cause les conceptions existantes dans ce domaine.

[138] Voir le chapitre 3.8, ci-dessous.

Chapitre 3. Qui est responsable des menaces, des risques, des préjudices et des torts causés par les technologies numériques avancées ?

Comme nous l'avons vu au chapitre précédent, les technologies numériques avancées sont sources de risques graves pour nos valeurs et nos intérêts individuels et collectifs et pourraient entraîner des préjudices substantiels et systématiques, y compris des atteintes aux droits de l'homme. Pris ensemble, leurs effets menacent de saper les bases morales et sociales collectives des sociétés démocratiques. Ce chapitre s'interroge donc sur les responsabilités en matière de prévention, de gestion et d'atténuation de ces effets, ainsi que de réparation en cas de dommages et d'atteintes aux droits des individus, des groupes et de la société. Les réflexions qui suivent examinent les liens entre la notion de responsabilité et l'émergence des technologies numériques avancées (dont l'IA), notamment à la lumière de leurs incidences sur les droits de l'homme protégés par la CEDH évoqués au chapitre 2.

Ces réflexions procèdent en plusieurs phases.

Premièrement, nous précisons ce que nous entendons par « responsabilité » et expliquons son importance, en soulignant son rôle vital pour la garantie et la concrétisation de l'État de droit et pour une coopération pacifique au sein de la société.

Deuxièmement, nous examinons les deux grands thèmes abordés dans le débat contemporain sur les risques associés aux technologies d'IA : d'une part, la promulgation par le secteur du numérique lui-même de « codes d'éthique » qu'il s'engage à respecter, et d'autre part, le « problème de contrôle » qui découlerait de la capacité des systèmes pilotés par IA à fonctionner plus ou moins indépendamment de leurs créateurs.

Troisièmement, nous identifions une série de « modèles de responsabilité » pouvant servir à répartir les responsabilités pour différents types d'impacts négatifs des systèmes d'IA : modèles fondés sur l'intention/la culpabilité, sur le risque/la négligence, responsabilité absolue et régimes d'assurance obligatoires. Notre étude est axée sur les incidences sur les droits de l'homme ; pour les atteintes à ces droits, la responsabilité est largement envisagée comme « absolue » (dès lors qu'une atteinte à un droit de l'homme est établie, il n'est pas nécessaire de prouver une faute). En revanche, les obligations de réparation en cas de dommage matériel (atteinte à la santé ou aux biens) peuvent être attribuées juridiquement, selon divers modèles rétrospectifs. En cas de dommage matériel causé par des systèmes d'IA, l'attribution de la responsabilité rétrospective revêt aussi une dimension prospective, puisqu'elle aide à identifier la nature et l'étendue des obligations de ceux qui développent, produisent et mettent en œuvre les systèmes d'IA. Ces deux types de responsabilité sont donc brièvement présentés.

Quatrièmement, nous attirons l'attention sur l'énorme défi que représente l'attribution des responsabilités lorsqu'entrent en jeu des systèmes sociotechniques complexes et en interaction, auxquels contribuent de multiples acteurs, organisations, composantes électroniques, algorithmes et utilisateurs, souvent dans des environnements complexes et en perpétuelle évolution.

Cinquièmement, nous attirons l'attention sur une série de mécanismes extrajudiciaires destinés à établir la responsabilité prospective et rétrospective pour les effets négatifs des systèmes d'IA, dont les analyses d'impact, les audits et les dispositifs techniques de protection.

Sixièmement, nous soulignons le rôle et les obligations des États à l'égard des risques associés aux technologies numériques avancées, et notamment leur obligation d'assurer une protection effective des droits de l'homme.

Nous affirmons, pour finir, la nécessité de renouveler le discours sur les droits de l'homme à l'ère du numérique, afin de préserver et d'entretenir les bases sociotechniques indispensables à la liberté d'action et à la responsabilité humaines, sans lesquelles les droits de l'homme et les libertés ne peuvent s'exercer véritablement.

3.1 Qu'est-ce que la responsabilité et en quoi est-elle importante ?

En énonçant les objectifs de cette étude, nous avons déjà relevé que les conceptions et les pratiques d'une société en matière de responsabilité étaient d'une importance cruciale car elles garantissent, dans le cadre de régimes démocratiques constitutionnels, que les individus et les organisations aient à répondre des effets négatifs de leurs actions sur des tiers. Bien que le thème de la responsabilité ait fait l'objet de nombreux écrits philosophiques et juridiques, assez peu de chercheurs se concentrent sur le rôle fondamental de la responsabilité pour les individus et pour la société. Mais en filigrane, tous ces travaux reconnaissent que la notion de responsabilité sert deux fonctions cruciales, que le spécialiste de philosophie morale Gary Watson nomme les « deux faces de la responsabilité[139] ». La première face est essentielle à notre sentiment d'« être au monde » en tant qu'agents moraux, c'est-à-dire en tant qu'auteurs de nos propres vies, agissant sur la base de certaines raisons. Pour citer Watson :

> La responsabilité compte lorsqu'il s'agit de construire sa vie, et même d'avoir une vie au sens biographique, et de connaître la qualité et le caractère de cette vie. Ces aspects correspondent à l'une des faces de la responsabilité : celle de l'accomplissement[140].

Mais Watson identifie une seconde face, celle des pratiques visant à ce que chacun rende des comptes[141]. Pour lui,

> lorsque nous disons qu'un comportement devrait être « censuré », qu'il est « répréhensible », « intolérable » ou « inconscient » (voire « mauvais »), nous sous-entendons qu'une *réaction* à l'encontre de son auteur est souhaitable (en principe). C'est la pratique consistant à tenir les personnes pour moralement responsables : un juge (le plus souvent ; sinon, d'autres membres de la communauté morale) est habilité (en principe) à réagir de diverses manières.

Le scénario suivant illustre la différence entre ces deux faces de la responsabilité, qu'on peut nommer d'une part la perspective de la « réalisation de soi » et d'autre part celle de la « responsabilisation morale » :

> Si quelqu'un trahit ses idéaux en optant pour un emploi ennuyeux mais sûr au détriment d'une activité plus risquée, mais potentiellement plus enrichissante, ou si elle compromet bêtement quelque chose de très important pour sa vie (par exemple, en ne dormant pas assez ou en buvant trop avant une échéance importante), on dit qu'elle a mal agi – par lâcheté, faiblesse ou, du moins, manque de sagesse. En la jugeant ainsi, nous constatons qu'elle est responsable, mais nous

139 Watson 2004.
140 Watson 2004 : 262-263.
141 Watson 2004 : 264.

> n'engageons pas sa responsabilité. Pour cela, il nous faudrait penser qu'elle est responsable envers nous-même ou envers autrui. Or dans de nombreux cas, nous estimons qu'un tel comportement « ne regarde personne ». Sauf si nous jugeons que cette personne nous doit, ou doit à autrui, de vivre la meilleure vie possible – et c'est une question morale –, nous ne lui demanderons pas de comptes. La question se posera différemment si son comportement timoré ou irréfléchi porte préjudice à autrui, et donc viole les exigences associées aux relations interpersonnelles[142].

On trouve une idée similaire dans le concept de « responsabilité fondamentale », forgé et développé par le juriste John Gardner, selon lequel notre responsabilité fondamentale se trouverait au cœur de notre sentiment d'être au monde. Elle est essentielle à notre identité d'agents rationnels, c'est-à-dire d'êtres agissant sur la base de raisons et qui, en tant qu'individus, veulent que leur vie ait un sens – qu'elle enrichisse la série des « *quoi* », mais aussi celle des « *pourquoi*[143] ».

Watson avance que la maîtrise occupe une place centrale dans les pratiques de responsabilisation qui caractérisent la seconde face de la responsabilité.

> Parce que certaines de ces pratiques – notamment celle de la responsabilisation morale – imposent des exigences aux personnes, elles soulèvent des enjeux d'équité qui n'existent pas dans l'accomplissement de soi. C'est ce souci d'équité qui fait de la maîtrise (ou de l'évitabilité) l'une des conditions de la responsabilisation morale. « Tenir pour responsable » peut être considéré comme équivalant à « demander des comptes ». Mais « tenir » ici ne doit pas être confondu avec *croire* (comme dans « Je tiens qu'elle est responsable de *x* »). Tenir quelqu'un pour responsable englobe la volonté de réagir d'une certaine manière à son comportement. Devoir « assumer » telle ou telle chose, c'est s'exposer à certaines réactions parce qu'on n'a pas agi comme on était censé le faire. Requérir ou exiger d'une personne un comportement donné, c'est prévoir que dans le cas contraire, elle s'exposera à un traitement défavorable ou non souhaité. Par commodité, j'appellerai « sanctions » les diverses formes de traitement défavorable. Tenir pour responsable suppose donc l'idée d'exposition à des sanctions. Par conséquent, qui est habilité à formuler des exigences l'est aussi à définir les critères d'application de sanctions[144].

Comme la présente étude cherche à situer les responsabilités à l'égard des risques potentiels et avérés, des dommages et des atteintes aux droits de l'homme, individuels et collectifs, découlant des technologies numériques avancées, elle se concentre sur la seconde face de la responsabilité : le fait de « demander des comptes ». Néanmoins, il existe un lien crucial entre les deux faces de la responsabilité, qui réside dans le statut d'*agent moral*, d'individu capable de choisir et de décider activement, y compris en affectant autrui ou en risquant de porter préjudice à autrui ou de lui causer des torts. Comme l'écrit Gardner, « nous ne sommes des agents moraux que dans la mesure où nous sommes fondamentalement responsables[145] ». La responsabilité fondamentale est donc cruciale pour les deux faces de la responsabilité. Comme l'observe Gardner, chaque fois que nous causons du tort ou commettons des erreurs, nous nous cherchons des justifications et des excuses ; non seulement parce qu'en tant qu'êtres rationnels, nous voulons éviter d'en assumer des conséquences (désagréables) (« perspective

142 Watson 2004 : 265-266.
143 Gardner 2003.
144 Watson 2004 : 272-273.
145 Gardner 2008 : 140.

de Hobbes »), mais aussi pour une raison plus profonde (que Gardner nomme « perspective d'Aristote ») : toujours en tant qu'êtres rationnels, nous souhaitons tous affirmer notre responsabilité fondamentale, et donc être en mesure de nous expliquer[146].

Responsabilité et État de droit

La responsabilité fondamentale est essentielle pour que nous nous appréhendions nous-mêmes non seulement comme auteurs de nos propres vies, mais aussi comme *membres d'une communauté de sujets moraux*. Les sujets ou « agents » moraux ont la capacité et la liberté d'opérer des choix sur leurs décisions et leurs actions, et cela d'une manière qui peut être répréhensible ou causer un dommage, soit à d'autres individus soit aux conditions indispensables à la stabilité et à la coopération sociale qui rendent possible la vie en communauté. Notre responsabilité fondamentale, et les pratiques via lesquelles les membres d'une communauté se demandent mutuellement des comptes, caractérisent une communauté politique comme une communauté largement *morale* (c'est-à-dire composée de sujets moraux). Le respect mutuel et l'autodiscipline pratiqués par les membres d'une communauté morale revêtent une importance cruciale, puisqu'ils rendent possible et perpétuent la vie en communauté et forment en fait le socle d'un idéal contemporain : l'État de droit[147]. Une société, si elle est dépourvue de mécanismes institutionnalisant ses pratiques de responsabilisation des personnes ayant nui à d'autres (y compris en leur portant préjudice ou en violant leurs droits de l'homme), se prive des fonctions protectrices vitales que de telles institutions offrent, et qui sont indispensables à une coopération et à une coordination fiables et pacifiques. Autrement dit, les mécanismes d'attribution des responsabilités jouent un rôle essentiel pour la structure sous-jacente de coopération sociale sans laquelle le droit ne peut l'emporter sur la force. Parallèlement, il est important de reconnaître que la stabilité et la continuité de ces bases sociales dépendent, en dernière analyse, du respect mutuel et de l'autodiscipline de chaque membre de la communauté morale et non d'un système de contrainte et de contrôle technologique. Ce sont précisément ce respect mutuel et cette autodiscipline qui manquent à la société ostensiblement heureuse, stable, efficace et ordonnée imaginée par Huxley dans *Le meilleur des mondes*[148]. Les habitants de ce monde n'ont ni droits ni libertés. Ils ne vivent pas dans une communauté morale, mais dans une société dont les membres ne sont que des objets passifs, aux pensées et aux actions pilotées et contrôlées par le pouvoir technologique exercé par un dictateur, et où les notions de liberté, d'autonomie et de droits de l'homme non seulement dépérissent, mais se trouvent simplement vidées de leur sens et de leur prix[149].

Responsabilisation, réponses et transparence

L'importance cruciale des structures de responsabilité institutionnalisées pour préserver les bases sociales de l'État de droit pointe la nécessité, dans toute communauté morale et politique s'engageant à respecter les droits de l'homme, de créer et d'appliquer des mécanismes institutionnels permettant de demander des comptes aux membres de la communauté. Bien que le concept de responsabilisation fasse débat, il peut être défini dans le cadre de notre étude comme « le fait de demander à une personne d'expliquer et de justifier,

146 Pour Gardner, ces explications n'ont pas à s'adresser à quelqu'un en particulier mais peuvent, et devraient, s'adresser à tous. Il rejette par conséquent l'idée que la responsabilité serait nécessairement « relationnelle ».

147 Galligan 2006.

148 Huxley 1932 ; Yeung 2017b.

149 Yeung 2011.

à l'aune de certains critères, ses décisions ou ses actes, puis de réparer ses éventuelles fautes ou erreurs[150] ». Dans cette perspective, les mécanismes de responsabilisation englobent quatre aspects : définition des normes à l'aune desquelles juger les faits, établissement des faits, jugement, et décision sur les conséquences qui devraient s'ensuivre (le cas échéant). Le concept de responsabilisation est d'une importance particulière dans les relations entre supérieur et subordonné, le second étant censé agir pour le compte et au nom du premier, et donc être capable de rendre compte – de *répondre de ses actes* – auprès de son supérieur. La transparence est directement liée à la responsabilisation, dans la mesure où cette dernière suppose que les intéressés puissent expliquer les motifs de leurs actions et les justifier à l'aune de règles ou de critères donnés. La transparence est donc importante dans deux buts au moins : permettre aux personnes affectées par des actes ou par des décisions d'en connaître les motifs, et de pouvoir apprécier la qualité de ces motifs[151].

Les mécanismes de responsabilisation revêtent une importance particulière pour l'exercice du pouvoir public dans les sociétés démocratiques libérales, étant donné que les agents de l'État y sont considérés comme au service des citoyens, dont ils tirent leur pouvoir, et agissent en leur nom. Cependant, la responsabilisation est importante chaque fois que l'exercice d'un pouvoir peut avoir des effets nocifs sur autrui. Par conséquent, le pouvoir, l'ampleur et les effets des systèmes sociotechniques fondés sur les technologies d'IA ont donné lieu à toute une série d'alertes, ayant en commun d'appeler à « responsabiliser les algorithmes », d'autant plus que ces systèmes sont opaques et que certains de leurs usages peuvent avoir des conséquences très lourdes pour les individus, les groupes et la société en général[152]. Il est essentiel de désigner les personnes responsables et devant rendre des comptes en cas d'atteintes aux droits de l'homme et d'autres effets néfastes découlant de ces technologies. Bien que le droit existant, notamment le droit constitutionnel et les lois sur la protection des données, la protection des consommateurs et la concurrence, qui protègent les droits de l'homme dans les systèmes juridiques nationaux, puissent jouer un rôle significatif pour établir les différents aspects de la responsabilisation algorithmique, leurs contributions respectives dépassent le champ de la présente étude. Les réflexions qui suivent s'intéressent aux incidences des technologies numériques avancées (dont les systèmes d'IA) sur la *notion de responsabilité*, en se concentrant sur les atteintes aux droits de l'homme ; elles s'appuient à la fois sur le droit et sur la philosophie morale.

3.2 Les différents aspects de la responsabilité

L'acception de la responsabilité comme le fait d'« assumer ses actes » a été maintes fois abordée par les spécialistes du droit et de la philosophie, et l'analyse qui suit se fonde sur une sélection d'ouvrages dans ces domaines. La « responsabilité » peut revêtir de nombreux sens différents[153] ; aux fins de la présente étude, il convient d'en souligner l'aspect temporel. On distingue :

a) la responsabilité rétrospective (ou historique), tournée vers le passé, cherchant à établir les responsabilités pour des comportements et événements qui se sont déjà produits. Comme nous le verrons, l'attribution de la responsabilité rétrospective pour les dommages et torts causés par des systèmes d'IA soulève des difficultés considérables ; et

150 Oliver 1994 : 245. Voir aussi Bovens 2007 et les auteurs qu'il cite.

151 Yeung et Weller 2018b. Zalnieriute et al. 2019.

152 Yeung 2017.

153 Hart 1968 : 211-230.

b) la responsabilité prospective, tournée vers l'avenir, définissant les obligations associées aux différents rôles et tâches afin de favoriser les résultats désirables au détriment des indésirables. Les responsabilités prospectives jouent un important rôle d'orientation. Comme l'écrit Cane, « l'une des raisons les plus importantes de notre intérêt pour la responsabilité et les concepts qui lui sont associés tient au rôle qu'ils jouent dans le raisonnement pratique sur nos droits et obligations envers les autres, et sur la manière dont nous devrions nous comporter avec eux [154] ». Dans le contexte des actions des systèmes autonomes robotiques/d'IA et de leurs conséquences, l'idée d'une « responsabilité par rôle[155] » a parfois été avancée.

Toute réponse légitime et effective aux risques, aux dommages et aux violations de droits entraînés par les technologies numériques avancées mettra probablement l'accent sur leurs conséquences pour les individus et la société, afin d'attribuer de manière juste et équitable *à la fois* la responsabilité prospective – pour prévenir et atténuer les risques – et la responsabilité rétrospective au regard des effets négatifs du fonctionnement des systèmes sociotechniques complexes dans lesquels ces technologies s'inscrivent. Ce n'est que si ces deux aspects sont couverts que les individus et la société auront l'assurance, d'une part que des efforts sont engagés pour éviter la survenue de torts et de préjudices, et d'autre part que des mécanismes institutionnels existent pour assurer une réparation appropriée et éviter, le cas échéant, que les mêmes effets néfastes ne se reproduisent. Il faudra pour cela s'intéresser à la fois à ceux qui développent, déploient et mettent en œuvre ces technologies, aux utilisateurs individuels et aux groupes affectés par ces technologies et aux mesures prises par l'État (et par les États, agissant collectivement et en coopération) pour assurer la création et le maintien des conditions nécessaires pour mettre les citoyens à l'abri des risques inacceptables, garantissant ainsi une protection adéquate des droits de l'homme. En d'autres termes, l'examen attentif de la responsabilité des technologies et systèmes d'IA doit tenir compte de la situation morale de ceux qui agissent comme de ceux qui subissent, ainsi que de la communauté morale au sens large, pour répondre aux questions : *la responsabilité, envers qui et pour quoi[156] ?*

3.3 Liens entre technologies numériques avancées (dont l'IA) et conceptions actuelles de la responsabilité

Nous avons précisé ce que nous entendions par « responsabilité » et souligné la nécessité d'en envisager les aspects prospectif et rétrospectif. Nous pouvons maintenant poser la question : qui est responsable des effets négatifs et des risques associés au développement et à la mise en œuvre des technologies d'IA, dont les atteintes aux droits de l'homme et les autres torts et préjudices découlant de leur fonctionnement ? Si poser la question est relativement facile, y répondre soulève de considérables difficultés conceptuelles. Comme l'observe le Groupe européen d'éthique[157], les technologies d'IA posent

> […] des questions de responsabilité morale de l'homme. Où se situe l'entité moralement pertinente dans les systèmes sociotechniques dynamiques et complexes de l'IA et des composants robotiques avancés ? Comment attribuer et

154 Cane 2002 : 45.

155 Hart 1968 : 211-230.

156 Liu et Zawieska 2017 ; Cane 2002.

157 Groupe européen d'éthique 2017.

répartir la responsabilité morale, et qui est responsable (et en quels termes) en cas de résultats indésirables ?

Autrement dit, la complexité des technologies elles-mêmes, et celle des contextes sociotechniques dans lesquels elles sont appliquées, peut brouiller les lignes de la responsabilité morale, en particulier en cas de fonctionnement inattendu aboutissant à des dommages ou à des atteintes aux droits. Cependant, il ne faut pas oublier que la responsabilité morale et la responsabilité juridique, bien que liées, sont des concepts distincts. Contrairement à la morale, le droit s'appuie sur un système très développé destiné à institutionnaliser et à faire respecter les responsabilités (y compris en appliquant des sanctions dans certaines circonstances), car il sert à trancher des différends dans le monde réel, ce qui suppose à la fois des jugements définitifs et une sécurité juridique[158]. Aucune société ne peut compter que sur la propension des individus à « bien agir ». L'absence de mécanisme institutionnel destiné à faire appliquer les normes d'éthique (y compris en sanctionnant légitimement les infractions), qui laisserait le champ libre à la seule volonté, n'offrirait pas les fondements sociaux stables et fiables nécessaires à une coopération pacifique au sein des sociétés contemporaines. Le droit joue donc un rôle essentiel pour déterminer et institutionnaliser les responsabilités afin de protéger les droits et d'assurer la réalisation des devoirs juridiquement reconnus. Comme nous allons le voir, les systèmes juridiques ont une conception de la responsabilité rétrospective traditionnellement plus sensible aux intérêts des victimes et de la société et à la protection des personnes et des biens, tandis que la philosophie morale tend à insister sur le sujet moral et sur ceux de ses comportements qu'il convient de blâmer. Cependant, appliquer ces conceptions morales et juridiques de la responsabilité aux technologies numériques avancées d'aujourd'hui (dont l'IA) n'a rien d'évident. Ces technologies et ces systèmes peuvent en effet accomplir des tâches qui étaient auparavant impossibles, ce qui remet en question la notion juridique, morale et sociale de responsabilité telle qu'elle existe aujourd'hui ; notamment du fait de certaines caractéristiques de ces technologies, examinées au chapitre 2.1. Elles sont en effet :

- opaques et impénétrables
- complexes et évolutives
- fondées sur l'apport d'êtres humains, leur libre arbitre et les interactions avec eux
- de nature généraliste
- interconnectées, applicables et déployables dans le monde entier
- fondées sur de grands ensembles de données
- fonctionnant automatiquement et en continu, souvent en temps réel
- capables d'extraire des connaissances « cachées » à partir des masses de données fusionnées entre elles
- capables d'imiter avec fidélité des traits humains
- associées à des logiciels de plus en plus complexes (notamment exposés aux attaques et vulnérables aux dysfonctionnements)
- capables de « personnaliser » et de configurer un environnement adapté à chaque usager

- capables de répartir les risques, les avantages et les inconvénients entre les groupes et les individus, via des systèmes d'optimisation par IA, pour reconfigurer les choix et les environnements sociaux ;
- faisant obstacle à l'action collective.

158 Cane 2002.

Avant de poursuivre, il est important de marquer la distinction conceptuelle entre deux types d'effets négatifs que les systèmes d'IA peuvent entraîner (et ont déjà entraînés) :

(a) les **violations des droits de l'homme** (dont les droits protégés par la CEDH, mais pas uniquement), et
(b) les **dommages matériels** sur la santé humaine, les biens ou l'environnement.

Il s'agit de concepts différents, aux implications différentes. Il peut y avoir violation des droits de l'homme sans aucun dommage matériel, et inversement. En 2016, par exemple, Facebook a retiré une photographie symbole de la guerre du Vietnam, celle d'une fillette de 9 ans fuyant un bombardement au napalm, au motif que les « Standards de la communauté » interdisaient la nudité ; ce retrait peut être compris comme une atteinte à l'article 10 (liberté d'expression et d'information), bien qu'il n'ait pas causé de dommage matériel substantiel[159]. Inversement, si une voiture autonome heurte et blesse un animal sauvage, il y a un dommage, mais pas d'atteinte aux droits de l'homme. Cependant, un événement ou une série d'événements peut toujours entraîner à la fois des dommages matériels et une violation des droits de l'homme. Par exemple, si un véhicule autonome blesse mortellement un piéton, il y aura à la fois violation de l'article 2 (droit à la vie) et infliction d'un dommage matériel[160].

La présente étude examine avant tout les incidences des systèmes d'IA sur la responsabilité sous l'angle des droits de l'homme. Par conséquent, elle analyse les responsabilités en cas d'atteintes aux droits de l'homme plutôt qu'en cas de dommage matériel. Les réflexions qui suivent se concentrent sur ceux qui créent, développent, mettent en œuvre et régissent les systèmes d'IA. Peuvent-ils être tenus pour responsables des éventuelles conséquences néfastes de ces systèmes ? Pour commencer à répondre, nous examinerons deux thèmes clés qui ressortent des analyses contemporaines sur la responsabilité des risques posés par les technologies d'IA : premièrement, l'adoption volontaire par le secteur des technologies de « codes d'éthique » qu'il s'engage publiquement à respecter ; deuxièmement, l'affirmation selon laquelle, parce que les systèmes d'IA agissent de manière autonome, leurs créateurs ne seraient pas responsables de leurs décisions et de leurs éventuels effets négatifs. Nous décrirons plusieurs « modèles de responsabilité » pouvant servir à déterminer les responsabilités de ceux qui développent et mettent en œuvre les systèmes d'IA, avant d'étudier les obligations de l'État face à ces effets négatifs à l'aune de ces différents modèles.

3.3.1 Responsabilité prospective : codes d'éthique volontaires et « robotique responsable »

L'inquiétude croissante du public et le récent « retour de bâton anti-techno[161] », réaction à la montée en puissance des pratiques et des politiques des géants du numérique – en particulier depuis le recours au microciblage politique et le scandale Cambridge Analytica, ont précipité chez les entreprises concernées l'adoption de nombreuses mesures d'« éthique ». Typiquement, ces initiatives consistent à promulguer une série de règles et de normes, qu'une entreprise isolée ou un groupe d'entités (y compris des organisations à but non lucratif[162] ou

159 Voir Scott et Isaac 2016.

160 L'étendue des effets négatifs considérés comme un « dommage » juridiquement reconnu varie selon les pays. Dans les systèmes de *common law*, par exemple, certaines formes de dommage immatériel (comme l'angoisse ou la détresse émotionnelle) peuvent être juridiquement reconnues comme un dommage appelant une indemnisation dans des affaires de préjudice corporel (Gilliker 2000).

161 The Economist 2018b.

162 Par exemple, le mouvement « pour une intelligence artificielle bénéfique » est soutenu par le Future of Life Institute ; voir Conn 2017.

un organisme de normalisation technique[163]) s'engagent publiquement et volontairement à respecter (souvent nommés « codes d'éthique[164] » ou « codes de déontologie »). Ces initiatives peuvent être vues comme les éléments d'un mouvement vers ce que Liu et Zawieska nomment le « projet IA/robotique responsable[165] ».

Deux caractéristiques de ces initiatives méritent d'être soulignées. Premièrement, elles s'intéressent à la *responsabilité prospective*. Ceux qui les adoptent cherchent à attribuer des « responsabilités par rôles » (ou par domaines d'obligations) aux personnes qui participent à chaque étape de la conception, du développement et du déploiement de ces technologies, en vue de montrer au public le sérieux de leur engagement à répondre aux préoccupations éthiques[166]. Point à noter, ces initiatives évitent soigneusement d'aborder les *responsabilités rétrospectives* des personnes concernées lorsque les choses tournent mal. Elles évitent aussi de dire *à qui revient la faute* en cas de conséquences néfastes et de reconnaître une *obligation de dédommager* les personnes touchées. Comme l'explique Liu, la responsabilité par rôles traduit plutôt « une conception de la responsabilité attachée à un individu en vertu du poste qu'il/elle occupe ou du rôle qu'il/elle est censé-e remplir, et donc dépendante de l'exécution d'obligations pouvant être définies à l'avance[167] ». Une personne qui a rempli les devoirs associés à son rôle ou à son poste est donc considérée comme s'étant dûment acquitté de ses responsabilités[168].

Deuxièmement, ces initiatives pour une « IA/robotique responsable » peuvent être caractérisées comme un mouvement émergent d'autogouvernance professionnelle s'inscrivant dans la lignée d'un phénomène plus ancien, souvent nommé « responsabilité sociale des entreprises ». La nature « sociale » (et non juridique) de ces « codes d'éthique », et le fait qu'ils ne reposent que sur le volontariat, font que les obligations et engagements qui y sont inscrits ne peuvent être invoqués devant un tribunal. Ces initiatives ne prévoient pas non plus la création et le maintien d'institutions et de mécanismes de mise en œuvre habilitant une entité extérieure indépendante à apprécier le respect des engagements ou à imposer des sanctions en cas de manquement. Ainsi, bien que bienvenues car elles montrent que l'industrie reconnaît l'éthique des technologies numériques avancées comme un sujet d'intérêt public méritant son attention et son action, ces initiatives sont dépourvues de tout mécanisme institutionnel formel permettant de les faire appliquer et de sanctionner les violations. Par ailleurs, le public n'est pas toujours représenté lors de la définition des normes. Par conséquent, ces initiatives ont été largement critiquées comme une forme d'« éthique-washing[169] » ne prenant pas les enjeux éthiques réellement au sérieux[170].

Si ces codes s'accompagnaient de mécanismes institutionnels *alignés sur le droit*, avec notamment une participation extérieure à la définition et à l'évaluation des normes et une

---

163 Voir par exemple les différentes directives et recommandations élaborées par la Global Initiative for Ethical Considerations in AI and Autonomous Systems de l'IEEE (2017).

164 Par exemple les « Objectifs de l'intelligence artificielle » définis par Google ; voir Pichai 2018.

165 Liu et Zawieska 2017.

166 Liu et Zaweiska 2017 ; Loui et Miller 2007 ; Eschelman 2016.

167 Cane critique la définition étroite de la responsabilité par rôles, cantonnée à des tâches ou à des fonctions spécifiques, et observe qu'« être une personne responsable suppose de prendre au sérieux les responsabilités prospectives, quelles qu'elles soient, de toute activité à laquelle on se livre à n'importe quel moment » (Cane 2002 : 32).

168 Liu 2016 : 336.

169 Wagner 2019 ; Metzinger 2019.

170 Green et al. 2019 ; Hagendorff 2019.

supervision extérieure indépendante du respect des règles par chaque entreprise, les personnes affectées (et la société en général) auraient de plus solides raisons de croire à l'existence de garanties *réelles* et *démocratiquement légitimes* visant à prévenir et à atténuer certains des risques éthiques associés aux technologies (voir le chapitre 3.7, plus loin[171]). Insister sur les droits de l'homme, c'est insister sur le besoin de garanties réelles et effectives[172]. Par ailleurs, les approches prospectives ne peuvent assurer la juste attribution de la *responsabilité rétrospective* en cas de dommage ou d'acte répréhensible. Comme l'avancent Liu et Zawieska, bien que louable, le « projet robotique/IA responsable » laisse un « vide de responsabilité » car il ne s'intéresse qu'aux rôles, et non aux causes. Contrairement à la responsabilité par rôles, la responsabilité causale est une forme de responsabilité rétrospective. Elle consiste à rechercher et à établir les liens de cause à effet. Elle est donc de nature *rétrospective* et par essence tournée vers l'extérieur et vers les relations, car elle met en avant le patient moral (c'est-à-dire la ou les personne(s) affectée(s) par l'acte en question[173]). À l'inverse, l'attribution de la responsabilité par rôles se concentre, de manière prospective, sur les fonctions des personnes identifiées comme les agents responsables. Elle crée un « vide de responsabilité », car s'acquitter de ses responsabilités prospectives ou par rôles ne garantit pas nécessairement la bonne attribution de la responsabilité causale[174]. Autrement dit, la responsabilité par rôles ne garantit ni que les personnes assumeront rétrospectivement leurs responsabilités, ni que les coupables pourront être désignés, car elle ne concerne que l'accomplissement d'obligations pré-établies et non les comptes à rendre et les mesures réparatrices en cas de conséquences négatives[175].

3.3.2 Machines autonomes et « problème de contrôle »

(a) Le prétendu « problème de contrôle »

Face aux appels à situer les responsabilités en cas d'incidences négatives des technologies numériques avancées, on entend souvent dire que, parce que ces systèmes fonctionnent de façon plus ou moins autonome et sans intervention ni contrôle humains directs depuis l'extérieur, il serait injuste d'imputer à ceux qui les développent et les appliquent la responsabilité des décisions, des actions et des conséquences de ces systèmes. Ce point de vue est défendu par Matthias[176], selon lequel

> l'agent ne peut être considéré comme responsable que s'il connaît les faits spécifiques qui entourent son action et s'il est capable de prendre librement la

171 Nemitz 2018.

172 Voir AHRC, note n° 10 ci-dessus. Comme affirmé par David Kaye, Rapporteur spécial sur la promotion et la protection du droit à la liberté d'opinion et d'expression, devant l'Assemblée générale des Nations Unies, « la mise en place de codes de déontologie et de structures institutionnelles d'accompagnement pourrait constituer un complément important aux mesures de protection des droits de l'homme, mais ils ne sauraient s'y substituer. Les codes et principes directeurs publiés par les organismes des secteurs public et privé devraient souligner le fait que c'est le droit des droits de l'homme qui établit les règles fondamentales de la protection des personnes dans le contexte de l'intelligence artificielle ». Assemblée générale des Nations Unies, 2018 : 20.

173 En droit, on tend à employer les termes de « victime » ou « victime potentielle » plutôt que celui de « patient moral », ce dernier étant plus courant en philosophie.

174 Liu et Zaweiska 2017.

175 Liu 2016.

176 Matthias 2004.

> décision d'agir et de sélectionner une série d'actions parmi celles envisageables compte tenu des faits[177].

Or, de plus en plus de machines, que Matthias nomme les « agents artificiels autonomes », relèvent d'une catégorie capable de poursuivre certains objectifs (souvent très limités) en se déplaçant seules dans un « espace » et en agissant *sans supervision humaine*. Cet agent peut être un logiciel se mouvant dans un espace informationnel (robot d'indexation, par exemple), mais aussi avoir une présence physique (robot de compagnie, par exemple) et se déplacer dans le temps et l'espace. Ces agents sont délibérément conçus pour agir et, inévitablement, interagissent avec d'autres objets, personnes et entités sociales (lois, institutions, attentes). Au moins pour ceux qui ont une présence physique et peuvent tirer des enseignements de leurs interactions directes dans un environnement réel, ils peuvent, en retour, manipuler directement cet environnement et le partager avec les êtres humains.

Selon Matthias, cela crée un « vide de responsabilité », car l'agent humain qui a programmé les agents artificiels de ce type n'a plus la maîtrise directe de leur comportement : cette maîtrise est peu à peu transférée à la machine elle-même. Et il serait injuste de tenir les êtres humains pour responsables des actions de machines qu'ils ne peuvent suffisamment contrôler[178]. Matthias offre plusieurs exemples de ce type d'agents artificiels, reposant notamment sur les phénomènes suivants :

a) *réseaux neuronaux artificiels* : au lieu d'une représentation symbolique claire et distincte d'informations dont on maîtrise le flux, on est parfois en présence d'une très grande matrice synaptique qui ne peut être directement interprétée. Les connaissances et comportements stockés dans un réseau neuronal ne peuvent être que déduites indirectement, par expérimentation et application de tests, après la fin de l'entraînement du réseau ;

b) *apprentissage par renforcement* : habituellement fondé sur les mêmes principes de réseau neuronal, il traduit en outre la distinction entre phase d'entraînement et phase de production. Les systèmes d'apprentissage par renforcement explorent leur marge d'action tout en fonctionnant dans leur environnement opérationnel, ce qui constitue à la fois leur caractéristique centrale (grâce à laquelle ils peuvent s'adapter à des environnements en évolution constante) et un inconvénient majeur en termes de prévisibilité. Les informations stockées dans le réseau ne peuvent être totalement vérifiées, même indirectement, car elles ne cessent de changer. Même si l'on peut prouver mathématiquement que les performances

---

177 Matthias 2004 : 175.

178 La position de Matthias a été déterminante dans le débat, qui n'a pas remis en question la théorie de la responsabilité morale « par les choix » sur laquelle ses arguments reposent. Au contraire, dans les milieux académiques, on a cherché à contrer ses arguments en mettant en avant un individualisme méthodologique et moral : toute action serait attribuable en dernier ressort à des individus humains et quel que soit le rôle joué par des objets non humains dans un résultat particulier, ces objets demeureraient accessoires (Hanson 2009 : 92). Sous cet angle, les technologies d'IA sont conçues comme un outil employé par des êtres humains, si bien qu'un humain sera toujours responsable en cas de faute (programmeur, codeur, fabricant, développeur, utilisateur, etc.) (Johnson 2006 ; Bryson 2010 ; Sullins 2005). D'autres ont réagi en considérant l'IA comme un exemple de personne morale ou juridique au statut ontologique indépendant (par ex. Gunkel 2017), allant jusqu'à reconnaître aux agents informatiques une part de sens moral (Dennett 1997 ; Sullins 2005). Toutefois, la grande majorité des spécialistes nie que des entités non humaines puissent avoir en elles-mêmes une responsabilité morale, car elles n'ont pas les qualités mentales (et ne peuvent donc remplir la condition épistémique) généralement reconnues comme nécessaires à la responsabilité morale, laquelle – du moins dans les écrits philosophiques – est souvent décrite en termes d'intentionnalité, de capacité à agir volontairement, de conscience de ses actions et d'anticipation de leurs conséquences (Johnson 2006 ; Kuflick 1999 ; Sparrow 2007 ; Asaro 2014 et Hanson 2009 : 93).

globales d'un tel système finiront par converger vers un optimum, ce dernier ne sera atteint qu'avec des *erreurs inévitables.* Le créateur d'un tel système (qui, d'après Matthias, n'est pas vraiment un programmeur au sens classique) ne peut éliminer ces erreurs, car il faut les autoriser explicitement pour que le système puisse rester opérationnel et s'améliorer ;

c) *méthodes de programmation génétiques* : un niveau supplémentaire de code généré par la machine sépare le programmeur et le produit de sa programmation. Contrairement aux réseaux neuronaux, où le concepteur définit toujours les paramètres de fonctionnement du système (architecture du réseau, apports et produits, interprétation) et, au moins, le langage utilisé et la sémantique des symboles, le programmeur génétique perd jusqu'à ce degré minimal de contrôle, car il crée une machine qui se programme elle-même.

Dans le même temps, Matthias observe que les agents autonomes privent le programmeur du lien spatial avec l'agent artificiel qu'ils créent. L'agent artificiel opère hors du champ de vision du programmeur, qui peut même être incapable d'intervenir manuellement (en cas de faute ou d'erreur, laquelle peut se produire très longtemps après que l'agent a commencé à fonctionner). Ainsi, de tels processus font peu à peu perdre au concepteur la maîtrise de ses machines, qui passe progressivement aux machines elles-mêmes, si bien que – selon Matthias – le programmeur « n'est plus *codeur*, mais *créateur* d'organismes logiciels ». À mesure que décroît l'influence du créateur de la machine, celle de l'environnement augmente, jusqu'à ce que le programmeur ne puisse plus contrôler le produit, mais uniquement l'environnement (en particulier pour les machines qui continuent d'apprendre et de s'adapter au milieu dans lequel elles sont mises en œuvre). En particulier parce que ces agents devront interagir avec une diversité potentiellement croissante de personnes (utilisateurs) et de situations, leur créateur ne peut ni prédire, ni maîtriser l'influence de l'environnement. D'après Matthias, cela revient à dire que ces machines *échappent au contrôle de leurs créateurs* et peuvent donc causer des dommages dont il serait injuste de les tenir pour responsables. Pourtant, Matthias avance que nous ne pouvons nous passer de tels systèmes, et devons donc trouver des moyens de « combler le vide de responsabilité dans la pratique morale et dans la législation[179] ».

(b) Théories de la responsabilité morale fondées sur les choix

L'affirmation de Matthias – il serait « injuste » de tenir les créateurs de machines autonomes pour responsables des actions de ces dernières – repose sur une vision de la responsabilité morale « par les choix » qui tend à dominer le débat académique contemporain sur les implications morales et éthiques de l'IA. Selon cette conception, un comportement peut être blâmé à juste titre lorsqu'il constitue une faute, cette faute étant comprise comme librement choisie[180]. De ce point de vue, un agent (X) n'est moralement responsable d'un résultat indésirable (Y) que si X « a causé » Y. Pour qu'il soit reconnu que X a causé Y, X doit avoir eu un comportement engageant sa responsabilité causale. L'établissement de ce lien de cause à effet suppose que X ait *volontairement choisi* le comportement en question, même si ce comportement produit des conséquences et des effets que X ne prévoyait ou ne souhaitait pas. D'après Matthias, les développeurs d'agents informatiques capables de prendre leurs propres décisions d'une manière non programmée à l'avance n'ont pas le degré de maîtrise

179 Matthias 2004 : 183.

180 Wallace 1994, cité par Cane 2002.

nécessaire, et ne sont donc pas moralement responsables des décisions de ces agents informatiques et de leurs conséquences[181].

Cette affirmation, selon laquelle le caractère autonome des agents informatiques briserait la chaîne des causes reliant les actes de leurs développeurs aux décisions prises par ces agents, est hautement discutable[182]. Pour commencer, il est important de reconnaître que les théories de la responsabilité morale fondées sur le choix sont très mal adaptées à l'identification des responsabilités en cas d'*atteintes aux droits de l'homme*. De par leur nature même, les droits en général, et les droits de l'homme en particulier, protègent des valeurs d'une telle importance que toute violation entraîne une responsabilité en soi, *sans qu'une faute ne soit prouvée*[183]. Reprenons l'exemple du retrait par Facebook, en 2016, de la célèbre photographie d'une fillette vietnamienne. Soumis à une législation nationale obligeant les acteurs étatiques et non étatiques à respecter les droits de l'homme, Facebook serait considéré comme juridiquement responsable d'atteinte à la liberté d'expression, sans nécessité de démontrer que l'entreprise avait la main sur le retrait ou non de cette image. En d'autres termes, il y aurait eu atteinte au droit à la liberté d'expression même si la décision de retrait a été prise par un système algorithmique automatique agissant de manière indépendante et sans intervention humaine directe, et même si les concepteurs de ce système n'avaient pas prévu ou souhaité la suppression automatique de la photographie en question.

3.4 Modèles d'attribution de la responsabilité

Tandis que les violations des droits de l'homme sont largement comprises comme relevant d'un modèle de responsabilité « absolue », ou « responsabilité sans faute », les obligations de réparation en cas de *dommage matériel* à la santé ou aux biens peuvent être attribuées juridiquement selon plusieurs modèles différents. Parce que le fonctionnement des systèmes d'IA peut aboutir à la fois à des violations des droits de l'homme et à des dommages aux personnes et/ou aux biens, et parce que l'attribution de la responsabilité rétrospective en cas de dommages sert d'orientation à ceux qui conçoivent, développent, produisent et mettent en œuvre ces systèmes en précisant la nature et l'étendue de leurs obligations, nous allons brièvement présenter ces modèles. Lorsqu'un comportement nuit à autrui, la diversité des modèles juridiques qui peuvent servir à attribuer et à répartir les responsabilités montre clairement à quel point il serait erroné d'attendre d'un modèle de responsabilité unique qu'il puisse s'appliquer à tous les types de conséquences négatives que peuvent entraîner les technologies numériques avancées. Comme nous l'avons déjà noté, contrairement à l'acception philosophique de la responsabilité, qui tend à se concentrer sur les agents aux dépens des « victimes » et de la société, les modèles juridiques de la responsabilité[184] sont *relationnels*, au sens où ils s'intéressent non seulement à la situation des individus dont le comportement *donne lieu* à une responsabilité (les agents moraux), mais aussi à l'*impact* de ce

181 Matthias 2004. Pour une réaffirmation récente, voir Gunkel 2017.

182 Déterminer la responsabilité causale pour une action ou un événement est en soi une affaire d'interprétation, et non de « vérité » scientifique.

183 Voir Représentant spécial du Secrétaire général de l'ONU 2011 (« principes de Ruggie »).

184 En dehors du droit, la notion de « responsabilité » désigne beaucoup plus couramment le « comportement humain et les conséquences de ce comportement qui déclenchent des réactions », si bien que nous tendons à parler de « responsabilité morale » d'une part et de « responsabilité juridique » d'autre part, la seconde désignant avant tout les pénalités, amendes et sanctions formelles et institutionnalisées qui caractérisent le droit et les systèmes juridiques, mais non la morale (Cane 2002 : 1-2).

comportement sur autrui et sur la société en général[185]. Comme l'écrit le juriste et philosophe Peter Cane :

> La responsabilité ne dépend pas uniquement de la qualité de la volonté que manifeste un comportement, ni de la qualité de ce comportement. Elle tient aussi à l'intérêt que nous avons tous à assurer la sécurité des personnes et des biens, et au mode de répartition des ressources et des risques dans la société. La responsabilité est un phénomène relationnel[186].

En d'autres termes, la responsabilité juridique met en lumière les relations entre les agents moraux, les patients moraux et la société en général, au lieu de se centrer exclusivement sur le comportement des agents moraux et sur les responsabilités qui devraient ou non y être attachées. Par conséquent, les recherches sur les différentes manières, dans les ordres juridiques nationaux, de déterminer les responsabilités en cas de dommages ou d'autres événements néfastes (dont les atteintes aux droits, se traduisant ou non par des dommages), montrent que chacun des modèles appliqués trouve son propre équilibre entre les intérêts des agents moraux et des patients moraux (ou « victimes », selon l'appellation répandue dans le domaine du droit[187]). Cependant, la présente étude ne cherche pas à établir si les approches actuellement adoptées par les différents systèmes juridiques attribuent correctement les responsabilités au moyen des règles nationales de responsabilité civile, d'autant que la capacité du droit national à déterminer les responsabilités rétrospectives en cas de torts et de dommages entraînés par des systèmes d'IA n'a pas encore été pleinement éprouvée devant les tribunaux[188]. En revanche, les réflexions qui suivent exposent quatre grands modèles de responsabilité présents dans les systèmes juridiques anglo-américains : 1) les modèles fondés sur l'intention/la culpabilité, 2) les modèles fondés sur le risque/la négligence, 3) la responsabilité absolue et 4) les régimes d'assurance obligatoires[189], qui correspondent à différentes manières d'attribuer les responsabilités à l'égard des risques, des atteintes aux droits de l'homme et des dommages collectifs[190]. Ils n'ont qu'une vocation heuristique, celle de mettre en avant l'éventail de modèles qui pourraient servir à attribuer et à répartir les responsabilités à l'égard des risques et des dommages associés aux technologies numériques avancées[191]. Ces esquisses décrivent donc, de façon sélective, ce que j'appellerai le critère de

---

185 Cane 2002 : 4-5.

186 Cane 2002 : 109.

187 La Commission européenne entreprend actuellement de revoir ces thématiques. Voir par exemple Commission européenne 2018c.

188 Divers organismes s'efforcent actuellement d'évaluer la capacité des règles nationales en matière de responsabilité civile à couvrir correctement les dommages entraînés par des systèmes d'IA. Par exemple, la Commission européenne compte publier mi-2019 des orientations sur l'application de la Directive UE sur la responsabilité du fait des produits à l'intelligence artificielle, à la robotique et à l'internet des objets (Commission européenne 2018c).

189 Si cette étude dégage les différents modèles de responsabilité appliqués dans les systèmes juridiques anglo-américains, c'est simplement parce que l'auteure a été formée dans ce type de système et le connaît mieux. Il ne faut pas en déduire que ces modèles coïncideraient avec ceux utilisés dans d'autres systèmes juridiques, ni qu'ils seraient supérieurs aux modèles adoptés ailleurs.

190 D'après le Parlement européen dans son Rapport concernant des règles de droit civil sur la robotique, « la responsabilité civile des robots est une question cruciale à laquelle il importe de répondre au niveau de l'Union afin de garantir le même niveau de transparence, de cohérence et de sécurité juridique dans toute l'Union, dans l'intérêt tant des consommateurs que des entreprises ». Commission des affaires juridiques du Parlement européen, 2017 : 11.

191 Dans les systèmes juridiques anglo-saxons, la distinction entre droit civil et droit pénal est d'une importance capitale. Le premier objectif du droit pénal est d'imposer des peines et sanctions aux auteurs de comportements criminels, si bien que la responsabilité pénale dépend avant tout du comportement de l'auteur présumé et de son état mental. En revanche, le premier objectif du droit civil est de déterminer et

« contrôle » et le critère de « connaissance », applicables à chaque modèle, et n'abordent pas en détail la teneur de chaque modèle. Comme on le verra, chaque modèle ménage un équilibre différent entre notre intérêt à pouvoir agir librement et notre intérêt, en tant que victimes, à préserver nos droits, notre sécurité et celle de nos biens[192]. Par conséquent, identifier lequel de ces modèles (s'il y en a un) convient le mieux à l'attribution des responsabilités pour les différents risques découlant des technologies numériques avancées ne va pas de soi[193] ; au contraire, cette répartition des risques relève d'un *choix politique engageant toute la société*.

3.4.1 Modèles fondés sur l'intention/la culpabilité

Les modèles fondés sur l'intention/la culpabilité, qui sous-tendent tout le droit pénal, se concentrent sur le caractère volontaire du comportement de l'agent. Ils peuvent être interprétés comme exigeant que deux critères soient remplis. Premièrement, le critère de « contrôle » : l'agent doit avoir causé le comportement illicite en choisissant librement et sciemment de se comporter ainsi ; deuxièmement, le critère de « connaissance » : une « faute » doit être prouvée, ce qui suppose plus généralement que l'agent ait conscience des faits particuliers entourant les conséquences négatives de son comportement, et que ses actes puissent être analysés comme fondés sur ces faits[194]. C'est un modèle de ce type, fondé sur l'intention/la culpabilité, qui sous-tend les conceptions de la responsabilité morale « par les choix » qui prédominent dans le débat philosophique cherchant à savoir si les développeurs d'agents informatiques autonomes sont moralement responsables des actions de ces agents. Pour l'heure du moins, parce que ces agents informatiques sont dépourvus de connaissance subjective, de conscience et d'intention, ces modèles de responsabilité s'appliquent mal à de tels agents, qui ne peuvent remplir le critère de connaissance[195]. En revanche, ils peuvent s'appliquer aux développeurs ou aux utilisateurs de ces agents informatiques. Des individus développant et déployant sciemment des technologies d'IA à des fins dangereuses ou malveillantes, par exemple pour commettre une escroquerie ou s'approprier des biens, rempliraient clairement les critères d'établissement de la responsabilité en vertu d'un modèle fondé sur l'intention/la culpabilité. En pareilles circonstances[196], il y aurait à première vue violation des droits de l'homme (la preuve d'une intention subjective pourrait être apportée, mais serait inutile car la responsabilité juridique pour la violation de tels droits est généralement « absolue ») ainsi que, probablement, une responsabilité pénale pour atteinte à la personne (ou aux biens) et des obligations civiles de réparation ou d'indemnisation.

---

de répartir les obligations légales de réparation entre ceux reconnus comme juridiquement responsables de tel ou tel dommage. La responsabilité en droit civil a donc deux faces : elle s'intéresse non seulement au comportement de l'agent, mais aussi à l'impact de ce comportement sur autrui. Nous n'entrerons pas davantage dans les détails sur le fonctionnement des modèles de responsabilité fondés sur les fautes collectives, la négligence et la responsabilité absolue, ni sur les différences entre ces modèles en droit civil et en droit pénal. Pour un examen approfondi, voir Cane 2002.

192 Cane 2002 : 98.

193 Danaher 2016.

194 En droit anglo-américain, les éléments mentaux du critère de la faute juridique sont l'intention, la négligence, la connaissance/conviction et la malveillance. Voir Cane 2002 : 79.

195 Hildebrandt 2013 ; Himma 2009 ; Solum 1991 ; Gless et al. 2016 ; Andrade et al. 2007.

196 Le recours à des technologies d'IA pour commettre une infraction pénale pourrait être considéré à juste titre comme une circonstance aggravante : voir 6 2002. Voir aussi Hallevy 2015.

3.4.2 Modèles fondés sur le risque/la négligence

En droit anglo-américain, les modèles de responsabilité fondés sur le risque/la négligence sous-tendent l'obligation générale de prendre des mesures raisonnables pour éviter les dommages prévisibles. Ces modèles de responsabilité sont traditionnellement appliqués pour déterminer si des agents sont juridiquement tenus d'indemniser ceux qui ont subi un dommage du fait de leur manquement à cette obligation de diligence générale. Ce modèle englobe un « critère de contrôle » proche de celui qui s'applique aux modèles fondés sur l'intention/la culpabilité (avec quelques modifications[197]), dans la mesure où il doit être démontré que l'agent a causé le dommage ou le préjudice en question. Cependant, le critère de connaissance est beaucoup moins strict dans les modèles fondés sur le risque/la négligence que dans ceux applicables aux modèles fondés sur l'intention/la culpabilité. Par exemple, la responsabilité juridique d'une négligence en droit anglo-américain ne requiert pas la preuve de l'état mental de l'agent, et cherche à ménager un juste équilibre entre l'intérêt des agents (à agir librement) et l'intérêt des victimes (à assurer leur sécurité). Comme l'ont souligné les philosophes du droit, la responsabilité morale d'un agent peut être engagée sans qu'il ait une connaissance subjective des conséquences de son comportement[198]. John Oberdiek explique que les faits ont un poids moral : ils sont dotés d'une force normative qui rend plus ou moins tolérables les actions à venir – à condition qu'il soit raisonnablement possible de les découvrir[199]. Au moment de décider d'une action, Oberdiek souligne qu'on peut moralement attendre d'une personne ordinaire qu'elle s'assure d'une « connaissance suffisante » : elle ne peut être supposée connaître tous les faits ; mais elle ne peut non plus enfouir sa tête dans le sable et ne se fier qu'à sa vision subjective, sans chercher à connaître les faits pertinents.

Par conséquent, on ne peut appliquer un modèle de responsabilité fondé sur le risque/la négligence au développeur d'un agent ou système informatique que si les dommages engendrés par ce dernier étaient une *conséquence raisonnablement prévisible* des actions et décisions du système. Dans le droit anglo-américain en matière de négligence, la responsabilité juridique en cas de dommage ne pèse que sur ceux qui sont soumis à une obligation de diligence. Cette obligation existe lorsque, pour parler très globalement, une action risque de manière raisonnablement prévisible de porter préjudice à un tiers immédiat. La prévisibilité permet donc à la fois de définir les types de risques pour lesquels une personne peut être considérée comme juridiquement responsable, et d'encadrer les dommages pour lesquels sa responsabilité peut être engagée[200].

La « prévisibilité raisonnable » contribue également à définir *comment* une personne est censée agir. Qui se comporte en personne ordinaire, prenant raisonnablement soin d'éviter les risques prévisibles, remplit son obligation de diligence[201]. La prévisibilité raisonnable sert donc de pierre de touche pour déterminer si des activités à risque (comme la conduite), pouvant entraîner des dommages matériels pour autrui, donnent lieu à une obligation légale de diligence. Comme l'observe Oberdiek, cette norme de *common law* constitue aussi une norme

197 L'application du principe des « dommages éloignés » peut nier les liens de cause à effet en cas de négligence (Horsey et Rackley 2015, chapitre 9).

198 Hart 1968.

199 Oberdiek 2017 : 57.

200 Lorsque le dommage a été causé par une omission ou par l'inaction, ces critères se manifestent de façon particulière. Par exemple, nous sommes tenus de protéger autrui des risques engendrés par une source de danger que nous avons créée, ou parce que nous avons assumé la responsabilité des intérêts d'un tiers. Voir Lunney et Oliphant 2013, chapitre 9.

201 Oberdiek 2017 : 40.

morale juste et appropriée car dans le cas d'activités à risque, il est important que nous puissions nous demander mutuellement des comptes sur notre appréciation du risque. En d'autres termes, nous devons être capables de justifier cette appréciation d'une manière qui résiste à un examen moral[202].

Cependant, quand on cherche à savoir s'il est ou non raisonnablement prévisible qu'une action risquée provoque des dommages, on se heurte au problème dit de la « catégorie de référence ». Comme l'explique Oberdiek,

> le problème de la catégorie de référence tient […] essentiellement à la possibilité de redescription […]. Tout risque peut être redécrit à l'infini […] il n'y a pas de catégorie de référence correcte unique sur lesquelles les convictions crédibles pourraient s'appuyer.

En 2018, par exemple, une femme poussant un vélo, des sacs de courses suspendus au guidon, est morte après avoir été heurtée par un véhicule Uber. Le véhicule, qui roulait en pilotage automatique depuis 19 minutes, a pris cette personne pour une voiture (censée donc rester dans sa file), avant de reconnaître son erreur et de remettre les commandes à la conductrice quelques secondes avant la collision – qui n'a pas pu être évitée[203]. Il semble peu probable que les développeurs du véhicule aient pu raisonnablement prévoir que son système de capteurs par IA confondrait une femme poussant un vélo chargé de sacs avec un autre véhicule. En revanche, il paraît entrer dans les limites du prévisible que les capteurs du véhicule ne classifient pas correctement les objets aux formes inhabituelles rencontrés dans des conditions de conduite normales, et que des erreurs de ce type puissent entraîner des collisions fatales.

Dans le même temps, savoir si tel ou tel événement associé au fonctionnement d'un objet technologique donné est « raisonnablement prévisible » ne peut être que le produit de notre expérience et de notre exposition à ces événements. Pendant les premières phases de déploiement d'une nouvelle technologie, les attentes quant à son comportement (et à ses conséquences) sont relativement incertaines[204]. Cependant, à mesure que le temps passe, ses actions et ses comportements peuvent devenir plus familiers pour les développeurs, et donc plus susceptibles d'être considérés comme raisonnablement prévisibles. Par conséquent, les développeurs de ces technologies devraient être tenus pour responsables si, par négligence, ils ne prennent pas les mesures qui auraient évité des dommages[205]. Même ainsi, se pose la question de nos attentes envers l'industrie du numérique lorsqu'elle décide de faire fonctionner des technologies émergentes dans le monde réel : nous appliquons à juste titre

---

202 Oberdiek 2017 : 48.

203 Smith 2018.

204 Par exemple, Tay, robot expérimental lancé par Microsoft, était conçue pour apprendre à mener une discussion en termes humains en observant les utilisateurs de Twitter et en dialoguant avec eux. Elle était censée s'améliorer au fil des interactions et, ainsi, livrer des enseignements sur la capacité des programmes d'IA à entrer en conversation avec les internautes. Hélas, en imitant les utilisateurs de Twitter, elle a très vite appris à lancer des bordées d'injures et de propos antisémites et haineux, conduisant Microsoft à fermer son compte. Les développeurs de Tay n'avaient pas anticipé ce phénomène ; pourtant, on peut avancer qu'il était raisonnablement prévisible, compte tenu du volume et de la fréquence des posts injurieux sur Twitter. Voir The Guadian 2016.

205 Liu et Zaweiska 2017.

des régimes de contrôle stricts aux nouveaux médicaments, ne devrions-nous pas procéder de même avec les technologies numériques avancées lorsqu'elles sont risquées[206] ?

D'autres questions se présentent quant au degré minimum de diligence dont les développeurs de systèmes d'IA devraient faire preuve lors de la conception et de la mise en œuvre de systèmes informatiques autonomes. Prenons à nouveau la collision fatale entre un véhicule Uber et une piétonne poussant un vélo, qu'il avait classée à tort comme un véhicule arrivant en sens inverse. Dans les débats d'aujourd'hui, on entend souvent ce refrain : les voitures autonomes seront « plus sûres » que celles pilotées par des êtres humains, ce qui suggère que le bon point de comparaison serait un conducteur humain raisonnable. Est-il adéquat, cependant, d'appliquer à des dommages involontaires résultant des actions d'une voiture autonome le même modèle de responsabilité et la même exigence de diligence qu'à un conducteur ordinaire, au volant d'une voiture traditionnelle ? Ou devrions-nous plutôt appliquer, pour régir le développement et le fonctionnement des véhicules autonomes, le modèle de responsabilité habituellement utilisé pour les fabricants de produits, qui prévoit dans les systèmes juridiques européens d'aujourd'hui une responsabilité absolue (abordée ci-dessous) en cas de défauts ? En d'autres termes, des choix politiques importants doivent être opérés, et il n'est en aucun cas évident que le conducteur humain ordinaire constitue la référence la plus adaptée[207].

3.4.3 Responsabilité absolue

Comme nous l'avons déjà relevé, le modèle de responsabilité juridique applicable aux violations de droits (dont les droits de l'homme et les libertés fondamentales) est celui de la responsabilité absolue – « *strict legal liability* » en droit anglo-américain. Dans ce cas, la responsabilité pèse sur l'agent *sans* qu'une faute soit prouvée. Ceux qui causent des atteintes aux droits en sont juridiquement responsables *qu'ils aient ou non* enfreint une norme de comportement spécifiée par la loi et indépendamment de l'état mental qui a pu justifier ou accompagner leurs actions[208]. Des quatre variantes de la responsabilité absolue identifiées par Cane, trois sont directement pertinentes pour notre étude : la responsabilité absolue fondée sur les droits, sur les résultats et sur les activités.

(a) Responsabilité absolue *fondée sur les droits* : en cas d'atteinte à des droits reconnus par la loi. Toute ingérence dans la sphère de protection prévue par la loi déclenche la responsabilité. Exemple classique, l'intrusion : parce qu'elle porte atteinte au droit du propriétaire à être seul maître chez lui, toute intrusion sans son accord constitue une ingérence illicite même si la personne qui a pénétré sur le terrain n'était pas blâmable. Comme déjà noté, les atteintes aux droits de l'homme relèvent de cette catégorie.

(b) Responsabilité absolue *fondée sur les résultats* : en cas de résultats négatifs (conséquences imprévues, par exemple), avec ou sans faute. Les législations européennes contemporaines sur les produits défectueux reposent sur ce modèle, qui prévoit la responsabilité absolue des fabricants si des produits défectueux causent des dommages aux personnes ou aux biens[209]. S'agissant des technologies numériques avancées, la question est de savoir ce qui constitue un « défaut ». Reprenons la collision fatale entre un

206 Nemitz 2018 ; Thomas 2017a ; Thomas 2017b.
207 Thomas 2017b.
208 Cane 2002 : 82.
209 Voir Union européenne 1985.

véhicule Uber et une piétonne poussant un vélo : après avoir classé la piétonne comme un autre véhicule, le système a reconnu son erreur et aussitôt rendu les commandes à l'opératrice, mais trop tard pour qu'elle puisse éviter la collision. On pourrait avancer qu'en pareil cas, le véhicule n'était pas « défectueux », dans la mesure où il a fonctionné exactement comme ses développeurs l'avaient voulu. En revanche, si par « défectueux » on entend « inadapté au but visé », l'échec du véhicule à catégoriser correctement la piétonne et à s'en écarter pour éviter la collision mérite tout à fait la qualification de défaut[210]. Une approche similaire s'applique souvent lorsque le risque de dommages tient au comportement imprévisible de certaines sources de danger, comme les animaux. En pareil cas, la responsabilité revient aux personnes censées surveiller l'animal, considérées comme les mieux placées pour adopter des mesures visant à prévenir ou à atténuer le risque de dommages.

(c) Responsabilité absolue *fondée sur les activités* : en cas d'activité spécifique, comme dans les diverses infractions « de possession » : lois interdisant de posséder des armes à feu, des couteaux, des substances illicites, etc. En droit anglo-américain, la responsabilité du fait d'autrui est une forme importante de responsabilité absolue fondée sur les activités. L'activité concernée est avant tout définie en termes de relations avec une autre personne, qui, si elle viole la loi, engage la responsabilité absolue de la première personne. La responsabilité du fait d'autrui s'applique dans les relations de travail ; si un employé enfreint la loi dans le cadre de ses fonctions, l'employeur est responsable. Certains pays adoptent une approche de responsabilité absolue à l'égard de ceux qui mènent des activités dangereuses (qui dirigent une centrale nucléaire ou pilotent un avion, par exemple) ou de ceux qui sont responsables en dernier recours de cette activité dangereuse (le propriétaire du véhicule, par exemple). Ici, le raisonnement sous-jacent est que cette personne a créé un risque et, dans le même temps, tire des avantages économiques de l'activité en question[211].

Ces diverses formes de responsabilité absolue répartissent les risques associés aux activités potentiellement nuisibles entre agents et victimes en accordant un poids considérable aux intérêts des victimes (préserver leur sécurité et celle de leurs biens). Elles montrent ainsi que la responsabilité n'est pas seulement fonction de la qualité de la volonté d'un agent, telle qu'elle se manifeste dans son comportement, ni de la qualité de ce comportement : elle concerne aussi l'intérêt que nous avons tous à assurer la sécurité de notre personne et de nos biens, ainsi que les modes de répartition des ressources et des risques dans notre société, qui définissent les limites de nos responsabilités[212].

3.4.4 Régime d'assurance obligatoire

Au lieu d'insister sur l'attribution des responsabilités à ceux qui peuvent être vus comme ayant contribué aux dommages que peuvent générer les technologies numériques avancées, nous pourrions décider au contraire d'accorder la priorité à l'indemnisation financière de toutes les personnes lésées. Cela passerait par un régime d'assurance obligatoire (qui pourrait reposer

---

210 En cas de dommages causés par des robots autonomes, le Rapport de Parlement européen concernant des règles de droit civil sur la robotique privilégie la responsabilité sans faute (Commission des affaires juridiques du Parlement européen, 2017).

211 Commission européenne 2018b.

212 Cane 2002 : 108-109.

sur le principe « même en l'absence de faute »), avec instauration d'une caisse d'assurance à laquelle toutes les personnes lésées par ces technologies pourraient avoir recours[213]. Ce régime pourrait être financé de diverses manières, y compris par des contributions du secteur des technologies, les demandes étant traitées par une autorité indépendante ou publique. On pourrait aussi simplement obliger les entreprises intervenant dans la chaîne de valeur concernée à souscrire une assurance en responsabilité civile[214]. Bien qu'apprécier la pertinence de tels régimes n'entre pas dans le champ de notre étude, ils présentent l'avantage de permettre aux personnes affectées par ces technologies de prétendre à une indemnisation lorsqu'il est difficile d'identifier précisément quelles entreprises devraient être tenues pour responsables des dommages, ou lorsque les entreprises en question ne sont plus solvables. Cette idée pourrait avoir de beaux jours devant elle ; en effet, nous comptons de plus en plus sur des systèmes intelligents autonomes qui continuent de fonctionner longtemps après la disparition de leurs développeurs ou des entreprises qui les ont créés, si bien que nos sociétés vont peut-être devoir mettre en place des institutions inscrites dans la durée, comme un régime d'assurance collectif, pour que les victimes ne restent pas systématiquement sans indemnisation[215]. Dans ce contexte, certains ont proposé de conférer un statut juridique aux machines intelligentes de manière à faciliter l'administration des versements des indemnités aux victimes[216].

3.5 Défis liés à la complexité des systèmes sociotechniques

À des fins d'analyse, nous avons jusqu'ici supposé que, s'agissant de situer les responsabilités à l'égard des effets négatifs des technologies numériques avancées, les liens de cause à effet étaient aisément identifiables. En pratique cependant, ces technologies constituent un élément essentiel de systèmes sociotechniques hautement complexes et sophistiqués, si bien que l'identification des responsabilités causales, morales et juridiques soulève d'énormes défis. Trois d'entre eux sont brièvement présentés ci-dessous : le problème des acteurs multiples, celui des « humains dans la boucle » et les effets imprévisibles des dynamiques complexes qui peuvent se créer lorsque de multiples systèmes algorithmiques interagissent entre eux.

---

213 La Commission des affaires juridiques du Parlement européen a recommandé ce type de solution pour les dommages causés par certaines catégories de robots : un régime d'assurance obligatoire devrait être mis en place ; il pourrait reposer sur l'obligation des entreprises de souscrire une assurance pour les robots autonomes qu'elles fabriquent, et être complété par un fonds afin de garantir un dédommagement y compris en l'absence de couverture (Commission des affaires juridiques du Parlement européen, 2017, par. 56-57).

214 Commission européenne 2018b.

215 6 : 2001, 429.

216 Par exemple, la Commission des affaires juridiques du Parlement européen a appelé la Commission européenne à envisager la création, à terme, d'une personnalité juridique spécifique aux robots, pour qu'au moins les robots autonomes les plus sophistiqués puissent être considérés comme des personnes électroniques responsables de réparer tout dommage causé à un tiers ; il serait envisageable de considérer comme une personne électronique tout robot qui prend des décisions autonomes ou qui interagit de manière indépendante avec des tiers (Commission des affaires juridiques du Parlement européen, 2017, par. 59). Le département du Parlement européen sur les droits des citoyens et les affaires constitutionnelles (Commission des affaires juridiques) s'est fermement opposé à cette proposition (Nevejans 2016, pp. 16-18). De telles propositions occupent une place à part, distincte du débat académique qui cherche à établir si les robots devraient être considérés comme des agents *moraux* dont les droits devraient être protégés. L'examen du statut juridique et moral adapté aux systèmes d'IA en tant qu'agents indépendants dépasse le champ de notre étude. Voir Solum 1991 ; Koops 2010 ; Teubner 2006 ; Teubner 2018.

3.5.1 Le problème des acteurs multiples

Sauf sous certaines formes de responsabilité absolue, l'attribution des responsabilités pour les risques potentiels ou avérés, les dommages et les atteintes aux droits (y compris les droits de l'homme) suppose de déterminer s'ils peuvent être compris comme *causés* par l'agent. Or, dès qu'on cherche à établir les responsabilités causales ayant conduit à un événement néfaste[217] pouvant être raisonnablement considéré comme une conséquence directe du fonctionnement d'un système sociotechnique complexe (qu'il utilise ou non l'IA), on se heurte au problème des « acteurs multiples[218] ». Ce problème se pose lorsqu'on adopte un modèle de responsabilité fondé sur l'intention/la culpabilité. D'abord identifié dans le contexte des technologies de l'information par Helen Nissenbaum, philosophe des technologies[219], le problème des acteurs multiples ne se limite pas aux ordinateurs, aux technologies numériques, aux algorithmes ou à l'apprentissage automatique. Il se présente dès qu'un éventail complexe d'individus, d'organisations, de composantes et de processus participe au développement, au déploiement et à la mise en œuvre de systèmes complexes ; lorsque ces systèmes dysfonctionnent ou entraînent des dommages, il devient très difficile d'identifier les fautifs, car les concepts utilisés s'inscrivent traditionnellement dans une conception individualiste de la responsabilité[220]. En d'autres termes, lorsque des systèmes technologiques complexes entrent en jeu, la responsabilité causale ne peut être que fractionnée, et le lien de cause à effet se dilue en une simple influence[221].

Le problème des acteurs multiples devient particulièrement critique lorsqu'on cherche à situer les responsabilités à l'égard des dommages ou des torts résultant du développement et du fonctionnement des systèmes d'IA. Ils dépendent en effet de plusieurs composantes essentielles, à savoir :

(a) les *modèles* développés pour représenter l'espace de caractéristiques du système et l'objectif global qu'il doit atteindre de façon optimale ;

(b) les *algorithmes*, fondés sur ces modèles, qui analysent les données pour produire des résultats pouvant déclencher telle ou telle « action » ou décision ;

(c) les *données* (comprenant ou non des données personnelles) fournies à ces algorithmes pour les entraîner ;

(d) les *développeurs* qui conçoivent ces systèmes, et doivent prendre des décisions engageant certaines valeurs concernant les modèles, les algorithmes et les données utilisées pour les entraîner. Les personnes chargées d'étiqueter les données servant à entraîner les algorithmes sont aussi concernées[222] ; et enfin,

(e) le *contexte et le système sociotechnique plus larges* dans lesquels le système algorithmique s'inscrit et fonctionne.

---

217 Cet événement négatif peut être un risque/dommage structurel, un dommage individuel ou une violation individuelle des droits de l'homme, n'entraînant pas nécessairement de perte ou dommage matériel ou d'atteinte à des intérêts collectifs.

218 Thompson 1980.

219 Nissenbaum 1996.

220 Thompson 1980.

221 Liu et Zaweiska 2017.

222 Zalnieriute et al. 2019.

Même en supposant que nous puissions attribuer de façon satisfaisante les responsabilités morales pour les impacts négatifs de chacun de ces éléments, il est peu probable que nous puissions facilement retracer les responsabilités morales en cas de conséquences négatives imprévues lorsque ces éléments s'associent dans un système complexe, intégré et en évolution. Ces difficultés sont encore renforcées par le fait que les produits et services numériques font l'objet d'extensions, de mises à jour et de correctifs après leur déploiement. Tout changement dans le logiciel du système peut affecter le comportement de tout le système ou d'une ou l'autre de ses composantes, en enrichissant leurs fonctionnalités, et modifier le profil de risque opérationnel du système, dont l'éventualité qu'il entraîne des dommages ou porte atteinte aux droits de l'homme[223].

Pour relever ces défis, on gagnera à garder à l'esprit trois considérations. Premièrement, les problèmes d'attribution des responsabilités juridiques pour des dommages découlant d'activités impliquant de multiples acteurs ne sont pas nouveaux, si bien que de nombreux systèmes juridiques disposent d'un ensemble assez étoffé de principes et de procédures visant à déterminer les responsabilités lorsqu'il existe de nombreux accusés potentiels[224]. Comme l'a récemment observé la Commission européenne, pour que les victimes soient indemnisées, il n'est peut-être pas utile d'identifier la répartition des obligations de réparation entre les multiples acteurs de la chaîne de valeur qui produit les technologies numériques émergentes, même si répondre à ces questions présente un intérêt du point de vue des politiques générales, pour offrir une sécurité juridique à ceux qui participent à la production et à la mise en œuvre de ces technologies[225]. Deuxièmement et de ce fait, si le droit s'avère capable de trouver des réponses pratiques au problème des acteurs multiples, pourtant inextricable en apparence, c'est en partie parce qu'il insiste sur l'intérêt légitime des patients moraux à préserver la sécurité de leur personne, au lieu de se centrer presque exclusivement sur l'agent moral comme le font les théories de la responsabilité morale par le choix (sous-jacentes au problème des acteurs multiples). Troisièmement, étant donné que ce rapport s'intéresse aux *violations des droits de l'homme* plutôt qu'aux *dommages*, il est particulièrement important de veiller à ce que nous disposions de mécanismes effectifs et légitimes destinés à *prévenir* et à *empêcher* les violations des droits de l'homme associées au fonctionnement des technologies numériques avancées, d'autant que beaucoup de ces violations ne se traduisent pas nécessairement en atteintes matérielles à la santé ou aux biens. La vitesse et l'étendue que ces technologies ont acquises aujourd'hui rendent cette approche préventive encore plus nécessaire. Les effets cumulés des atteintes aux droits de l'homme causées par les systèmes d'IA pourraient gravement saper les bases sociales nécessaires aux ordres moraux et démocratiques, eux-mêmes préalables essentiels à l'existence de ces droits ; d'où la nécessité, peut-être, de renouveler les approches existantes de la protection des droits de l'homme à l'heure des réseaux et des données[226].

3.5.2 L'interaction humain-machine

Bien que de nombreux individus, entreprises et autres organisations participent au développement et à la mise en œuvre des technologies numériques avancées, ces dernières

223 Thomas 2015.

224 Voir par exemple les modèles de responsabilité partagée appliqués par les plates-formes d'hébergement en ligne : De Streel, Buiten et Peitz 2018 ; Helberger et al. 2018.

225 Commission européenne 2018b : 20-21

226 Voir le chapitre 3.8, plus loin.

sont souvent conçues pour fonctionner sans intervention humaine active[227]. Compte tenu des interactions complexes entre les êtres humains et les machines, répartir correctement l'autorité et la responsabilité entre eux soulève de sérieux défis. En particulier, beaucoup de tâches auparavant assurées par des êtres humains sont désormais réalisées par des machines, et pourtant, des humains participent inévitablement à différents points de la chaîne de développement, de tests, de mise en œuvre et de fonctionnement. Comme l'observe la Royal Academy of Engineering :

> Il y aura toujours des êtres humains dans la chaîne, mais en cas de préjudice, il est difficile de savoir lequel sera responsable – le concepteur, le fabricant, le programmeur ou l'utilisateur[228] ?

L'interaction entre des êtres humains et des machines au sein de systèmes sociotechniques complexes et en évolution pose des questions particulièrement épineuses sur le rôle de supervision joué par les humains. Préoccupation récurrente : pour veiller à ce que les systèmes sociotechniques de plus en plus complexes intégrant l'IA restent au service de l'humanité, ils devraient toujours être conçus de manière à ce qu'un opérateur puisse les désactiver. Pourtant, comme l'observe là encore la Royal Academy of Engineering :

> On pourrait penser qu'une intervention humaine est toujours nécessaire, mais ce sont parfois les systèmes autonomes qui sont nécessaires, lorsque des êtres humains risquent de faire de mauvais choix sous l'effet de la panique (en particulier dans des situations stressantes) ; ici, une reprise de contrôle par l'humain serait problématique. Les opérateurs humains n'ont pas toujours raison, et leurs intentions ne sont pas toujours bonnes. Se pourrait-il que dans certaines situations, les systèmes autonomes s'avèrent plus fiables que les opérateurs humains[229] ?

Par ailleurs, même si des êtres humains se trouvent toujours « dans la boucle » afin de surveiller les systèmes informatiques, les individus placés dans une telle situation hésitent à intervenir – et on peut les comprendre. Voici ce qu'observaient déjà Johnson et Powers[230] il y a plus de dix ans :

> Si le contrôle aérien devait être automatisé, […] il serait difficile de décider si et quand les contrôleurs humains doivent intervenir dans le contrôle par ordinateur. […] Les personnes qui assumaient auparavant la responsabilité par rôles pour les tâches qu'elles accomplissaient, soit seront remplacées par des préposés à l'entretien des technologies, soit le deviendront elles-mêmes. Dans un tel environnement, les êtres humains chargés d'interagir avec les systèmes

---

227 La supervision humaine peut être assurée par plusieurs mécanismes de gouvernance, comme ceux des humains « dans la boucle » (*human-in-the-loop*, HITL), « sur la boucle » (*human-on-the-loop*, HOTL) ou « aux commandes » (*human-in-command*, HIC). L'approche HITL désigne la possibilité d'une intervention humaine à tous les cycles décisionnels du système, intervention qui n'est dans de nombreux cas ni possible ni souhaitable. L'approche HOTL désigne la possibilité d'une intervention humaine au cours du cycle de conception du système et du suivi de son fonctionnement. L'approche HIC désigne la possibilité de superviser l'activité globale du système d'IA (y compris son impact économique, sociétal, juridique et éthique au sens large) et celle de décider quand et comment utiliser le système dans telle ou telle situation. Cela peut englober la décision de ne pas utiliser de système d'IA dans une situation donnée, de prévoir certains niveaux de latitude humaine lors de l'utilisation du système ou de permettre de passer outre une décision prise par le système. Voir Groupe d'experts de haut niveau de l'UE 2019a : 16.

228 Royal Academy of Engineering 2009 : 2.

229 Royal Academy of Engineering 2009 : 3.

230 Johnson et Powers 2005 : 106.

> « automatiques » pourraient percevoir toute intervention comme moralement risquée. Ils pourraient raisonner ainsi : mieux vaut laisser l'informatique agir et ne pas s'en mêler. Intervenir dans le comportement de systèmes informatiques automatisés, c'est mettre en doute la sagesse des concepteurs du système et les « connaissances » du système lui-même. Dans le même temps, quiconque choisit d'intervenir dans le système prend sur elle ou sur lui le lourd poids de la responsabilité morale ; les contrôleurs seront donc incités à laisser le système fonctionner en mode automatique. Les êtres humains ne voudront pas être responsables, ce qui marquera, d'une certaine manière, un transfert de la responsabilité vers le système informatique[231].

Pourtant, alors que nous devenons de plus en plus dépendants du large éventail de services et de systèmes que l'automatisation rend possibles, et en particulier à mesure que les technologies numériques gagnent en puissance et en sophistication, continuer à insister pour qu'un être humain se trouve « dans la boucle » pour assurer une supervision risque de transformer les intéressés en « amortisseurs moraux », en totems dont le rôle central deviendra de prendre la faute sur eux, même s'ils ne maîtrisent que partiellement le système, et susceptibles de servir de boucs émissaires aux entreprises et organisations cherchant à se dégager de leurs responsabilités[232]. Comme le montre l'étude d'Elish et Twang sur les contentieux autour du pilote automatique dans l'aviation, les aéronefs modernes sont aujourd'hui largement contrôlés par des logiciels et pourtant, les pilotes qui se trouvent dans le cockpit restent juridiquement responsables de leur fonctionnement. Cependant, nos perceptions culturelles semblent montrer un « parti pris pour l'automatisation », une croyance en la fiabilité et l'infaillibilité des technologies automatiques, les erreurs étant reprochées aux êtres humains (voir l'encadré 2[233]).

Encadré 2 : Parti pris pour l'automatisation et responsabilité des humains dans la boucle

La collision provoquée par une voiture Tesla en mode semi-automatique illustre la tendance à reprocher les conséquences imprévues aux êtres humains placés dans la boucle, et non à l'environnement sociotechnique de ces personnes.

En mai 2016, une Tesla en mode semi-automatique est rentrée dans un camion que sa fonction Autopilot n'avait pas détecté. L'enquête officielle a révélé que le pilote automatique avait fonctionné de manière conforme à sa conception mais n'avait pas détecté le camion. Le

231 Sur la question de savoir jusqu'où les êtres humains peuvent raisonnablement transférer des prises de décisions à un ordinateur sans se réserver la responsabilité de les superviser, voir Kuflik 1999.

232 Elish 2016.

233 Elish cite la tragédie du vol Air France 447, en 2009 (l'avion s'était abîmé dans l'océan Atlantique, tuant les 228 personnes à bord) comme exemple classique du rôle d'« amortisseurs moraux » dévolu aux pilotes. Parti du Brésil pour gagner la France, l'avion avait traversé une tempête et des cristaux de glace s'étaient formés sur les sondes Pitot, éléments du système d'avionique qui mesure la vitesse de l'air. Les sondes gelées avaient transmis des données erronées au pilote automatique qui, en retour, avait réagi exactement comme il était censé le faire en l'absence de données : il s'est automatiquement déconnecté, rendant la maîtrise de l'aéronef aux pilotes. Pris par surprise, les pilotes ont dû faire face à une avalanche d'informations – voyants qui clignotent, alarmes qui retentissent et affichages confus sur les instruments ; le rapport officiel français conclut à la « perte du contrôle cognitif de la situation », une série d'erreurs et de manœuvres incorrectes de la part des pilotes ayant conduit à l'accident fatal. Elish observe que la couverture médiatique réservée au rapport sur l'accident a souligné les erreurs des pilotes, mais n'a pas attiré l'attention sur le fait que ces erreurs s'expliquaient au moins partiellement par l'automatisation, qui modifie la nature du contrôle pouvant être exercé par un opérateur humain et ouvre la voie à de nouveaux types d'erreurs (Elish 2016, *ibid.*).

conducteur n'avait pas réagi ; l'enquête a conclu qu'il s'était trop fié à l'automatisation et au suivi du couple du volant, qui ne constituaient pas des méthodes efficaces pour maintenir l'attention du conducteur.

L'autorité chargée de l'enquête a conclu que la collision ne résultait pas d'un défaut spécifique du pilote automatique, et que Tesla n'était donc pas responsable de l'accident. Étant donné que Tesla avait dûment averti ses clients, indiquant que le système Autopilot devait rester sous la supervision du conducteur et que ce dernier devait garder les mains sur le volant et les yeux sur la route, la responsabilité revenait au conducteur. Par ailleurs, les conditions d'utilisation de Tesla comportaient des dispositions mentionnant la nature semi-autonome du pilote « automatique », et affirmant que le conducteur devait reprendre le contrôle du véhicule dans les 4 secondes s'il remarquait un comportement problématique.

Source : Commission européenne, Staff Working Document, 'Liability for Emerging Digital Technologies' (avril 2018), 14-15.

3.5.3 Interactions imprévisibles et changeantes entre systèmes sociotechniques complexes

La situation se complique encore lorsqu'il s'agit d'identifier, d'anticiper et de prévenir les effets néfastes des *interactions* entre des systèmes sociotechniques complexes et guidés par des algorithmes, fonctionnant à une vitesse et à une échelle qui n'étaient tout simplement pas possibles avant l'ère du numérique et des réseaux. Le « krach éclair » de 2010, qui a vu le cours de la Bourse chuter vertigineusement pendant cinq minutes avant de se reprendre sans raison apparente, en donne une très bonne illustration [234]. Tandis que les agents d'IA individuels, capables d'apprendre à partir de leur environnement et de s'améliorer par itération, peuvent faire l'objet de tests et de vérifications mathématiques, l'*interaction* de multiples algorithmes avec d'autres agents algorithmiques au sein d'un écosystème complexe et dynamique risque de générer des résultats imprévisibles et potentiellement dangereux. En d'autres termes, ces interactions créent des risques dont nous commençons à peine à prendre la mesure[235]. Le défi consiste donc à mettre au point des solutions pour prédire de façon fiable, modéliser et prévenir les résultats indésirables, et potentiellement catastrophiques, des interactions entre des systèmes sociotechniques complexes et changeants – ce qui fixe à la recherche en informatique un horizon nouveau et de plus en plus urgent. Shadbolt et Hampson, spécialistes renommés des sciences informatiques, mettent en garde contre des « systèmes hypercomplexes et super-rapides » générant de nouveaux risques considérables, et face auxquels

> nous devons nous montrer vigilants, intelligents et inventifs. Tant que nous le resterons, nous garderons le contrôle sur les machines et en tirerons de grands bénéfices. Pour cela, il nous faut développer des cadres politiques. Au-delà des dangers, c'est tout un monde de possibilités qui s'ouvre[236].

234 Akansu 2017.

235 Smith 2018.

236 Shadbolt et Hampson 2018.

3.6 Obligation de l'État d'assurer une protection effective des droits de l'homme

L'une des plus grandes préoccupations entourant l'émergence des systèmes algorithmiques tient au pouvoir croissant des géants du numérique, et notamment au radical déséquilibre de pouvoir entre ces entreprises et les individus qui leur sont soumis[237]. Le pouvoir de déployer des systèmes algorithmiques réside presque entièrement entre les mains de ces entreprises. En revanche, l'obligation de protéger les droits de l'homme en droit international repose avant tout sur les États, étant donné que la protection des droits de l'homme est avant tout pensée comme verticale – il s'agit de protéger les individus contre les ingérences injustifiées de la part de l'État. Il est bien établi par la jurisprudence de la CEDH, par ailleurs, que les droits protégés par la Convention imposent aux États membres des obligations positives, consistant à agir pour assurer à toute personne relevant de leur juridiction les droits et libertés définis par la Convention[238]. La CEDH impose donc aux États d'adopter la législation et les autres politiques nationales nécessaires pour veiller à ce que les droits inscrits dans la CEDH soient dûment respectés, y compris en les protégeant des ingérences de la part de *tiers* (y compris les entreprises), qui peuvent donc être juridiquement contraints de respecter les droits de l'homme[239]. C'est grâce à ces obligations juridiques contraignantes ancrées dans la protection des droits de l'homme, dont le droit à un recours effectif, que la Convention offre une base solide pour imposer des mécanismes effectifs à même d'amener les auteurs d'atteintes aux droits de l'homme à rendre des comptes, bien au-delà de ce qu'on peut attendre des discours sur l'« éthique de l'IA » et de l'autorégulation de l'industrie du numérique[240].

Le débat sur les différents modèles d'attribution de la responsabilité rétrospective, présenté au chapitre 3.2, puise largement dans les approches juridiques anglo-américaines, soit instaurées par la législation (et appliquées par les tribunaux), soit développées par les tribunaux lorsqu'ils interprètent et appliquent le droit existant pour déterminer les responsabilités en cas de dommages et autres préjudices. L'un des défauts des voies judiciaires est qu'elles conviennent mieux à la réparation de dommages importants subis par quelques-uns qu'à celle de dommages moins importants subis par le grand nombre. Avec l'IA, la difficulté à détecter les dommages et à déterminer et prouver les liens de cause à effet complique encore le recours aux tribunaux, sans parler des considérables obstacles pratiques rencontrés par les personnes souhaitant porter plainte et des mesures prises pour les en dissuader[241]. Dans le même temps, nous avons déjà souligné la capacité des systèmes d'IA, dans un monde interconnecté, à faire obstacle à l'action collective, et donc l'importance d'organismes nationaux de mise en œuvre dotés des pouvoirs nécessaires ; cette situation suggère également que des mécanismes de réclamation accessibles et pratiques pourraient s'avérer nécessaires pour que les atteintes aux droits de l'homme provoquées par des systèmes d'IA soient dûment sanctionnées. Il est important de reconnaître, par ailleurs, qu'il existe aux côtés des voies de justice classique de nombreux *autres mécanismes de gouvernance institutionnelle* qui pourraient aider à assurer la conformité aux droits de l'homme dans le développement et la mise en œuvre des technologies numériques avancées. Le chapitre qui suit offre donc un aperçu des autres mécanismes institutionnels envisageables

237 Ibid, Schwab et al. 2018 ; The Economist 2018b.

238 Rainey, Wicks et Ovey 2014 : 102.

239 La nature et l'étendue de la protection requise dependent du droit concerné (*ibid*.).

240 La procédure de l'arrêt pilote de la Cour européenne des droits de l'homme offre un mécanisme institutionnel permettant de demander aux États d'adopter des mesures correctives dans leur ordre juridique interne afin de mettre un terme aux violations constatées par la Cour, sous la supervision du Comité des Ministres. Voir Glas 2014.

241 Mantelero 2018 : 55.

(en dehors des initiatives d'autorégulation qui sont en train d'apparaître) pouvant aider à mieux définir les responsabilités prospectives et rétrospectives à l'égard des risques, dommages et autres torts suscités par les technologies numériques avancées. Il met brièvement en avant plusieurs mécanismes et institutions de gouvernance qui pourraient jouer un rôle précieux dans la responsabilisation en cas d'atteintes aux droits de l'homme, et compléter les mécanismes juridiques existants.

3.7 Rôle des mécanismes extrajudiciaires

Bien que les mécanismes de régulation puissent être catégorisés de nombreuses manières différentes, deux grands traits sont à souligner dans le cadre de cette étude. Premièrement, il existe des mécanismes *a priori*, qui surveillent et évaluent un objet, un processus ou un système avant sa mise en œuvre en milieu réel et assurent donc avant tout la responsabilité prospective. Deuxièmement, les mécanismes *a posteriori* s'appliquent pendant ou après la mise en œuvre et assurent donc avant tout la responsabilité rétrospective. Comme cette étude l'a déjà souligné, le développement et la mise en œuvre responsables de systèmes d'IA supposent de prendre en compte ces deux aspects. Cependant, notre étude s'intéresse avant tout aux incidences de ces technologies sur les droits de l'homme ; et il est extrêmement important de *prévenir* et d'*empêcher* les atteintes à ces droits à travers des mécanismes effectifs et légitimes, surtout compte tenu de la vitesse et de l'échelle auxquelles les systèmes d'IA fonctionnent aujourd'hui, combinées à la culture de l'« avancer vite et casser les codes » qui caractérise la stratégie opérationnelle des géants du numérique. Cette stratégie consiste à prendre de l'avance en innovant sans cesse, sans prêter grande attention aux risques potentiels : on préfère traiter les « effets indésirables » après coup, alors qu'il est devenu pratiquement impossible de retirer les innovations technologiques mises sur le marché[242]. Deuxièmement, il faut s'intéresser au caractère *juridiquement contraignant* des institutions et mécanismes de régulation, pour savoir si, et dans quelle mesure, ils peuvent être considérés comme des mécanismes facultatifs que les entreprises sont libres d'appliquer sélectivement ou d'ignorer totalement, ou s'ils se fondent sur le droit, les contrevenants s'exposant à des sanctions judiciaires substantielles[243]. Troisièmement, bien que les mécanismes de régulation émanent traditionnellement de la société, les dispositifs de protection technique, fondés sur ce qu'on appelle parfois la « régulation dès la conception[244] », peuvent être tout aussi importants (sinon plus) dans le présent contexte. Le chapitre suivant leur est consacré.

3.7.1 Dispositifs de protection techniques

Avec la prise de conscience croissante des questions éthiques et juridiques soulevées par l'usage des technologies d'IA, l'un des champs de recherche les plus prometteurs parmi ceux qui sont apparus réside dans les réponses *techniques* visant à « implanter » certaines valeurs dans la conception et le fonctionnement des algorithmes[245]. L'une des caractéristiques souvent associées à ces dispositifs de régulation « dès la conception » est leur aptitude à fonctionner en temps réel, plutôt qu'a priori ou a posteriori[246]. Les premiers travaux dans ce domaine – l'utilisation de mesures techniques pour protéger des intérêts et des valeurs à travers les TCI –

242 Taplin 2018 ; Vaidhyanathan 2011.
243 Nemitz 2018.
244 Yeung 2015.
245 *Ibid*.
246 *Ibid*.

se centraient sur la protection des droits de propriété intellectuelle[247] ; mais en parallèle, d'autres se sont penchés sur la confidentialité des données, donnant lieu au mouvement de la « vie privée dès la conception » ou de la « protection des données dès la conception ». Ces travaux reconnaissent que la technologie peut être mise au service d'intérêts et de valeurs au lieu de les menacer, et cherchent à renforcer le poids des normes juridiques en matière de propriété intellectuelle et de protection des données en implantant ces normes dans l'architecture même des systèmes d'information[248]. Outre les travaux d'« ingénierie de la vie privée », des recherches plus récentes sur l'apprentissage automatique et les logiciels prolongent cette approche en cherchant à assurer la « protection des droits de l'homme dès la conception ». Ce sont notamment les suivants.

(a) IA explicable (XAI) : les progrès des systèmes d'apprentissage automatique, dont ceux reposant sur les réseaux neuronaux, sont souvent utilisés pour assister les prises de décisions[249] ; or, leur logique n'est ni facile à expliquer (nous ignorons pourquoi ils ont opéré un choix plutôt qu'un autre), ni facile à interpréter (ils sont incapables d'expliquer ou de présenter leurs résultats de manière compréhensible pour un être humain). C'est pourquoi on admet de plus en plus la nécessité de pouvoir rendre les résultats générés par les systèmes d'IA intelligibles pour les utilisateurs[250], ce qui a donné lieu à un important domaine de recherche en informatique : l'« IA explicable » (XAI[251]).

(b) Équité, responsabilité et transparence de l'apprentissage automatique (FATML) (pour *Fairness, Accountability and Transparency in Machine Learning*) : de même, une communauté grandissante de chercheurs en apprentissage automatique s'intéresse au développement de techniques visant à identifier et à surmonter les problèmes de « discrimination numérique[252] », c'est-à-dire les partis pris et la discrimination découlant de l'exploitation des données et des autres techniques d'apprentissage automatique (techniques d'apprentissage automatique « équitables » ou « anti-discrimination[253] »).

3.7.2 Instruments et techniques de régulation

Des instruments de gouvernance plus classiques, de nature sociale et organisationnelle, sont également apparus face au risque que les technologies d'IA contreviennent à des valeurs importantes ; certains de ces instruments visent explicitement à s'assurer que ces systèmes technologiques respectent les droits de l'homme. Deux d'entre eux sont brièvement abordés ici : les analyses de l'impact sur les droits de l'homme et les techniques d'audit des algorithmes[254].

(a) Analyses de l'impact des algorithmes / sur les droits de l'homme : plusieurs spécialistes et organismes ont proposé diverses formes d'« études d'impact » appliquées aux algorithmes. Ces modèles d'analyse des risques s'adressent à ceux qui souhaitent proposer

---

247 D'abord appelés *Electronnic Copyright Management Systems* (ECMS), ils ont ensuite pris le nom de *Digital Rights Management Systems* (DRMS).

248 Bygrave 2017.

249 Voir par exemple Doshi-Velez, Ge, & Kohane, 2013 ; Carton et al. 2016.

250 Weller 2017 ; Yeung et Weller 2018b.

251 Voir par exemple Samek et al. 2017 ; Wierzynski 2018.

252 Barocas et Selbst 2016 ; Criado et Such 2019 ; Zliobaite 2015.

253 Voir en particulier les rencontres annuelles organisées par le FAT/ML (http://www.fatml.org/) et les ressources correspondantes : http://www.fatml.org/resources/relevant-scholarship.

254 Voir Assemblée générale des Nations Unies 2018, où les deux techniques sont soutenues.

ou déployer des systèmes algorithmiques en identifiant les incidences de ces systèmes sur les droits de l'homme, l'éthique et la société, et prendre des mesures pour lever leurs inquiétudes quant à la conception et au fonctionnement des systèmes algorithmiques avant leur mise en œuvre. Plusieurs modèles d'analyses d'impact générales, mais aussi spécifiques à certains domaines, ont été proposés[255].

Ces modèles diffèrent largement sur plusieurs aspects :

- *critères d'évaluation :* le droit européen sur la protection des données impose désormais des « analyses d'impact relatives à la protection des données(AIPD[256]) » dans certaines circonstances, dans le prolongement des « analyses d'impact sur la vie privée » déjà existantes, toujours largement centrées sur l'évaluation des impacts sur la qualité et la sécurité des données. D'autres modèles, comme l'« analyse de l'impact sur les droits de l'homme[257] », évaluent de manière plus générale l'impact d'un système proposé sur les droits de l'homme[258] ;

- *partie chargée de l'analyse :* certains modèles sont conçus pour être appliqués par le contrôleur des données (comme les AIPD), tandis que d'autres proposent que l'analyse soit effectuée par un tiers extérieur ou un organisme d'accréditation, approche retenue par les Principes directeurs des Nations Unies relatifs aux entreprises et aux droits de l'homme concernant la « diligence raisonnable en matière de droits de l'homme[259] » ;

- *caractère obligatoire ou facultatif :* certaines analyses de l'impact des algorithmes/sur les droits de l'homme proposées sont de nature facultative : le contrôleur des données choisit de les entreprendre ou non et détermine, le cas échéant, les mesures à adopter à l'issue de l'analyse[260]. D'autres, comme les AIPD, sont imposées par la loi dans certaines circonstances[261] ;

- *champ de l'évaluation :* tandis que l'analyse de l'impact sur les droits de l'homme examine un large éventail d'activités pour évaluer leur conformité avec les normes de droits de l'homme, d'autres formes d'analyse de l'impact, comme les AIPD, sont beaucoup plus étroites et se concentrent sur une seule activité de traitement des données.

---

255 Par exemple, concernant l'utilisation de systèmes de prise de décision algorithmique par le secteur public, voir AI Now Institute : 2018. Ce rapport dresse un cadre destiné aux entités du secteur public des États-Unis pour qu'elles mènent des « études d'impact algorithmiques » avant d'acquérir ou de déployer un système de décision automatique. Concernant l'évaluation des risques dans le domaine pénal, voir Selbst 2018 et Oswald et al. 2018. Concernant les risques pour les droits de l'homme et les registres de noms de domaines internet, voir ARTICLE 19 : 2017.

256 L'article 35 du Règlement général de l'UE sur la protection des données (RGPD) requiert une analyse de l'impact sur la protection des données personnelles lorsque le traitement de ces données est susceptible d'engendrer un « risque élevé » pour les droits et libertés des personnes physiques.

257 Plusieurs modèles d'étude d'impact sur les droits de l'homme peuvent être vus comme des formes spécifiques de la « diligence raisonnable en matière de droits de l'homme », issue des Principes directeurs des Nations Unies relatifs aux entreprises et aux droits de l'homme, pour lesquels la « diligence raisonnable » constitue une première étape indispensable pour identifier, atténuer et réparer les impacts négatifs de l'IA sur les droits de l'homme (Rasso et al. 2018 : 53). Voir aussi Toronto Declaration on Machine Learning 2018.

258 Mantelero 2018.

259 Raso et al. 2018 : 53.

260 Mantelero 2018.

261 Voir plus haut, note n° 241 ; Mantelero 2018 ; Edwards et Veale 2017.

Les techniques d'analyse d'impact peuvent jouer un rôle précieux en attirant l'attention sur les divers risques d'ingérence dans les droits de l'homme associés à une activité proposée, qui pourraient autrement être ignorés ou minimisés. Cependant, pour que les approches par l'analyse d'impact offrent une protection réelle et substantielle, il faudra développer une approche méthodologique claire et rigoureuse, susceptible d'être adoptée par des entreprises et organisations déterminées à identifier les risques pour les droits de l'homme et non considérée comme un travail bureaucratique tenant du rituel, permettant d'afficher une conformité de façade sans aucune préoccupation véritable pour le respect des droits de l'homme[262].

(b) Audit des algorithmes : contrairement aux analyses d'impact, qui ont lieu *avant* la mise en œuvre du système, les techniques d'audit des algorithmes visent à tester et à évaluer les systèmes algorithmiques en fonctionnement. L'audit des algorithmes est un domaine de recherche appliquée émergent, fondé sur une série de nouveaux outils et techniques de recherche destinés à détecter, à étudier et à diagnostiquer les effets négatifs indésirables des systèmes algorithmiques[263]. Il a été proposé de formaliser et d'institutionnaliser les techniques de ce type au sein d'un cadre de gouvernance prévu par la législation, prévoyant que les systèmes algorithmiques (ou du moins ceux jugés « à haut risque » en raison de l'ampleur et de la gravité des conséquences en cas de panne ou d'effets imprévus) soient régulièrement contrôlés par une autorité extérieure dotée d'un personnel technique dûment qualifié. Par exemple, Cukier et Mayer-Schonenberg estiment qu'il faudrait créer pour assumer ce rôle une nouvelle profession (les « algorithmistes »), proche des métiers du droit, de la médecine, de la comptabilité et de l'ingénierie ; les algorithmistes pourraient se charger d'auditer les algorithmes soit en tant que prestataires extérieurs indépendants, soit « en interne », en tant que salariés chargés de surveiller les algorithmes développés et déployés par leur organisation, qui pourraient ensuite être soumis à un contrôle extérieur[264].

3.7.3 Définition, suivi et application de normes

Les techniques et approches décrites ci-dessus sont très prometteuses pour l'objectif de situer les responsabilités prospectives et rétrospectives à l'égard des systèmes fondés sur les technologies numériques avancées. Mais pour exploiter tout leur potentiel, nous devons aussi tenir compte des cadres de gouvernance juridique et institutionnelle dans lesquels elles s'inscrivent. Par exemple, les divers courants de recherches techniques mentionnés au chapitre 3.7.2 pourraient énormément faciliter l'attribution des responsabilités prospectives entourant les systèmes numériques ; en effet, ils montrent que les milieux techniques reconnaissent – ce qui est tout à fait bienvenu – que ces systèmes ne sont pas « neutres », mais imprégnés de valeurs et susceptibles de fonctionner de façon contraire aux droits de l'homme. Il est important non seulement d'encourager ces travaux, mais aussi de leur donner un caractère interdisciplinaire, associant les milieux techniques et ceux du droit, de la philosophie, des lettres et des sciences sociales, afin d'approfondir la question de la concrétisation des valeurs des droits de l'homme en dispositifs techniques de protection et de

---

262 Power 1997.

263 Desai et Kroll 2017 ; Sandvig et al. 2014. Voir par exemple les ressources citées dans *Auditing Algorithms*.

264 Cette approche ressemblerait à celle du contrôle des comptes classique : la comptabilité des entreprises est soumise à la fois à des experts-comptables internes à l'entreprise et à des commissaires aux comptes, en externe, qui vérifient les comptes conformément à la loi et certifient qu'ils sont sincères et véritables (Mayer-Schonberger et Cukier 2013 : 180). Voir aussi Crawford et Schultz 2014 ; Citron 2008.

résoudre les conflits de valeurs à travers une approche fondée sur les droits de l'homme. Tout aussi important, nous devons définir le *statut juridique* de ces techniques. Face aux problèmes éthiques, le secteur du numérique tend à proposer des solutions techniques ; mais faire aveuglément confiance à ces solutions reviendrait à valider une nouvelle forme d'« éthique-washing[265] ». En d'autres termes, si ces approches techniques ne se fondent pas sur le droit, et si leur validité et leur fonctionnement ne font pas l'objet d'une évaluation et d'une supervision transparentes de la part d'une autorité indépendante compétente, elles risquent d'échouer à protéger réellement les droits de l'homme. Comme le soulignent les spécialistes des politiques de régulation, il est crucial que les trois éléments du processus soient présents : définition de normes, collecte d'informations et suivi des activités requises pour respecter les normes, et mesures de mise en œuvre avec application de sanctions en cas de non-conformité[266]. Pour être effective et légitime, la gouvernance de la régulation doit à la fois associer les acteurs concernés à la définition des normes et reposer sur une autorité indépendante, dotée de ressources suffisantes et compétente pour collecter des informations, enquêter sur les cas de non-conformité et sanctionner les violations[267]. Pour être sûrs que les dispositifs techniques de protection visant à assurer le respect des droits de l'homme sont bien appliqués aux processus numériques, nous devons nous doter de solides mécanismes de surveillance capables d'enquêter sur leur fonctionnement. C'est pourquoi les normes techniques elles-mêmes devraient être élaborées en toute indépendance (et dans l'idéal, à travers un processus participatif associant tous les acteurs concernés) et soumises à un contrôle externe, et la conformité à ces normes peut et doit être vérifiée par une instance extérieure compétente pour imposer des sanctions (ou pour demander que des sanctions soient appliquées). Sans véritable contrôle indépendant, de tels dispositifs n'offriront pas de base suffisante pour assurer le respect des droits de l'homme. De plus en plus, les pouvoirs publics nationaux et locaux admettent la nécessité d'une prise en compte et d'une évaluation plus formelle, institutionnalisée et systématique des systèmes algorithmiques. L'existence de groupes de travail et organismes publics chargés d'examiner et/ou de superviser des systèmes sociotechniques fondés sur les données en témoigne[268].

3.8 Renouveler le discours sur les droits de l'homme à l'heure des réseaux numériques

Avec le début dans le monde d'une nouvelle ère, celle des réseaux numériques, la protection des droits de l'homme et des valeurs qui les sous-tendent acquiert une importance encore plus cruciale. La conception actuelle des droits de l'homme, et les mécanismes destinés à les mettre en œuvre, sont-ils adaptés à ce nouveau paysage sociotechnique ? Les puissantes technologies numériques en réseau apparues ces dernières années ont rendu possibles des pratiques et des actions qui ne l'étaient pas auparavant, et donc engendré des risques et des formes de préjudices nouveaux, suscitant une réflexion sur la nécessité de nouveaux droits de l'homme et de nouveaux régimes de gouvernance institutionnelle pour répondre à ces risques dans la pratique[269]. Bien que la structure et le cadre institutionnel de base en matière de protection des droits de l'homme, bien établis et universellement reconnus, soient susceptibles de répondre effectivement à une bonne part des défis associés à la montée de l'automatisation numérique et de l'intelligence artificielle, un renouvellement du discours actuel sur les droits et des mécanismes de mise en œuvre pourrait s'imposer, et ce pour

265 Greene et al. 2019.

266 Morgan et Yeung 2007 ; Lodge et Wegrich 2014.

267 Nemitz 2018.

268 Pour une synthèse des initiatives nationales dans toute l'Europe, voir Access Now 2018.

269 Brownsword, Scotford et Yeung 2017.

plusieurs raisons. Premièrement, beaucoup des droits des personnes concernées sont difficiles à faire valoir dans la pratique, notamment du fait de l'opacité de beaucoup des systèmes sociotechniques dans lesquels les technologies numériques s'inscrivent. Deuxièmement, la teneur des droits existants et leur champ d'application ont été définis avant l'ère des réseaux. De ce fait, les droits en question risquent de ne pas protéger complètement contre les menaces que ces technologies font peser sur les individus, en particulier les efforts illégitimes pour tromper et manipuler les individus rendus possibles par les « technologies persuasives », ou encore les problèmes de discrimination (voir plus haut). Par exemple, bien que le droit à la protection des données permette aux intéressés d'insister pour obtenir une intervention humaine, de donner leur avis ou de contester une décision entièrement automatique ayant sur eux un impact significatif, aucun de ces droits ne s'applique en cas de décision *partiellement* automatique. Ils ne garantissent pas non plus, en pratique, qu'une personne affectée puisse facilement comprendre qu'elle a subi une inégalité de traitement et, si oui, savoir s'il s'agissait d'une discrimination, par conséquent illégale. Troisièmement, et peut-être surtout dans le contexte des nouvelles technologies numériques, les intéressés sont libres de *renoncer* à certains de leurs droits et à la protection qui les accompagne en acceptant des pratiques qui, autrement, constitueraient une atteinte à ces droits[270]. Par exemple, s'il ne fallait compter que sur l'article 8 pour protéger les droits et les intérêts liés à la prestation de services fondés sur les données, et dans un contexte du « tout gratuit », il existerait un fort risque que les intéressés renoncent trop vite à leurs droits, c'est-à-dire consentent à livrer leurs données personnelles en échange d'un accès « gratuit » aux services numériques et aux facilités qu'ils offrent[271]. En revanche, les principes de protection des données au cœur des régimes européens contemporains dans ce domaine, dont la Convention 108 modernisée (principes reflétés par la jurisprudence de la Cour européenne des droits de l'homme sur l'article 8) comprennent des obligations imposées à ceux qui traitent les données et auxquelles les détenteurs des droits ne peuvent renoncer : caractère légal du traitement, exposé de son objectif, minimisation des données, etc., offrant une protection plus constante des valeurs et des intérêts collectifs que ces régimes cherchent à protéger.

Au-delà de ces faiblesses potentielles cependant, les conceptions contemporaines des droits de l'homme et leurs mécanismes de mise en œuvre, parce qu'ils sont axés sur l'individu, peuvent passer à côté des menaces que ces technologies font peser sur des biens communs, dont notamment le besoin de préserver et d'alimenter les bases sociotechniques qui rendent possibles les choix moraux et l'exercice des droits de l'homme. Mireille Hildebrandt, grande philosophe du droit et des technologies, parle des conditions techniques censées être réunies pour que le droit (et la notion actuelle d'État de droit) remplisse ses fonctions[272]. Au sein

---

270 La mesure dans laquelle la Cour européenne des droits de l'homme est prête à admettre la possibilité, pour un individu, de renoncer à des droits protégés par la CEDH, et les conditions qui doivent entourer une telle renonciation dépendent du droit en question et du contexte précis dans lequel l'intéressé a renoncé à son droit. Par exemple, dans *Scoppola c. Italie* (n° 2), 17 septembre 2009, n° 10249/03, par. 135, la Cour affirme : « Ni la lettre ni l'esprit de cette disposition n'empêchent une personne d'y renoncer de son plein gré de manière expresse ou tacite. Cependant, pour entrer en ligne de compte sous l'angle de la Convention, ladite renonciation doit se trouver établie de manière non équivoque et s'entourer d'un minimum de garanties correspondant à son importance […]. De plus, elle ne doit se heurter à aucun intérêt public important […] ». S'agissant du droit privé et des relations contractuelles entre acteurs non étatiques, la Cour est susceptible d'aborder la question des renonciations sous l'angle de l'obligation positive qu'ont les États de prendre des mesures raisonnables pour protéger les individus des violations de la Convention par d'autres acteurs privés, dont l'obligation de garantir (par la loi ou par d'autres mesures) un exercice « concret et effectif » des droits qui y sont affirmés.

271 Solove 2012.

272 Hildebrandt 2015.

d'environnements « intelligents », qui prélèvent sans cesse des données numériques sur le monde matériel afin d'en déduire, pour les prédire et les anticiper, les comportements à venir des choses, des personnes et des systèmes, ces conditions techniques sont à la fois supplantées et augmentées, altérant la possibilité même de choisir, de penser et de raisonner au sens où nous l'entendons aujourd'hui – parce que les technologies intelligentes fonctionnent de manière continue et immanente, et parce qu'elles sont conçues pour apprendre et produisent des résultats que leurs concepteurs n'avaient pas prévus[273].

La juriste nord-américaine Julie Cohen développe les réflexions d'Hildebrandt en se fondant, à la fois, sur l'étude du droit et sur le corpus grandissant de travaux en sociologie des sciences regroupés sous l'appellation STS (*Science and Technology Studies*[274]). Selon Cohen, pour que les droits de l'homme s'exercent à l'heure des environnements intelligents, il faut « prendre les invites au sérieux », sous peine de vider nos droits de leur substance. D'après la théorie des invites, la conception même des objets et des environnements technologiques conditionne et restreint nos possibilités, notamment à travers la série d'actions et de réactions qu'elle « invite » de la part de l'utilisateur. Admettre que les technologies numériques intelligentes s'interposent continuellement, et de par leur nature même, entre nous-mêmes et nos choix, c'est reconnaître comme incomplet le discours juridique actuel sur les droits de l'homme (dont le respect de la vie privée). Cohen affirme donc, en termes très convaincants, qu'il ne suffit pas d'« élargir » le discours sur les droits. Il faut à la fois *repenser les droits* et les *exprimer en d'autres termes*, en reconnaissant que les configurations sociotechniques ont le pouvoir de faciliter ou d'entraver les libertés et les possibilités dont nous jouissons dans les faits[275]. En particulier, notre discours sur les droits part de présupposés rarement contestés sur les qualités inhérentes à l'environnement, avec ses limites (par exemple, il serait matériellement impossible de surveiller toute une population) ou son absence de limites (par exemple, il existerait toutes sortes de lieux où se réunir dans différents buts, dont les manifestations démocratiques). Pourtant, les avancées des technologies numériques en réseau remettent en question ces présupposés, et nous sommes tout juste en passe de comprendre que les contraintes et invites pertinentes ne concernent pas que le monde physique, mais aussi les flux de données et d'informations, qui se répercutent directement sur nos droits et nos libertés. Nous devons donc revoir notre discours sur les droits pour y englober notre architecture sociotechnique et ses invites, avec leurs incidences pratiques, afin que ce discours « s'applique avec force aux nouveaux types de considérations matérielles et opérationnelles[276] ».

En d'autres termes, l'incapacité des droits à apporter une réponse complète aux menaces créées par les technologies d'IA tient surtout aux limites inhérentes aux approches fondées sur les droits, axées sur les individus détenteurs de droits plutôt que sur les dommages structurels avant tout ressentis à un niveau collectif et sociétal. Par exemple, la création d'un nouveau « droit aux rapports humains authentiques[277] », bien que tentante, ne suffirait peut-être pas à lever les peurs de déshumanisation structurelle et sociétale engendrées par la place croissante des technologies informatiques dans nos vies. Ce sont les effets cumulés de ces technologies

273 Cohen 2017, 3, citant Hildebrandt 2015, 88-102.

274 Cohen 2017.

275 Cohen 2017 : 7.

276 Cohen 2017 : 9.

277 Abordé ci-dessus, chapitre 2.2.2.

au fil du temps, et leur ampleur, qui risquent de saper les bases sociotechniques que la notion même de droits de l'homme présuppose et dans lesquelles ils sont ancrés[278].

Parce que les technologies numériques intelligentes sont « radicalement différentes par nature » des autres types de technologies, le défi pour la société consiste à s'adapter à leur différence, ainsi qu'à leur puissance[279]. Insister sur les incidences structurelles de ces technologies suscite une manière de voir que Cohen qualifie d'« intrinsèquement collective ». Cette perspective met en lumière la responsabilité des États, ainsi que notre responsabilité collective en tant que communauté morale, dans l'entretien des bases sociotechniques de la liberté morale et démocratique, ainsi que les effets cumulés que pourraient avoir les phénomènes sociaux inquiétants cités dans notre étude, effets susceptibles de saper les « biens communs moraux et démocratiques[280] » sans lesquels les droits de l'homme et les libertés fondamentales ne peuvent s'exercer en pratique. Ces bases sociales doivent, au minimum, assurer les conditions nécessaires à l'existence de la responsabilité et de l'appréciation morale ; car sans elles, il n'y a pas de liberté et les droits de l'homme perdent tout leur sens[281]. Pourtant, nous ne possédons pas d'institution chargée de veiller à la santé des bases sociotechniques dans lesquelles sont ancrés les droits de l'homme et les libertés démocratiques. Il nous faut donc, peut-être, développer à la fois un nouveau « vocabulaire » des droits et des mécanismes institutionnels destinés à assurer la santé et la pérennité de ces bases, afin d'assurer une véritable protection des droits de l'homme à l'ère du numérique et de l'hyperconnexion[282].

3.9 Résumé

Ce chapitre a souligné toute l'importance de veiller à ce que les responsabilités puissent être attribuées, de manière prospective comme rétrospective, à l'égard des conséquences négatives réelles et potentielles associées au développement et au fonctionnement des technologies numériques avancées. L'attribution juste et effective des responsabilités à l'égard de ces risques et de ces impacts négatifs est vitale, non seulement pour protéger les droits de l'homme et préserver le bien-être des individus, des groupes et de toute la société, mais aussi – plus fondamental encore – pour veiller à ce que notre société reste une communauté morale. Or, trouver les responsables des risques et des effets négatifs associés aux technologies numériques, de plus en plus puissantes et sophistiquées, soulève des défis considérables, parce que de très nombreux individus et organisations participent à leur développement et à leur mise en œuvre et parce que leur fonctionnement peut prendre des tournures inattendues.

La prolifération des codes d'éthique adoptés par les acteurs de l'industrie du numérique, qui s'engagent publiquement à les respecter, marque une reconnaissance bienvenue de la nécessité de prendre au sérieux les responsabilités associées aux risques et aux autres effets négatifs des technologies numériques avancées. Cependant, ces initiatives d'autorégulation ne s'appuient sur aucun mécanisme institutionnel destiné à assurer une véritable participation du public à la définition des normes pertinentes, et ne s'accompagnent d'aucun mécanisme

278 Yeung 2011.
279 Hildebrandt 2015 ; Cohen 2017.
280 Yeung 2011 ; Yeung 2017b.
281 Brownsword 2005.
282 Yeung 2011.

externe de mise en œuvre et de sanction ; elles n'offrent donc pas des garanties légitimes et efficaces.

Bien que l'aptitude des systèmes numériques avancés à fonctionner de manière plus ou moins autonome ait été présentée comme exonérant leurs développeurs de toute responsabilité, cette affirmation repose sur une vision de la responsabilité morale très spécifique et très étroite. Nous avons vu que plusieurs modèles de responsabilité pouvaient convenir pour situer les responsabilités en cas d'impacts négatifs des systèmes d'IA, et noté que concernant les atteintes aux droits de l'homme, la responsabilité était absolue, sans nécessité de prouver une faute. Les États, auxquels il appartient en premier lieu d'assurer une protection effective des droits de l'homme, sont juridiquement tenus d'adopter des cadres législatifs nationaux imposant des devoirs aux acteurs non étatiques. En outre, il est de plus en plus admis que la nature absolument fondamentale des droits de l'homme crée des effets horizontaux sur les acteurs non étatiques, développeurs compris[283]. Bien que les recours en justice offrent aux personnes affectées par des technologies d'IA un important moyen de demander réparation, nous avons aussi identifié une série d'autres instruments de gouvernance (dont des dispositifs techniques de protection) pouvant servir à assurer une véritable responsabilisation, et qui méritent notre intérêt.

Pourtant, bien que plusieurs mécanismes de gouvernance (décrits plus haut) puissent aider, s'ils s'appuient sur le droit, à protéger effectivement les droits de l'homme, il est peu probable qu'ils fournissent à eux seuls une protection pertinente et complète. En particulier, les technologies numériques avancées sont aujourd'hui d'une puissance et d'une sophistication telles qu'elles peuvent être comprises comme « radicalement différentes par nature » des autres types de technologies, en particulier vu leurs implications profondes pour l'architecture collective et partagée de nos sociétés sur les plans technique, social, démocratique et moral. Nous devons donc renouveler le discours existant sur les droits de l'homme et les instruments pertinents pour affirmer notre responsabilité collective dans l'entretien des bases sociotechniques de la liberté morale et démocratique et pour tenir compte des effets cumulés que pourraient avoir les phénomènes sociaux inquiétants cités dans notre étude, effets susceptibles de saper les « biens communs moraux et démocratiques » sans lesquels les droits de l'homme et les libertés fondamentales ne peuvent s'exercer dans la pratique.

[283] Pour le secteur privé, on doit l'approfondissement le plus complet de cet aspect à M. Ruggie, Rapporteur spécial des Nations Unies, qui a « codifié » l'obligation, découlant de la responsabilité sociale des entreprises, de respecter les droits de l'homme et d'agir en conséquence même dans les pays où la législation nationale ne l'exige pas.

Chapitre 4. Conclusion

On peut prévoir que les techniques dites d'« intelligence artificielle » vont continuer à progresser et gagner en puissance et en sophistication. Les réussites relativement récentes de l'IA, associées à l'apparition d'infrastructures de données mondiales et interconnectées, ont permis la prolifération de services et de systèmes numériques. Ces derniers sont déjà sources d'avantages considérables, en particulier parce qu'ils offrent plus de commodité et d'efficacité dans un très large éventail de domaines et d'activités, bien que l'accès à ces services reste largement l'apanage des habitants des pays riches et industrialisés. Ils sont porteurs de promesses extraordinaires et pourraient améliorer substantiellement notre bien-être individuel et collectif, y compris en renforçant notre capacité à exercer nos droits et nos libertés. Cependant, le public s'inquiète de plus en plus, et à juste titre, de leurs conséquences négatives pour la société, dont le risque qu'ils sapent la protection des droits de l'homme, ce qui pourrait déstabiliser – comme notre étude l'a souligné – les bases mêmes de notre capacité à agir en êtres moraux. C'est pourquoi nous avons cherché à examiner les incidences des technologies numériques avancées (dont l'IA) sur la notion de responsabilité sous l'angle des droits de l'homme. Notre étude a identifié au sein de ces technologies une série de propriétés « pertinentes pour la responsabilité », mis en avant plusieurs des impacts négatifs qu'elles pourraient avoir et cherché à savoir comment les responsabilités en matière de prévention, de gestion et d'atténuation de ces impacts (dont le risque d'atteintes aux droits de l'homme) pourraient être attribuées et réparties.

Cette étude a montré que les réponses aux risques potentiels et avérés, aux dommages et aux atteintes aux droits associés aux technologies numériques avancées seraient probablement d'autant plus efficaces et légitimes qu'elles se concentreraient sur les conséquences pour les individus et la société, en veillant à attribuer équitablement les *responsabilités prospectives*, afin de prévenir et d'atténuer les risques associés à ces technologies, et les *responsabilités rétrospectives*, au cas où ces risques dégénéreraient en dommages et/ou en atteintes aux droits. C'est à cette condition que nous aurons l'assurance, d'une part que des efforts soutenus et systématiques sont engagés pour éviter la survenue de torts et de préjudices, et d'autre part qu'il est mis un terme aux activités préjudiciables et que des mécanismes institutionnels effectifs et légitimes existent pour assurer une réparation et une prévention appropriées le cas échéant. Il faudra pour cela s'intéresser à la fois à ceux qui développent, déploient et mettent en œuvre ces technologies, aux utilisateurs individuels et aux intérêts collectifs affectés par ces technologies, et au rôle des États dans la mise en place des conditions nécessaires pour mettre les citoyens à l'abri des risques et dans la garantie d'une véritable protection des droits de l'homme.

Quatre constats issus de cette étude méritent d'être soulignés :

1. Il est particulièrement important de veiller à l'existence de mécanismes, effectifs et légitimes, capables de *prévenir et d'empêcher* les atteintes aux droits de l'homme, d'autant que beaucoup de violations des droits de l'homme associées aux technologies numériques avancées n'entraînent pas de dommage matériel. La vitesse et l'échelle auxquelles fonctionnent ces technologies rendent plus pressante encore la nécessité d'une approche préventive, tout comme le risque réel que de telles violations ne sapent les bases sociotechniques collectives indispensables à l'existence même de la liberté, de la démocratie et des droits de l'homme. Ces constats ont plusieurs implications. Premièrement, les États ont une responsabilité importante, celle de préserver l'environnement sociotechnique dans lequel les droits de l'homme sont ancrés. Deuxièmement, de plus solides mécanismes de réclamation collective pourraient être nécessaires pour lever les obstacles à l'action collective qui peuvent empêcher les individus de réagir aux violations de droits générées par des systèmes

d'IA. Troisièmement, à l'heure des réseaux et des données, il nous faut peut-être renouveler notre conception existante des droits de l'homme pour tenir compte de la reconfiguration de l'environnement sociotechnique que ces technologies peuvent entraîner et des menaces qu'elles font peser sur les valeurs et les biens communs.

2. Le modèle de responsabilité juridique s'appliquant aux *violations des droits de l'homme* est largement considéré comme un modèle de « responsabilité absolue », qui s'applique même sans qu'une faute n'ait été prouvée. En revanche, les obligations de réparation en cas de *dommage matériel* peuvent être attribuées juridiquement, selon divers modèles de responsabilité (brièvement présentés au chapitre 3.4). Lorsqu'un comportement nuit à autrui, la diversité des modèles juridiques qui peuvent servir à attribuer et à répartir les responsabilités montre clairement à quel point il serait erroné d'attendre d'un seul modèle de responsabilité qu'il s'applique à tous les types de conséquences négatives que peuvent entraîner les technologies numériques avancées. Les modèles juridiques de responsabilité insistent sur les relations entre les agents moraux, les patients moraux et la société dans son ensemble, contrairement à beaucoup d'analyses philosophiques de la responsabilité appliquées aux systèmes d'IA, qui se concentrent sur le comportement des agents moraux et sur les responsabilités qu'il entraîne à l'égard des patients moraux (les « victimes ») et de la société. Les différents modèles juridiques de responsabilité ménagent chacun un équilibre différent entre notre intérêt à pouvoir agir librement et notre intérêt, en tant que victimes, à préserver notre sécurité et celle de nos biens. Identifier lequel de ces modèles (s'il y en a un) convient le mieux à l'attribution des responsabilités pour les différents risques découlant des technologies numériques avancées n'a rien d'évident ; au contraire, cette répartition des risques relève d'un *choix politique engageant toute la société*. Dans des sociétés démocratiques, qui se sont engagées à respecter les droits de l'homme, l'État a une responsabilité cruciale : veiller à ce que ces choix soient opérés de manière transparente et démocratique et de manière à ce que les politiques adoptées à terme protègent effectivement les droits de l'homme.

3. Les divers courants de recherches techniques peuvent fortement contribuer à situer les responsabilités prospectives et rétrospectives à l'égard des technologies numériques avancées, via la mise au point de méthodes rendant possibles à la fois des dispositifs techniques de protection et un véritable « audit des algorithmes ». Il convient de poursuivre et de soutenir ces recherches, qui devraient être interdisciplinaires – technique, droit, philosophie, lettres et sciences sociales – en vue de mieux traduire la sauvegarde des droits de l'homme en dispositifs techniques de protection et de comprendre en quoi une approche par les droits de l'homme peut résoudre les problèmes de conflits de valeurs.

4. Une véritable protection des droits de l'homme à notre époque d'hyperconnexion numérique exige que nous nous dotions de mécanismes, instruments et institutions de gouvernance efficaces et légitimes pour surveiller et superviser le développement, la mise en œuvre et le fonctionnement de nos systèmes sociotechniques. On trouvera en annexe A quelques suggestions sur les moyens d'instituer des mécanismes de gouvernance capables d'assumer ce rôle. Les entreprises de technologies ont adopté volontairement des « codes d'éthique » qu'elles s'engagent publiquement à respecter, reconnaissant ainsi, ce qu'il faut saluer, que les technologies qu'elles développent peuvent avoir des effets néfastes et qu'elles portent une part de responsabilité. Ces codes, cependant, ne suffiront pas à protéger les droits de l'homme. Un développement et une mise en œuvre responsables de l'IA supposent, au minimum, que les normes pertinentes soient définies à travers une démocratie participative et

> que des autorités indépendantes, dotées des ressources et des compétences nécessaires, rassemblent méthodiquement les informations, enquêtent sur les cas de non-conformité et sanctionnent les violations. En particulier, pour être sûrs que les dispositifs techniques de protection visant à assurer le respect des droits de l'homme sont bien appliqués aux processus numériques, nous devons nous doter de mécanismes de surveillance externes solides et indépendants capables d'enquêter sur leur fonctionnement, faute de quoi de tels dispositifs ont peu de chances d'assurer une véritable responsabilisation à l'égard de l'IA. Il appartient aux États de veiller à ce que de tels mécanismes de gouvernance soient créés et mis en œuvre de manière à garantir la protection des droits de l'homme.

À l'heure de l'hyperconnexion numérique, si nous voulons vraiment protéger et promouvoir les droits de l'homme, nous ne pouvons tolérer que les technologies et systèmes numériques avancés et ceux qui les développent et les appliquent exercent de manière irresponsable un pouvoir de plus en plus grand. Un principe fondamental s'applique ici, celui de la réciprocité : ceux qui offrent des services (dont ils tirent des bénéfices) en profitant des avantages de ces technologies numériques avancées, y compris l'IA, doivent assumer leurs responsabilités en cas de conséquences négatives. Il est par conséquent crucial que les États, tenus de protéger les droits de l'homme, s'engagent aussi à faire en sorte que ceux qui exercent le pouvoir numérique (dont le pouvoir tiré de l'accumulation massive de données) aient à répondre des conséquences de leurs actes. L'engagement des États à protéger les droits de l'homme les oblige aussi à faire en sorte qu'il existe, en droit national, des structures de gouvernance assurant dûment l'attribution de la responsabilité prospective et rétrospective à l'égard des risques, des dommages et des atteintes aux droits engendrés par les technologies numériques avancées.

Annexe A

Cette annexe présente une série de mesures et de mécanismes institutionnels qui mériteraient une étude plus approfondie, car ils pourraient contribuer à la protection des droits de l'homme à notre époque de technologies numériques avancées et en réseau. Ces mesures ne se veulent pas des recommandations, mais des pistes de réflexion invitant à approfondir le débat.

Responsabilité prospective

Envisager des financements supplémentaires pour soutenir et encourager des recherches interdisciplinaires visant à développer des techniques, des dispositifs et des normes pouvant contribuer à une juste attribution des responsabilités prospectives, en vue de prévenir et d'atténuer les risques de dommages ou de torts entraînés par les technologies numériques avancées.

Envisager des mesures destinées à encourager les États à œuvrer, y compris en coopérant entre eux, au développement de mécanismes institutionnels de gouvernance fondés sur le droit destinés à faciliter la protection des droits de l'homme contre les risques potentiels et avérés que soulèvent les technologies numériques avancées. Ces mesures peuvent être les suivantes :

a. obligation légale d'entreprendre une « analyse de l'impact sur les droits de l'homme » (intégrant une analyse de l'impact des algorithmes) avant de déployer des technologies numériques avancées, comprenant une déclaration accessible au public expliquant comment l'architecture et le fonctionnement du système résolvent les conflits de valeurs et les ingérences potentielles dans les droits de l'homme ;

b. développer, en associant un large éventail d'acteurs, un code de bonnes pratiques sur la préparation des analyses de l'impact des technologies numériques avancées sur les droits de l'homme ;

c. clarifier l'étendue et la teneur des obligations légales de tous ceux qui participent au développement de services numériques (dont les développeurs de logiciels), en particulier les obligations ayant une incidence directe sur la protection des droits de l'homme ;

d. étudier la nécessité de soumettre les développeurs et les prestataires à l'obligation légale de mener des tests et des vérifications, à la pertinence démontrable, des systèmes informatiques complexes pouvant avoir un impact direct et substantiel sur les droits de l'homme, à la fois avant le déploiement de ces systèmes et à intervalles réguliers après leur mise en œuvre en milieu réel ;

e. encourager l'usage de dispositifs techniques de protection (« droits de l'homme dès la conception », techniques d'exploitation des données soucieuses d'équité ou IA explicable, par exemple), en identifiant en quoi ils peuvent contribuer au respect des droits de l'homme. Étudier la nécessité d'asseoir ces techniques sur le droit, y compris en les soumettant à un contrôle extérieur pour s'assurer qu'elles sont bien appliquées dans le respect des droits de l'homme ;

f. encourager des recherches supplémentaires sur le développement de techniques et de normes en faveur d'innovations responsables et respectant les droits de l'homme

dans l'industrie du numérique (modélisation, provenance et qualité des données, audit des algorithmes, validation, vérification et tests, etc.) ;

g. envisager la mise en place d'un système d'accréditation professionnelle pour des techniciens qualifiés et formés à l'audit des algorithmes, constituant une catégorie de professionnels chargés de vérifier et de certifier la conception et le fonctionnement des algorithmes et soumis à une obligation fiduciaire de loyauté et de bonne foi ;

h. élaborer un cadre méthodologique et une série de mesures visant à identifier et évaluer systématiquement l'ampleur et la gravité des risques pour les droits individuels (y compris pour les bases sociotechniques dans lesquelles sont ancrés les droits de l'homme et les libertés fondamentales) soulevés par les applications de l'IA proposées ou potentielles ;

i. déterminer si les applications de l'IA qui font peser des menaces graves et disproportionnées sur les droits de l'homme devraient être interdites, à moins de faire l'objet d'une consultation publique préalable et d'être approuvées par une autorité de contrôle indépendante et à la composition adéquate. Un tel cadre pourrait comprendre une catégorie d'applications de l'IA interdites d'emblée car elles feraient peser des menaces d'une gravité inacceptable et potentiellement catastrophique sur les droits de l'homme et les libertés fondamentales[284].

Responsabilité rétrospective

Envisager de soutenir l'élaboration d'orientations et de techniques pouvant aider à dûment déterminer les responsabilités rétrospectives en cas de dommages ou d'atteintes aux droits, individuels ou collectifs, entraînés par des technologies numériques avancées. On pourrait pour cela encourager les États, y compris en coopérant entre eux, à mettre en place des mécanismes de gouvernance institutionnelle fondés sur le droit. Les mesures prises pourraient être les suivantes :

a. passer en revue le système juridique national pour vérifier s'il peut déterminer les responsabilités en cas de dommage causé par des technologies numériques avancées, et repérer les lacunes potentielles appelant éventuellement une réforme législative (action à mener par chaque État membre) ;

b. étudier la nécessité d'élaborer des instruments normatifs pour, en cas de préjudice, clarifier et répartir par défaut les responsabilités entre ceux qui participent à la conception, au développement, au déploiement et à l'offre de systèmes numériques, ainsi que les propriétaires de ces systèmes. Cela pourrait englober l'obligation légale de dédommager les personnes affectées ou lésées par le fonctionnement de ces services, y compris en les indemnisant et en adoptant des mesures pour éviter que les mêmes phénomènes ne se reproduisent. Au moment d'élaborer un tel instrument, on pourrait étudier l'opportunité d'un critère de « diligence raisonnable », dans certaines circonstances clairement et étroitement définies, conduisant à réduire l'étendue de la responsabilité juridique du développeur en cas de tort ou de dommage ;

[284] Voir aussi Groupe d'experts de haut niveau de l'UE sur l'intelligence artificielle (2019a)

c. soutenir des recherches supplémentaires sur la définition et la répartition de l'autorité entre les êtres humains placés dans la boucle de systèmes informatiques complexes, à la lumière du problème reconnu du « parti pris en faveur de l'automatisation » et de la tendance à faire peser la responsabilité sur les individus placés dans la boucle plutôt que sur ceux qui développent et mettent en œuvre le système sociotechnique en question ;

d. étudier l'opportunité d'imposer au secteur du numérique un régime d'assurance obligatoire et celle de créer un régime d'assurance national, financé par ce secteur, pour veiller à ce que les victimes ne restent pas sans indemnisation ;

e. soutenir un renforcement des capacités en vue de créer de nouvelles institutions de gouvernance (et d'étendre les compétences des institutions existantes) pouvant mener des enquêtes rigoureuses sur les responsabilités prospectives et rétrospectives des développeurs et des prestataires de services numériques et faire appliquer ces responsabilités ;

f. étudier l'opportunité d'instaurer des mécanismes de réclamation collective et d'assouplir les règles existantes afin de lever les obstacles à l'action collective : de nombreux individus sont exposés à des atteintes aux droits mais risquent de ne pas avoir la motivation nécessaire pour réagir, même si les effets cumulés de ces atteintes peuvent s'avérer très lourds. À cette fin, chercher à savoir si la procédure de réclamations collectives adoptée pour rendre plus rapide et efficace la mise en œuvre de la Charte sociale européenne constitue un modèle approprié ;

g. réexaminer les ressources et les pouvoirs d'enquête, de sanction et de réparation des autorités répressives. Cela pourrait comprendre le développement et la consolidation de connaissances et de compétences techniques en apprentissage automatique et autres techniques de développement et d'évaluation des logiciels dans le secteur public.

<u>Renouveler le discours sur les droits de l'homme à l'heure des réseaux numériques</u>

Chercher à savoir comment la protection des droits de l'homme et le discours à leur sujet tels qu'ils existent aujourd'hui devraient évoluer pour assurer une protection effective des droits de l'homme à notre époque d'interconnexion mondiale, en reconnaissant la nécessité de préserver les bases sociotechniques sur lesquelles se fondent l'État de droit et notre communauté morale. Ces mesures peuvent être les suivantes :

a. étudier l'opportunité d'une nouvelle Convention sur les droits de l'homme à l'époque des réseaux numériques qui reconnaîtrait, au minimum, que les responsabilités prospectives et rétrospectives en cas de risques, de dommages et d'atteintes aux droits doivent être dûment attribuées et réparties ;

b. envisager la nécessité, dans le cadre d'une telle Convention (ou d'un autre instrument multilatéral), de reconnaître formellement le rôle de sauvegarde joué par les mécanismes institutionnels indépendants contre les risques collectifs que ces technologies font peser sur les bases des ordres démocratiques, dans lesquels les droits de l'homme sont ancrés ;

c. chercher à savoir si de nouveaux mécanismes collectifs de suivi et de prise de décision seraient nécessaires ou souhaitables afin de surveiller et d'évaluer les effets cumulés de ces technologies sur les droits de l'homme dans les États membres. À cette fin,

étudier la nécessité ou l'opportunité d'établir un « observatoire mondial » chargé d'assurer en permanence ce rôle de suivi et de signalement ;

d. appliquer le principe de précaution lorsque l'interaction de systèmes algorithmiques est susceptible d'avoir des conséquences catastrophiques qu'aucun prestataire de services numériques ne peut raisonnablement prévoir ; envisager d'interdire les catégories d'applications algorithmiques susceptibles d'avoir des effets catastrophiques ; étudier la nécessité de structures de suivi systématique et d'institutions spécialisées visant à éviter que de telles applications ne soient développées et déployées.

Références

6, P. (2001) 'Ethics, regulation and the new artificial intelligence, part I: accountability and power'. *Information, Communication & Society* 4(2):199-229.

6, P. (2001) 'Ethics, regulation and the new artificial intelligence, part II: autonomy and liability'. *Information, Communication & Society* 4(3): 406-434.

6, P. (2002) 'Who wants privacy protection, and what do they want?' *Journal of Consumer Behaviour: An International Research Review*. 2(1): 80-100.

Access Now (2018) *Mapping Regulatory Proposals for Artificial Intelligence in Europe*. Disponible à l'adresse suivante : https://www.accessnow.org/mapping-artificial-intelligence-strategies-in-europe/ (consulté le 7.11.18).

AI Now (2017) *AI Now 2017 Report*. Disponible à l'adresse suivante : https://ainowinstitute.org/AI_Now_2017_Report.pdf (consulté le 31.10.2018).

Akansu, A. N. (2017). 'The flash crash: a review.' *Journal of Capital Markets Studies* 1(1): 89-100.

Amnesty International (2017) *Artificial Intelligence for Good*. Disponible à l'adresse suivante : https://www.amnesty.org/en/latest/news/2017/06/artificial-intelligence-for-good/ (consulté le 2.11.2018).

Andrade, F., Novais, P., Machado, J. and Neves, J. (2007) 'Contracting agents: legal personality and representation' *Artificial Intelligence and Law*. 15(4): 357-373.

Angwin, J., Larson, J., Mattu, S. and Kirchner, L. (2016) 'Machine bias'. *ProPublica,* 23 May. Disponible à l'adresse suivante : https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (consulté le 5.11.2018).

ARTICLE 19, The Danish Institute for Human Rights and the Dutch Internet Domain-registry (2017) *Sample ccTLD Human Rights Impact Assessment Tool*. Disponible à l'adresse suivante : https://www.article19.org/wp-content/uploads/2017/12/Sample-ccTLD-HRIA-Dec-2017.pdf (consulté le 2.11.2018).

Asaro, P.M. (2014) 'A Body to Kick, but Still No Soul to Damn: Legal Perspectives on Robotics' in *Robot Ethics: The ethical and social implications of robotics.* Edited by P. Lin, K. Abney & G.A. Bekey. MIT Press.

Assemblée générale des Nations Unies (2018), Rapport du Rapporteur spécial sur la promotion et la protection du droit à la liberté d'opinion et d'expression. Soixante-treizième session. 29 août. A/73/348. Disponible sur https://undocs.org/fr/A/73/348 (consulté le 7.11.18)

Auditing Algorithms: Adding Accountability to Automated Authority. Disponible à l'adresse suivante : http://auditingalgorithms.science/ (consulté le 5.11.2018).

Australian Human Rights Commission (2018) *Human Rights and Technology Issues Paper.* July. Disponible à l'adresse suivante : https://tech.humanrights.gov.au/sites/default/files/2018-07/Human%20Rights%20and%20Technology%20Issues%20Paper%20FINAL.pdf (consulté le 5.11.18).

Barocas, S. et Selbst, A.D. (2016) 'Big data's disparate impact.' *Cal. L. Rev*. 104: 671.

Barr, S. (2018) 'Computer-Generated Instagram Account Astounds Internet'. *The Independent*. 1er mars. Disponible à l'adresse suivante : https://www.independent.co.uk/life-style/fashion/instagram-model-computer-generated-shudu-gram-internet-cameron-james-a8234816.html (consulté le 7.11.18)

Bennett Moses, L. et de Koker, L. (2017). 'Open Secrets: Balancing Operational Secrecy and Transparency in the Collection and Use of Data by National Security and Law Enforcement Agencies' *Melbourne University Law Review* 41(2): 530.

Bostrom, N. (2014) *Superintelligence: Paths, Dangers, Strategies.* Oxford: Oxford University Press.

Bovens, M. (2007) 'New forms of accountability and EU-governance' *Comparative European Politics* 5(1): 104-120.

Boyd, D., et Crawford, K. (2012) 'Critical Questions for Big Data'. *Information, Communication and Society* 15(5):662-79.

Brownsword, R. (2005) 'Code, control, and choice: why East is East and West is West.' *Legal Studies 25*(1): 1-21.

Brownsword, R., Scotford, E. et Yeung., K. (éd.) (2017). *Oxford Handbook on Law, Regulation and Technology*. Oxford: Oxford University Press.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B. et Anderson, H. (2018). *The malicious use of artificial intelligence: Forecasting, prevention, and mitigation.'* Disponible à l'adresse suivante : *arXiv preprint arXiv:1802.07228* (consulté le 5.11.2018).

Bublitz, J.C. (2013) 'My mind is mine!? Cognitive liberty as a legal concept' in *Cognitive Enhancement: An Interdisciplinary Perspective,* établi par E. Hildt et A.G Franke. Dordrecht: Springer, 233-264.

Bureau exécutif du président des États-Unis (2016), Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights. Disponible à l'adresse suivante : https://obamawhitehouse.archives.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf (consulté le 6.11.18).

Burrell, J. (2016) 'How the machine 'thinks': Understanding opacity in machine learning algorithms.' *Big Data & Society* 3(1):1-12.

Bryson, J. J. et A. Theodorou (2018) 'How Society can Maintain Human-Centric Artificial Intelligence'. In M. Toivonen-Noro and E. Saari (éd.) *Human-Centered Digitalization and Services*.

Bryson, J.J. (2010) 'Robots should be slaves' in *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues,* établi par Y Wilks. Amsterdam: John Benjamins Publishing, 63-74.

Bygrave, L.A. (2017) 'Hard-wiring Privacy' in Brownsword, R., Scotford, E. et Yeung., K. (éd.) (2017). *Oxford Handbook on Law, Regulation and Technology*. Oxford : Oxford University Press.

Cane, P. (2002) *Responsibility in Law and Morality*. Oxford : Hart Publishing.

Carton, S., Helsby, J., Joseph, K., Mahmud, A., Park, Y., Walsh, J., Cody, C., Patterson, C.P.T., Haynes, L. et Ghani, R., (2016) 'Identifying police officers at risk of adverse events.' In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 67-76. Disponible à l'adresse suivante : https://dl.acm.org/citation.cfm?id=2939698 (consulté le 5.11.18).

Cath, C. (2018) 'Governing artificial intelligence: ethical, legal and technical opportunities and challenges' 376 *Phil Trans A*: *Mathematical, Physical and Engineering Sciences*. https://doi.org/10.1098/rsta.2018.0080

Centre européen de stratégie politique (2018) 'The age of artificial intelligence: Towards a European Strategy for Human-Centric Machines'. Disponible à l'adresse suivante : https://ec.europa.eu/epsc/sites/epsc/files/epsc_strategicnote_ai.pdf (consulté le 6.11.2018). .

Chen, A. (2014) 'The Labourers Who Keep Dick Pics and Beheadings Out of Your Facebook News Feed', *Wired*, 23 octobre. Disponible à l'adresse suivante : https://www.wired.com/2014/10/content-moderation/ (consulté le 5.11.18).

Chesney, B. et D. Citron (2019) 'Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security.' *California Law Review* 107 : à paraître.

Citron, D. K. (2008) 'Technological Due Process.' *Washington University Law Review* 85*:* 1249-1313.

Cohen, J. E. (2017). 'Affording Fundamental Rights.' *Critical Analysis of Law.* 4(1): 76-90.

Comité économique et social européen, avis sur la « Communication de la Commission au Parlement européen, au Conseil européen, au Conseil, au Comité économique et social européen et au Comité des régions – L'intelligence artificielle pour l'Europe ». COM(2018) 237 final, disponible sur : https://eur-lex.europa.eu/legal-content/FR/TXT/PDF/?uri=CELEX:52018AE2369&from=FR (consulté le 5.11.2018)

Commission des affaires juridiques du Parlement européen (2017), *Rapport contenant des recommandations à la Commission concernant des règles de droit civil sur la robotique.* Rapporteure : M. Delvaux. 2015/2103 (INL)). Disponible sur : http://www.europarl.europa.eu/doceo/document/A-8-2017-0005_FR.html (consulté le 5.11.2018).

Commission européenne (2018a), *Un plan coordonné dans le domaine de l'intelligence artificielle.* Communication de la Commission au Parlement européen, au Conseil européen, au Conseil, au Comité économique et social européen et au Comité des régions sur l'intelligence artificielle pour l'Europe.

Commission européenne (2018b), Consumer market study on online market segmentation through personalised pricing/offers in the European Union, 19 juillet. Disponible à l'adresse suivante : https://ec.europa.eu/info/publications/consumer-market-study-online-market-segmentation-through-personalised-pricing-offers-european-union_en (consulté le 3 mai 2019).

Commission européenne (2018c), *Liability for emerging digital technologies. (Document de travail des Services de la Commission sur la responsabilité du fait des produits pour les nouvelles technologies numériques)* [SWD(2018)137]). COM (2018) 237 final. Disponible sur https://eur-lex.europa.eu/legal-content/FR/TXT/PDF/?uri=CELEX:52018AE2369&from=FR (consulté le 5.11.2018)

Commission européenne (2018d), Evaluation of Council Directive 85/374/EEC of 25 July 1985 on the approximation of the laws, regulations and administrative provisions of the Member States concerning liability. COM(2018) 246 final. (Évaluation de la Directive du Conseil relative au rapprochement des dispositions législatives, réglementaires et administratives des États membres en matière de responsabilité du fait des produits défectueux). Disponible à l'adresse suivante : https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52018SC0157&from=EN (SWD(2018)157 final. (85/374/CEE) (consulté le 7.11.18)

Commission européenne (2019b), Groupe d'experts de haut niveau sur l'IA, *A Definition of AI: Main Capabilities and Disciplines*.

Conn, A. (2017) 'Research for Beneficial Artificial Intelligence' *Future of Life Institute.* Disponible à l'adresse suivante : https://futureoflife.org/2017/12/27/research-for-beneficial-artificial-intelligence/?cn-reloaded=1&cn-reloaded=1 (consulté le 5.11.18)

Conseil de l'Europe. Recommandation CM/Rec(2018)2 du Comité des Ministres aux États membres sur les rôles et les responsabilités des intermédiaires d'internet Voir ce lien : https://search.coe.int/cm/Pages/result_details.aspx?ObjectId=0900001680790e37 (consulté le 7.11.18)

Conseil de l'Europe, Assemblée parlementaire (2017), Commission de la culture, de la science, de l'éducation et des médias. *La convergence technologique, l'intelligence artificielle et les droits de l'homme*. 10 avril. Doc 14288. Voir ce lien : https://assembly.coe.int/nw/xml/XRef/Xref-XML2HTML-fr.asp?fileid=23531&lang=fr (consulté le 7.11.18)

Convention de sauvegarde des droits de l'homme et libertés fondamentales (CEDH)

Cowley, J. (2018) 'Beijing subway to install facial recognition as fears grow of China surveillance powers.' 19 juin. *The Telegraph*. Disponible à l'adresse suivante : https://www.telegraph.co.uk/news/2018/06/19/beijing-subway-install-facial-recognition-fears-grow-china-surveillance/ (consulté le 5.11.18).

Crawford, K. et Schultz, J. (2014) 'Big data and due process: Toward a framework to redress predictive privacy harms.' *Boston College Law Review* 55:93.

Danaher, J. (2016) 'Robots, law and the retribution gap.' *Ethics and Information Technology* 18(4): 299-309.

Datta, A., Sen, S. et Zick, Y. (2016), Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *Security and Privacy (SP), 2016 IEEE Symposium on*, 598-617. IEEE.

Davidow, B. (2014). 'Welcome to Algorithmic Prison - The use of Big Data to to profile citizens is subtly, silently constraining freedom'. *The Atlantic*. 20 février.

Dennett, D.C., (1997), 'When HAL kills, who's to blame' in *HAL's Legacy: 2001's Computer as Dream and Reality,* établi par D.G. Stork. MIT Press.

Desai, D.R. et Kroll, J. (2017), 'Trust but Verify: A Guide to Algorithms and the Law'. *Harvard Journal of Law & Technology* 31:1-64

De Streel, A., Buiten, M. & Peitz, M. (2018), 'Liability of online hosting platforms: should exceptionalism end?' *Centre on Regulation in Europe Report.* Disponible à l'adresse suivante : http://www.cerre.eu/sites/cerre/files/180912_CERRE_LiabilityPlatforms_Final_0.pdf (consulté le 5.11.18)

Dietrich, W., Mendoza, C et Brennnan, T. (2016), 'Compass risk scales: Demonstrating accuracy, equity and predictive parity'. Northpointe.

Donahoe, E. (2016), 'So Software Has Eaten the World: What Does it Mean for Human Rights, Security and Governance?' 18 mars, *Just Security*. Disponible à l'adresse suivante : https://www.justsecurity.org/30046/software-eaten-world-human-rights-security-governance/ (consulté le 5.11.18).

Doshi-Velez, F., Ge, Y. et Kohane, I. (2014), 'Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis.' *Pediatrics 133*(1): e54-e63.

Draper, N. A. et J. Turow (2017), 'Audience Constructions, Reputations and Emerging Media Technologies: New Issues of Legal and Social Policy' in *The Oxford Handbook on Law, Regulation and Technology*, établi par R. Brownsword, E. Scotford & K. Yeung. Oxford : Oxford University Press.

Edwards, L. et Veale, M. (2017), 'Slave to the Algorithm: Why a Right to an Explanation Is Probably Not the Remedy You Are Looking for.' *Duke L. & Tech. Rev.* 16.

Ekbia, H. et B. Nardi (2014), 'Heteromation and its (dis)contents: The invisible division of labor between human and machine. *First Monday* 19(6). Disponible à l'adresse suivante : https://firstmonday.org/article/view/5331/4090#author (consulté le 7.11.18)

Elish, M.C. (2016) : 'Letting Autopilots Off the Hook: Why do we blame humans when automation fails?' 16 juin. Disponible à l'adresse suivante : http://www.slate.com/articles/technology/future_tense/2016/06/why_do_blame_humans_when_automation_fails.html (consulté le 5.11.18)

Engineering and Physical Sciences Research Council (EPSRC) : https://epsrc.ukri.org/ (consulté le 6.11.18)

Eschelman, A., (2016), 'Moral responsibility', *The Stanford Encyclopedia of Philosophy* (édition hiver 2016). Établi par Edward N. Zalta. Disponible à l'adresse suivante : https://plato.stanford.edu/archives/win2016/entries/moral-responsibility/ (consulté le 6.11.18)

Groupe d'experts de haut niveau sur l'IA (GEHN IA) de la Commission européenne, *Ethics Guidelines for Trustworthy Artificial Intelligence* (Lignes directrices en matière d'éthique pour une intelligence artificielle de confiance)

Groupe européen d'éthique des sciences et des nouvelles technologies (GEE) (2018). Déclaration sur l'intelligence artificielle, la robotique et les systèmes « autonomes ». Disponible sur : https://ec.europa.eu/research/ege/pdf/ege_ai_statement_2018_fr.pdf (consulté le 6.11.18)

Farr, C. (2016) 'If You Want Life Insurance, Think Twice Before Getting A Genetic Test'. 2 juillet. Disponible à l'adresse suivante : https://www.fastcompany.com/3055710/if-you-want-life-insurance-think-twice-before-getting-genetic-testing (consulté le 7.11.18)

Ferguson, A.G. (2016) 'Policing predictive policing.' *Wash. UL Rev.* 94: 1109.

Ferraris, V., Bosco, F. et D'Angelo, E. (2013), 'The impact of profiling on fundamental rights'. Disponible à l'adresse suivante : SSRN: https://ssrn.com/abstract=2366753 ou http://dx.doi.org/10.2139/ssrn.2366753 (consulté le 6.11.2018).

Financial Times (2018), FT Series , *The AI arms race.* Disponible à l'adresse suivante : https://www.ft.com/content/21eb5996-89a3-11e8-bf9e-8771d5404543

ForbrukerRadet. Norwegian Consumer Council (2018), *Deceived by Design*. 27 juin. Disponible à l'adresse suivante : https://fil.forbrukerradet.no/wp-content/uploads/2018/06/2018-06-27-deceived-by-design-final.pdf (consulté le 7.11.18).

Galligan, D.G. (1997), *Due Process and Fair Procedures*. Clarendon Press : Oxford.

Galligan, D.G. (2006), *Law in Modern Society*. OUP : Oxford.

Gandy, O.H. (1993), *The panoptic sort: a political economy of personal information.* Westview.

Gardner, J. (2003), 'The Mark of Responsibility.' *Oxford Journal of Legal Studies.* 23(2): 157-171.

Gardner, J. (2008), *Introduction to H.L.A. Hart, Punishment and Responsibility: Essays in the Philosophy of Law: Second Edition.* OUP Oxford.

The Guardian (2016), 'Microsoft 'deeply sorry' for racist and sexist tweets by AI chatbot'. Disponible à l'adresse suivante : : https://www.theguardian.com/technology/2016/mar/26/microsoft-deeply-sorry-for-offensive-tweets-by-ai-chatbot (consulté le 6.11.18)

Gilliker, P (2000), 'A "new" head of damages: damages for mental distress in the English law of torts'. *Legal Studies* 20: 19-41.

Girardin, F. et Blat, J. (2010), The co-evolution of taxi drivers and their in-car navigation systems. *Pervasive and Mobile Computing*. 6(4): 424-434.

Glas, L.R., 'The Functioning of The Pilot-Judgment Procedure Of The European Court Of Human Rights In Practice' (2016), *Netherlands Quarterly of Human Rights* 34(1): 41.

Gorton, W.A. (2016), 'Manipulating Citizens: How Political Campaigns' Use of Behavioral Social Science Harms Democracy' *New Political Science*, *38*(1), pp.61-80.

Greene, D., Hoffman, A.L et Stark, L., 'Better, Nicer, Clearer and Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning' (2019), Hawaii International Conference on System Sciences, DOI: 10.24251/HICSS.2019.258. Disponible à l'adresse suivante : http://dmgreene.net/wp-content/uploads/2018/09/Greene-Hoffman-Stark-Better-Nicer-Clearer-Fairer-HICSS-Final-Submission.pdf (consulté le 6.05.19)

Gunkel, D.J. (2017), 'Mind the gap: responsible robotics and the problem of responsibility', *Ethics and Information Technology*, pp.1-14.

Hagendorf, T. (2019), 'The Ethics of AI Ethics: An Evaluation of Guidelines'. Disponible à l'adresse suivante: https://arxiv.org/abs/1903.03425 (consulté le 6 mai 2019).

Hall, W. et Pesenti, J (2017), *Growing the Artificial Intelligence Industry in the UK*. Disponible à l'adresse suivante : https://www.gov.uk/government/publications/growing-the-artificial-intelligence-industry-in-the-uk (consulté le 7.11.18)

Hallevy, G. (2015), *Liability for crimes involving artificial intelligence systems*. Springer International Publishing.

Hanson, F.A. (2009), 'Beyond the skin bag: on the moral responsibility of extended agencies', *Ethics and information technology*, *11*(1), pp. 91-99.

Hart, H.L.A., ((1968) 2008), *Punishment and responsibility: Essays in the philosophy of law*. Oxford University Press: Oxford.

Helberger, N., Pierson, J. et Poell, T. (2018), Governing online platforms: From contested to cooperative responsibility. *The Information Society* 34(1): 1-14, DOI:10.1080/01972243.2017.1391913

Hildebrandt, M. et Gutwirth, S., (2008), Profiling the European Citizen. Springer : Pays-Bas.

Hildebrandt, M. (2013), 'Criminal Law and Technology in a Data-Driven Society' in M.D. Dubber et T. Hornle (éd.), *Oxford Handbook of Criminal Law*. Oxford : Oxford University Press 174-197.

Hildebrandt, M., (2015), *Smart technologies and the end (s) of law: novel entanglements of law and technology*. Edward Elgar Publishing.

Hildebrandt, M. (2016), 'Data-gestuurde intelligentie in het strafrecht' In : E.M.L. Moerel, J.E.J. Prins, M.Hildebrandt, T.F.E. Tjong Tjin Tai, G-J. Zwenne & A.H.J. Schmidt (éd.), *Homo Digitalis*. Nederlandse : Wolters Kluwer.

Himma, K.E. (2009), 'Artificial Agency, Consciousness, and the Criteria for Moral Agency: What Properties Must an Artificial Agent have to be a Moral Agent.' *Ethics and Information Technology* 11(1):24.

Horsey, K et Rackley, E (2014), *Tort Law*. 4th ed. Oxford University Press: Oxford.

Hutson, M (2018), 'Lip-reading artificial intelligence could help the deaf—or spies.' 31 juillet. *Science*. doi:10.1126/science.aau9601. Disponible à l'adresse suivante : http://www.sciencemag.org/news/2018/07/lip-reading-artificial-intelligence-could-help-deaf-or-spies (consulté le 6.11.18)

Huxley, A. (1932), *Le meilleur des mondes (Brave New World).* Chatto & Windus.

IEEE, Global Initiative for Ethical Considerations in AI and Autonomous Systems (2017). Disponible à l'adresse suivante : https://standards.ieee.org/industry-connections/ec/autonomous-systems.html (consulté le 7.11.18)

Irani, L. (2015), 'Difference and Dependence among Digital Workers: The Case of Amazon Mechanical Turk'. *The South Atlantic* Quarterly 114(1): 225-234.

Jasanoff, S. (2016), *The ethics of invention: technology and the human future*. WW Norton & Company.

Johnson, D.G. (2006), 'Computer systems: Moral entities but not moral agents.' *Ethics and information technology*. 8(4): 195-204.

Johnson, D.G. et Powers, T.M. (2005), 'Computer systems and responsibility: A normative look at technological complexity', *Ethics and information technology*, *7*(2), pp.99-107

Kaminski, M.E. et Witnov, S. (2014), 'The Conforming Effect: First Amendment Implications of Surveillance, Beyond Chilling Speech.' *U. Rich. L. Rev.*, 49: 465.

Keen, A (2018), *How to Fix the Future.* Atlantic Books : Londres.

Kitchin, R (2014), *The Data Revolution*. Sage : Los Angeles.

Korff, D. et Browne, I. (2013), 'The use of the Internet & related services, private life & data protection: trends, technologies, threats and implications', Conseil de l'Europe, T-PD(2013)07. Disponible à l'adresse suivante : SSRN: https://ssrn.com/abstract=2356797 (consulté le 6.11.18)

Koops, B.J., Hildebrandt, M et Jaquet-Chiffelle, D-O. (2010), 'Bridging the Accountability Gap: Rights for New Entities in the Information Society?' *Minnesota Journal of Law, Science and Technology* 497.

Kosinski, M., Stillwell, D & Graepel, T. (2013), 'Private traits and attributes are predictable from digital records of human behaviours.' *Proceedings of the National Academy of Science* 110: 5802-5805.

Kramer, A.D., Guillory, J.E. et Hancock, J.T. (2015), 'Experimental evidence of massive-scale emotional contagion through social networks.' *Proceedings of the National Academy of Sciences*, 8788–8790.

Kuflik, A. (1999), 'Computers in control: Rational transfer of authority or irresponsible abdication of autonomy?' *Ethics and Information Technology* 1(3): 173-184.

Lanzing, M. (2018), '"Strongly Recommended" Revisiting Decisional Privacy to Judge Hypernudging in Self-Tracking Technologies.' *Philosophy & Technology* https. doi.org/10.1007/s13347-018-0316-4.

Latonero, M (2019), *Governing Artificial Intelligence: Upholding Human Rights and Human Dignity*, Data & Society. Disponible à l'adresse suivante : https://datasociety.net/wp-content/uploads/2018/10/DataSociety_Governing_Artificial_Intelligence_Upholding_Human_Rights.pdf (consulté le 6 mai 2019).

Leonelli , S (2018), 'Rethinking Reproducibility as a Criterion for Research Quality' in Fiorito, L., Scheall, S., and Suprinyak, C.E. (éd.) *Including a Symposium on Mary Morgan: Curiosity, Imagination, and Surprise (Research in the History of Economic Thought and Methodology, Volume 36B)* Emerald Publishing Limited, 129 – 146.

Liu, H.Y. (2016), 'Refining responsibility: Differentiating two types of responsibility issues raised by autonomous weapons systems'. Établi par N. Bhuta, S. Beck, R. Geiss, H.Y. Liu et C. Kress. *Autonomous weapons systems—Law, ethics policy* at 325-344. CUP: New York.

Liu, H.Y. et Zawieska, K. (2017), 'From responsible robotics towards a human rights regime oriented to the challenges of robotics and artificial intelligence.' *Ethics and Information Technology*. 19(3):1-13.

Lodge, M. et K. Wegrich (2012), *Managing Regulation*. London, Palgrave Macmillan.

Lohr, J. Maxwell, W. et Watts, P. (2019), 'Legal practitioners' approach to regulating AI risks.' In *Algorithmic Regulation*. Établi par K. Yeung & M. Lodge. OUP: Oxford. Sous presse.

Loui, M. C. et Miller, K. W. (2008), 'Ethics and Professional Responsibility in Computing'. In *Wiley Encyclopedia of Computer Science and Engineering*. Établi par B. W. Wah. doi : 10.1002/9780470050118.ecse909

Lunney, M et Oliphant, K (2013), *Tort Law*. 5e édition. Oxford University Press : Oxford.

Mangan, D. (2017), 'Lawyers could be the next profession to be replaced by computers.' 17 février. Disponible à l'adresse suivante : https://www.cnbc.com/2017/02/17/lawyers-could-be-replaced-by-artificial-intelligence.html (consulté le 6.11.18).

Mantelero, A. (2018), 'AI and Big Data: A blueprint for a human rights, social and ethical impact assessment.' *Computer Law & Security Review* 34(4) : 754-772.

Mantalero, A. (2019), *Intelligence artificielle et protection des données : enjeux et solutions possibles.* Rapport preparé par le Conseil de l'Europe, Comité consultatif de la Convention pour la protection des personnes à l'égard du traitement automatisé des données à caractère personnel, T-PD(2018)09Rev. Guidelines on Artifical Intelligence and Data Protection. Disponible sur https://rm.coe.int/intelligence-artificielle-et-protection-des-donnees-enjeux-et-solution/168091f8a5 (consulté le 6 mai 2019).

Matthias, A. (2004), 'The responsibility gap: Ascribing responsibility for the actions of learning automata.' *Ethics and information technology* 6(3) : 175-183.

Mayer-Schönberger, V. et Cukier, K. (2013), *Big Data–A Revolution That Will Transform How We Live, Think and Work*. Londres, John Murray.

Menn, J. et D. Volz (2016), 'Exclusive: Google, Facebook quietly move toward automatic blocking of extremist videos.' Disponible à l'adresse suivante : https://www.reuters.com/article/us-internet-extremism-video-exclusive-idUSKCN0ZB00M. (consulté le 7.11.18)

McSherry, M. (2018), 'Will AI Widen or Weaken the Global Digital Divide?' *Medium,* 21 mai (consulté le 01.05.19).

Merton, R. K. (1942), 'The Normative Structure of Science'. In *The Sociology of Science: Theoretical and Empirical Investigations*. Établi par R. K. Merton. Chicago, Il, University of Chicago Press : 267-278.

Metcalf, J., & Crawford, K. (2016), 'Where are human subjects in Big Data research? The emerging ethics divide' *Big Data & Society*. https://doi.org/10.1177/2053951716650211

Metzinger (2019), 'Ethics Washing Made in Europe', *Der Taggespiegel*, 8 avril. Disponible à l'adresse suivante : https://www.tagesspiegel.de/politik/eu-guidelines-ethics-washing-made-in-europe/24195496.html (consulté le 06.05.19).

Michalski, R.S., Carbonell, J.G. et Mitchell, T.M. (éd.), (2013) *Machine learning: An artificial intelligence approach*. Springer Science & Business Media.

Miller, A.A. (2014), 'What Do We Worry about When We Worry about Price Discrimination -The Law and Ethics of Using Personal Information for Pricing' *J. Tech. L. & Pol'y* 19 :41.

Morgan, B. et Yeung, K., 2007, *An Introduction to Law and Regulation: Text and Materials*. Cambridge : Cambridge University Press.

Moses, L.B. et Koker, L.D. (2017), 'Open secrets: Balancing operational secrecy and transparency in the collection and use of data by national security and law enforcement agencies.' *Melb. UL Rev.* 41: 530.

Narula, G. (2018), 'Everyday Examples of Artificial Intelligence and Machine Learning.' 29 octobre. Disponible à l'adresse suivante : https://www.techemergence.com/everyday-examples-of-ai/ (consulté le 7.11.18).

Nemitz, P. (2018). 'Constitutional Democracy and Technology in the Age of Artificial Intelligence'. *Phil Trans. A*. 376.

Nevejans, N. (2016) Parlement européen, Commission des affaires juridiques (2016). *Règles européennes de droit civil en robotique.* Étude pour la commission juridique (JURI). Disponible sur : http://www.europarl.europa.eu/RegData/etudes/STUD/2016/571379/IPOL_STU(2016)571379_FR.pdf (consulté le 7.11.18)

Nilsson, N.J. (2014), *Principles of artificial intelligence*. Morgan Kaufmann.

Nissenbaum, H. (1996), 'Accountability in a computerized society.' *Science and Engineering Ethics*. 2(1): 25-42.

Nissenbaum, H. (1996), 'Accountability in a Computerized Society.' *Science and Engineering Ethics* 2 : 25-42.

Nissenbaum, H. (2010), *Privacy in Context: Technology, Policy and the Integrity of Social Life*. Stanford CA : Stanford Law Books.

Nissenbaum, H. (2011), 'A contextual approach to privacy online.' *Daedalus the Journal of the American Academy of Arts & Sciences*. 140(4) : 32–48.

Noto La Diega, G. (2018), 'Against the Dehumanisation of Decision-Making: Algorithmic Decisions at the Crossroads of Intellectual Property, Data Protection and Freedom of Information.' J*ournal of Intellectual Property, Information Technology and E-Commerce Law*. 9 (3) : 11-16.

Nuffield Foundation and the Leverhulme Centre for the Future of Intelligence (2019), *Ethical and Social Implications of Algorithms, Data and Artificial Intelligence: A Roadmap for Research*. Disponible à

l'adresse suivante : https://www.adalovelaceinstitute.org/nuffield-foundation-publishes-roadmap-for-ai-ethics-research/ (consulté le 06.05.19).

Oberdiek, J. (2017), *Imposing risk: a normative framework*. Oxford : Oxford University Press.

Olsen, M. (1965), *The Logic of Collective Action - Public Goods and the Theory of Groups*. Cambridge, MA, Harvard University Press.

Oliver, D., (1994), 'Law, politics and public accountability. The search for a new equilibrium.' *Public Law* 238-238.

O'Neil, C. (2016), *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

Oswald, M., Grace, J., Urwin, S. et Barnes, G.C., (2018), 'Algorithmic risk assessment policing models: lessons from the Durham HART model and 'Experimental' proportionality' *Information & Communications Technology Law*. 27(2):223-250.

Oxera (2018), *Consumer Data in Online Markets*. Article pour *Which?*, 5 juin.

Pasquale, F. (2015), *The Black Box Society: The secret algorithms that control money and information*. Harvard University Press.

Pariser, E. (2012), *The Filter Bubble*. Londres, Penguin Books.

Pasquale, F (2015), *The Black Box Society*. Boston : Harvard University Press.

Pichai, S. (2018), 'AI at Google: our principles'. 7 juin. Disponible à l'adresse suivante : https://www.blog.google/technology/ai/ai-principles/ (consulté le 6.11.18)

Polyakov, A. (2018), 'Seven Ways Cybercriminals Can Use Machine Learning'. Disponible à l'adresse suivante : https://www.forbes.com/sites/forbestechcouncil/2018/01/11/seven-ways-cybercriminals-can-use-machine-learning/#1e42a2a81447 (consulté le 6.11.18)

Power, M. (1997), *The Audit Society*. Oxford : Oxford University Press.

Powles, J. (2015), 'We are citizens, not mere physical masses of data for harvesting'. *The Guardian*. 11 mars. Disponible à l'adresse suivante : https://www.theguardian.com/technology/2015/mar/11/we-are-citizens-not-mere-physical-masses-of-data-for-harvesting (consulté le 5.11.18)

Prainsack, B. (2019), Logged out: Ownership, exclusion and public value in the digital data and formation commons.' *Big Data & Society*. https://doi.org/10.1177/2053951719829773.

Rainey, B., Wicks, E. and Ovey, C. (2014) *Jacobs, White and Ovey: the European Convention on Human Rights* (6e édition). Oxford : Oxford University Press.

Raso, F., Hilligoss, H,, Krishnamurthy, V., Bavitz,C., & Kim,. (2018) *Artificial Intelligence & Human Rights: Opportunities & Risks.* Berkman Klein Center for Internet & Society, Harvard University. Disponible à l'adresse suivante : https://cyber.harvard.edu/publication/2018/artificial-intelligence-human-rights (consulté le 5.11.2018)

Raz, J. (1986), *The Morality of Freedom.* Oxford : Oxford University Press.

Représentant spécial du Secrétaire général des Nations Unies (2011), *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.* Approuvé par le Conseil des droits de l'homme dans sa résolution 17/4 du 16 juin 2011. Disponible sur : https://www.ohchr.org/Documents/Publications/GuidingPrinciplesBusinessHR_FR.pdf. (consulté le 7.11.18)

Rieder, B. (2016), 'Big data and the paradox of diversity'. *Digital Culture & Society* 2(2) : 39-54.

Risse, M. (2018), 'Human Rights and Artificial Intelligence: An Urgently Needed Agenda'. *Harvard Kennedy School Faculty Research Working Paper Series*. RWP18-015. Disponible à l'adresse suivante : https://research.hks.harvard.edu/publications/getFile.aspx?Id=1664 (consulté le 6.11.18)

Royvroy, A. (2016), « Des données et des hommes ». Droits et libertés fondamentaux dans un monde de données massives. *Rapport pour le Bureau du comité consultatif de la Convention pour la protection des personnes à l'égard du traitement automatisé des données à caractère personnel*, Conseil de l'Europe, TD-PD-BUR. Disponible sur https://rm.coe.int/16806b1659%202015).pdf (consulté le 6.11.18).

Russell, S.J. et Norvig, P. (2016), *Artificial intelligence: a modern approach*. Malaisie : Pearson Education Limited.

SAE International (2018), *Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems*. Disponible à l'adresse suivante : https://www.sae.org/standards/content/j3016_201806/ (consulté le 6.11.18).

Samek et al. (2017), 'Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models.' *ITU Journal: ICT Discoveries*, numéro spécial n° 1 : 1-10.

Sandvig, C., Hamilton, K., Karahalios, K. et Langbort, C., (2014), Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry*, 1-23. Disponible à l'adresse suivante : http://www-personal.umich.edu/~csandvig/research/Auditing%20Algorithms%20--%20Sandvig%20--%20ICA%202014%20Data%20and%20Discrimination%20Preconference.pdf (consulté le 7.11.18)

Schut, M. et Wooldridge, M. (2000), juin. Intention reconsideration in complex environments. In *Proceedings of the fourth international conference on Autonomous agents*. 209-216. ACM.

Schwab, K., Davies, N. et Nadella, S. (2018), *Shaping the Fourth Industrial Revolution*. Forum économique mondial.

Scott, M. et Isaac, M. (2016), 'Facebook Restores Iconic Vietnam War Photo It Censored for Nudity'. Disponible à l'adresse suivante : : https://www.nytimes.com/2016/09/10/technology/facebook-vietnam-war-photo-nudity.html (consulté le 6.11.18)

Shadbolt, N. et Hampson, R. (2018), *The Digital Ape*. Scribe : Melbourne.

Skilton, M. et Hovsepian, F. (2017), *The 4th Industrial Revolution: Responding to the Impact of Artificial Intelligence on Business*. Springer.

Smith, A. (2018), 'Franken-algorithms: the deadly consequences of unpredictable code'. *The Guardian*. 30 août. Disponible à l'adresse suivante : https://www.theguardian.com/technology/2018/aug/29/coding-algorithms-frankenalgos-program-danger (consulté le 6.11.18)

Solove, D.J. (2012), 'Introduction: Privacy self-management and the consent dilemma'. *Harvard Law Review* 126 : 1880.

Solum, L.B. (1991), 'Legal personhood for artificial intelligences' *NCL Rev.*70: 1231.

Sparrow, R. (2007), 'Killer Robots', *Journal of Applied Philosophy* 24(1) : 62.

Su, Xiaoyuan, et Taghi M. Khoshgoftaar (2009), 'A survey of collaborative filtering techniques.' *Advances in artificial intelligence*.

Swan, M. (2015), Connected car: quantified self becomes quantified car. *Journal of Sensor and Actuator Networks* 4(1) : 2-29.

Sullins, J.P. (2005), 'Ethics and artificial life: From modelling to moral agents' *Ethics and Information technology*, *7*(3), p. 139.

Taplin, J. (2018), *Move fast and break things*. Politikens Forlag.

Teubner, G (2006), 'Rights of Non-Humans? Electronic Agents and Animals as New Actors in Politics and Law.' *Journal of Law and Society.* 33: 497-521.

Tebuner, G (2018), 'Digital Personhood? The Status of Autonomous Software Agents in Private Law'. Disponible via le réseau SSRN (consulté le 21.5.2019).

The Economist (2017), 'Imitating people's speech patterns could bring trouble'. Disponible à l'adresse suivante : https://www.economist.com/science-and-technology/2017/04/20/imitating-peoples-speech-patterns-precisely-could-bring-trouble (consulté le 7.11.18)

The Economist (2018a), 'Images aren't everything: AI, radiology and the future of work'. Disponible à l'adresse suivante : https://www.economist.com/leaders/2018/06/07/ai-radiology-and-the-future-of-work (consulté le 6.11.18)

The Economist (2018b), 'The techlash against Amazon, Facebook and Google - and what they can do'. Disponible à l'adresse suivante : https://www.economist.com/briefing/2018/01/20/the-techlash-against-amazon-facebook-and-google-and-what-they-can-do (consulté le 6.11.18)

The Royal Academy of Engineering (2009), *Autonomous Systems: Social Legal and Ethical Issues.* Août. Disponible à l'adresse suivante : https://www.raeng.org.uk/publications/reports/autonomous-systems-report (consulté le 6.11.18).

The Royal Society (2017), *Machine Learning: The power and promise of computers that learn by example.* Avril. Disponible à l'adresse suivante : https://royalsociety.org/~/media/policy/projects/machine-learning/publications/machine-learning-report.pdf (consulté le 6.11.18).

Thomas, M (2015), 'Should We Trust Computers?' *Gresham Lectures*: London. 20 Octobre. Disponible à l'adresse suivante : https://www.gresham.ac.uk/lectures-and-events/should-we-trust-computers (consulté le 03.05.19).

Thomas, M (2017a) 'Safety Critical Systems', *Gresham Lectures*: London. 10 Janvier. Disponible à l'adresse suivante : https://www.gresham.ac.uk/lectures-and-events/safety-critical-systems (consulté le 3 mai 2019).

Thomas, M (2017b), 'Is Society Ready for Driverless Cars?' Gresham Lectures, Londres, 24 octobre. Disponible à l'adresse suivante : https://www.gresham.ac.uk/lectures-and-events/is-society-ready-for-driverless-cars (consulté le 03.05.19)

Thompson, D. (1980), 'Moral Responsibility of Public Officials: The Problem of Many Hands', *The American Political Science Review* 74(4): 905-916. doi:10.2307/1954312

The Toronto Declaration: Protecting the rights to equality and non-discrimination in machine learning systems (2018). Disponible à l'adresse suivante : https://www.accessnow.org/cms/assets/uploads/2018/08/The-Toronto-Declaration_ENG_08-2018.pdf (consulté le 6.11.18)

Townley, C., Morrison, E., & Yeung, K. (2017), 'Big Data and Personalized Price Discrimination in EU Competition Law'. *Yearbook of European Law* 36(1): 683-748.

Tufekci, Z. (2015), 'Algorithmic harms beyond Facebook and Google: Emergent challenges of computational agency'. *J. on Telecomm. & High Tech. L.* 13: 203.

UK Competition and Markets Authority *(Autorité britannique de la concurrence et des marchés)* (2018), *Pricing Algorithms*. 8 Octobre. CMA 94. Disponible à l'adresse suivante : https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/746353/Algorithms_econ_report.pdf (consulté le 03.05.19)

UK Information Commissioner's Office *(Bureau du Commissaire à l'information du Royaume-Uni)* (2018), *Democracy Disrupted – Personal Information and Political Influence*. 11 July. Disponible à l'adresse suivante : https://ico.org.uk/media/2259369/democracy-disrupted-110718.pdf (consulté le 03.05.19).

UK Government *(Gouvernement du Royaume-Uni)* (2019) *Online Harms White Paper*, CP 57. Disponible à l'adresse suivante : https://www.gov.uk/government/consultations/online-harms-white-paper/online-harms-white-paper-executive-summary--2 (consulté le 03.05.19).

UK Department for Business, Energy and Industrial Strategy *(Ministère britannique des entreprises, de l'énergie et de la stratégie industrielle)* (2018), *Artificial Intelligence Sector Deal*. Disponible à l'adresse suivante : https://www.gov.uk/government/publications/artificial-intelligence-sector-deal (consulté le 6.11.18).

UK Department for Digital, Culture, Media and Sport *(Ministère britannique du numérique, de la culture, des médias et des sports)* (2018), 'Up to £50 million to develop world leading AI talent in the UK'. Disponible à l'adresse suivante : https://www.gov.uk/government/news/up-to-50-million-to-develop-world-leading-ai-talent-in-the-uk (consulté le 6.11.18).

UK House of Commons, Digital Culture Media and Sports Committee (*Chambre des communes du Royaume-Uni, Commission du numérique, de la culture, des médias et du sport*), *Disinformation and 'fake news' : Final Report*, Eighth Report of Session 2017-2019, 14 février, HC 1791. Disponible à l'adresse suivante : https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/1791/1791.pdf (consulté le 06.05.19).

Union européenne (1985), Directive 85/374/CEE du Conseil du 25 juillet 1985 relative au rapprochement des dispositions législatives, réglementaires et administratives des États membres en matière de responsabilité du fait des produits défectueux (OJ L 210, 7.8.1985, 29-33).

Université de Montréal (2017) *Déclaration de Montréal pour un développement responsable de l'intelligence artificielle*. Disponible sur https://www.declarationmontreal-iaresponsable.com/la-declaration (consulté le 6 mai 2019).

U.S. Citizenship and Immigration Service *(Service américain de la citoyenneté et de l'immigration)* (2018) *Meet Emma, Our Virtual Assistant*. Disponible à l'adresse suivante : https://www.uscis.gov/emma (consulté le 6.11.18).

U.S. Department of Transportation *(Département des transports des États-Unis)* (2017), *Automated Driving Systems: A Vision for Safety 2.0*. Disponible à l'adresse suivante : https://www.nhtsa.gov/sites/nhtsa.dot.gov/files/documents/13069a-ads2.0_090617_v9a_tag.pdf (consulté le 6.11.18).

Vaidhyanathan, S. (2011), The Googlization of Everything (And Why We Should Worry). University of California Press.

Van der Sloot, B. (2014), 'Do data protection rules protect the individual and should they? An assessment of the proposed General Data Protection Regulation'. *International Data Privacy Law* 4(4):307-325.

Van Est, R. et J.B.A. Gerritsen, avec l'aide de L. Kool (2017), *Human rights in the robot age: Challenges arising from the use of robotics, artificial intelligence, and virtual and augmented reality*. Rapport d'expert rédigé pour la Commission de la culture, de la science, de l'éducation et des médias de l'Assemblée parlementaire du Conseil de l'Europe (APCE). La Haye : Rathenau Instituut. Disponible à l'adresse suivante : https://www.rathenau.nl/sites/default/files/2018-02/Human%20Rights%20in%20the%20Robot%20Age-Rathenau%20Instituut-2017.pdf. (consulté le 6.11.18).

Veale, M. et Binns, R. (2017) 'Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data' *Big Data & Society* 4(2) doi : 10.1177/2053951717743530.

Wagner, B. (2017), « Étude sur les dimensions des droits humains dans les techniques de traitement automatisé des données (en particulier les algorithmes) et éventuelles implications réglementaires ». 6 octobre. Conseil de l'Europe, Comité d'experts sur les intermédiaires d'internet (MSI-NET). Disponible

sur https://rm.coe.int/algorithmes-et-droits-humains-etude-sur-les-dimensions-des-droits-huma/1680796d11 (consulté le 6.11.18)

Wagner, B. (2019), 'Ethics as an Escape from Regulation: From ethics-washing to ethics-shopping?' in *Being Profiling: Cogitas Ergo Sum.* Établi par M. Hildebrandt. Amsterdam University Press, Amsterdam. À paraître.

Wallace, R.J. (1994), *Responsibility and the Moral Sentiments*. Harvard University Press : Boston.

Watson, G. (2004), *Agency and Answerability: Selected Essays*. Clarendon Press : Oxford.

Weller, A. (2017), 'Challenges for transparency'. Document présenté en 2017, ICML Workshop on Human Interpretability in Machine Learning (WHI 2017), Sydney, NSW, Australie. Disponible à l'adresse suivante : *arXiv preprint arXiv:1708.01870*. (consulté le 6.11.18).

Which? (2018) *Control, Alt or Delete? Consumer research on attitudes to data collection and use.* Policy Research Report. Juin.

White, A. (2018), 'EU calls for $24 billion in AI to keep up with China, U.S.' Disponible à l'adresse suivante: https://www.bloomberg.com/professional/blog/eu-calls-24-billion-ai-keep-china-u-s/ (consulté le 6.11.18).

Wierzynski, C. (2018), 'The Challenges and Opportunities of Explainable AI'. 12 janvier. Disponible à l'adresse suivante : https://ai.intel.com/the-challenges-and-opportunities-of-explainable-ai/ (consulté le 27.3.18).

Yao, M. (2017), 'Chihuahua or muffin? My search for the best computer vision API'. Disponible à l'adresse suivante : https://medium.freecodecamp.org/chihuahua-or-muffin-my-search-for-the-best-computer-vision-api-cbda4d6b425d (consulté le 6.11.18).

Yearsley, L. (2017), "We Need to Talk About the Power of AI to Manipulate Humans." 5 juin. *MIT Technology Review.* Disponible à l'adresse suivante : https://www.technologyreview.com/s/608036/we-need-to-talk-about-the-power-of-ai-to-manipulate-humans/ (consulté le 7.11.18).

Yeung, K., (2011), 'Can we employ design-based regulation while avoiding brave new world?' *Law, Innovation and Technology*. 3(1) : 1-29.

Yeung, K. (2015), 'Design for Regulation.' In *Handbook of Ethics, Values and Technological Design,* établi par M. J. Van Den Hoven, P.E. Varmaas et I. van de Poel. Dordecht : Springer.

Yeung, K. (2016), '"Hypernudge": Big Data as a mode of regulation by design'. *Information, Communication & Society :* 1-19.

Yeung, K. (2017a), 'Algorithmic regulation: a critical interrogation', *Regulation & Governance.* doi 10.1111/rego.12158.

Yeung, K., (2017b), 'Blockchain, Transactional Security and the Promise of Automated Law Enforcement: The Withering of Freedom Under Law?' In *3THICS - The Reinvention of Ethics in a Digital Age,* établi par P. Otto and E. Graf, 132-146.

Yeung, K. (2018a), 'Five Fears About Mass Predictive Personalization in an Age of Surveillance Capitalism'. *International Data Privacy Law* 8: 258-269.

Yeung, K. et Weller, A. (2018b), 'How is 'transparency' understood by legal scholars and the machine learning community?' *Being Profiling. Cogitas Ergo Sum*. In E. Bayamlioglu, I. Baraliuc, L. W. Janssens et M. Hildebrandt (éd.) Amsterdam University Press.

Zalnieriute, M., et al. (2019), "The Rule of Law and Automation of Government Decision-Making." Modern Law Review 82: 397-424.

Zliobaite, I. (2015), A survey on measuring indirect discrimination in machine learning. Disponible à l'adresse suivante : *arXiv preprint arXiv:1511.00148* (consulté le 6.11.18).

Zook, M. et Grote M. H. (2017), 'The microgeographies of global finance: High-frequency trading and the construction of information inequality' *Environment and Planning A: Economy and Space* 49(1) : 121-140.

Zuboff, S., (2015), 'Big other: surveillance capitalism and the prospects of an information civilization', *Journal of Information Technology* 30(1) : 75-89.

Zweig, K. A., Wenzelburger, G. et Krafft, T. D (2018), 'On Chances and Risks of Security Related Algorithmic Decision-Making Systems'. *European Journal for Security Research*. 3(2) : 181-203.

******

Les technologies et services numériques de pointe, y compris les outils d'IA, sont extrêmement prometteurs, en particulier sous la forme d'une efficacité, d'une précision, d'une rapidité et d'une commodité accrues dans un large éventail de services. Mais l'émergence de ces technologies s'accompagne également d'une inquiétude croissante de l'opinion publique quant à leurs effets potentiellement préjudiciables pour les individus, pour les groupes vulnérables et pour la société en général.

Étant donné leur omniprésence dans la vie quotidienne, nous devons acquérir une meilleure compréhension de leur impact sur l'exercice des droits de l'homme et des libertés fondamentales, et nous devrions examiner avec soin l'attribution de la responsabilité en cas de conséquences défavorables. Si nous voulons prendre les droits de l'homme au sérieux dans une ère numérique mondialement connectée, nous ne pouvons permettre que la puissance de nos technologies et systèmes numériques avancés, et de ceux qui en tirent profit, s'accumule et s'exerce sans responsabilité.

Des mécanismes de gouvernance et d'application efficaces et démocratiquement légitimées doivent être mis en place pour s'assurer que la responsabilité des risques, des préjudices et des torts découlant de l'exploitation des technologies numériques avancées est dûment attribuée.

www.coe.int/freedomofexpression

PREMS 112019

FRA

www.coe.int

Le Conseil de l'Europe est la principale organisation de défense des droits de l'homme du continent. Il comprend 47 États membres, dont l'ensemble des membres de l'Union européenne. Tous les États membres du Conseil de l'Europe ont signé la Convention européenne des droits de l'homme, un traité visant à protéger les droits de l'homme, la démocratie et l'État de droit. La Cour européenne des droits de l'homme contrôle la mise en œuvre de la Convention dans les États membres.

COUNCIL OF EUROPE

CONSEIL DE L'EUROPE