Parliamentary Assembly
Assemblée parlementaire
https://pace.coe.int

COUNCIL OF EUROPE
CONSEIL DE L'EUROPE

Doc. 15151
29 septembre 2020

Prévenir les discriminations résultant de l'utilisation de l'intelligence artificielle

Rapport[1]
Commission sur l'égalité et la non-discrimination
Rapporteur: M. Christophe LACROIX, Belgique, Groupe des socialistes, démocrates et verts

Résumé

En permettant une généralisation massive des processus de prise de décision automatisés, l'intelligence artificielle (IA) permet des gains d'efficacité – mais en parallèle, elle est susceptible de perpétuer voire d'aggraver des discriminations. Il a déjà été démontré que l'utilisation de l'IA, tant par le secteur public que privé peut avoir un impact discriminatoire, et que les flux d'informations tendent à mettre en exergue les extrêmes et à fomenter la haine. L'utilisation de données biaisées, l'absence d'intégration dès le départ de la nécessité de protéger les droits humains, le manque de transparence des algorithmes et la difficulté à déterminer à qui revient la responsabilité de leur impact, ainsi que le manque de diversité au sein des équipes travaillant sur l'IA, sont autant de facteurs qui contribuent à ce phénomène.

Les États doivent agir dès aujourd'hui pour éviter que l'IA ait un impact discriminatoire au sein de nos sociétés, et œuvrer ensemble pour établir des normes internationales dans ce domaine.

Les parlements doivent par ailleurs contrôler activement l'utilisation des technologies basées sur l'IA et veiller à ce qu'elle soit soumise à un examen public. La législation anti-discrimination nationale devrait être revue et modifiée afin de garantir l'accès à un recours efficace aux victimes de discriminations causées par l'utilisation de l'IA, et les organes nationaux chargés de l'égalité devraient disposer de ressources leur permettant de traiter l'impact des technologies basées sur l'IA.

Le respect de l'égalité et de la non-discrimination doit être intégré dès la conception des systèmes basés sur l'IA et testé avant leur déploiement. Les secteurs public et privé devraient promouvoir activement la diversité ainsi que des approches interdisciplinaires dans les études et les professions liées aux technologies.

1. Renvoi en commission: Doc. 14808, renvoi 4434 du 12 avril 2019.

Sommaire **Page**

A. Projet de résolution[2]

1. L'intelligence artificielle (IA) est en train de transformer nos modes de vie. Permettant une généralisation massive des processus, elle est déjà utilisée par un large éventail d'entités privées et publiques, dans des domaines aussi divers que des procédures de sélection déterminant l'accès à l'emploi et à l'éducation, l'évaluation des droits d'un individu à des prestations sociales ou à un crédit, ou le ciblage des messages publicitaires et d'information.

2. Bon nombre d'utilisations de l'IA peuvent avoir un impact direct sur l'égalité d'accès aux droits fondamentaux, notamment le droit à la vie privée et la protection des données à caractère personnel ; l'accès à la justice et le droit à un procès équitable, particulièrement en ce qui concerne la présomption d'innocence et la charge de la preuve ; l'accès à l'emploi, à l'éducation, au logement et à la santé ; et l'accès aux services publics et à la protection sociale. Il s'avère que l'utilisation de l'IA est susceptible de provoquer ou exacerber la discrimination dans ces domaines, entraînant des refus d'accès aux droits qui touchent de manière disproportionnée certains groupes — souvent les femmes, les personnes appartenant aux minorités, ainsi que les personnes les plus vulnérables et les plus marginalisées. Son utilisation dans les flux d'informations a également été liée à la propagation de la haine en ligne qui a des retombées sur toutes les autres interactions sociales.

3. Le processus d'apprentissage par la machine utilisé pour produire des systèmes d'IA repose sur le recours à de vastes ensembles de données (*big data*) qui sont pour une grande part à caractère personnel. Des garanties effectives de la protection des données à caractère personnel revêtent donc une importance essentielle dans ce contexte. Parallèlement, les données sont biaisées par nature dans la mesure où elles reflètent les discriminations déjà présentes dans la société, ainsi que les préjugés des personnes chargées de leur collecte et de leur analyse. Le choix des données à utiliser et à ignorer dans un système basé sur l'IA, ainsi que l'absence de données sur des questions clés, l'utilisation de variables indirectes et les difficultés inhérentes à la quantification de concepts abstraits, peuvent également aboutir à des résultats discriminatoires. Les ensembles de données biaisés sont au cœur de nombreux cas de discrimination du fait de l'utilisation de l'IA et demeurent un problème majeur à résoudre dans ce domaine.

4. La conception et la finalité d'un système basé sur l'IA sont également cruciales. Les algorithmes optimisés en vue de l'efficacité, de la rentabilité ou d'autres objectifs, sans tenir dûment compte de la nécessité de garantir l'égalité et la non-discrimination, peuvent entraîner une discrimination directe ou indirecte (y compris une discrimination par association) fondée sur toute une série de motifs, dont le sexe, le genre, l'âge, l'origine nationale ou ethnique, la couleur, la langue, les convictions religieuses, l'orientation sexuelle, l'identité de genre, les caractéristiques sexuelles, l'origine sociale, l'état civil, le handicap ou l'état de santé. Il est donc particulièrement important, lorsque leur utilisation risque d'avoir une incidence sur l'accès à des droits fondamentaux, que les systèmes basés sur l'IA intègrent d'emblée le plein respect de l'égalité et de la non-discrimination dans leur conception et soient rigoureusement testés avant d'être déployés, mais aussi de façon régulière après leur déploiement, afin de s'assurer que ces droits soient garantis.

5. La complexité des systèmes d'IA et le fait qu'ils soient fréquemment développés par des entreprises privées et traités comme des éléments relevant de leur propriété intellectuelle peuvent conduire à de sérieux problèmes de transparence et de responsabilité concernant les décisions prises en utilisant ces systèmes. Ces caractéristiques peuvent rendre la discrimination extrêmement difficile à prouver et entraver l'accès à la justice, en particulier lorsque la charge de la preuve incombe à la victime et/ou lorsque la machine est supposée par défaut avoir pris la bonne décision, ce qui viole la présomption d'innocence.

6. Le manque de diversité au sein de nombreuses entreprises et professions technologiques augmente le risque que des systèmes d'IA soient développés sans tenir compte de leurs impacts potentiellement discriminatoires sur certains individus et groupes de la société. L'accès des femmes et des minorités aux professions scientifiques, technologiques, d'ingénierie et de mathématiques (STIM) doit être amélioré, et une véritable culture de respect de la diversité développée de toute urgence au sein de ces milieux professionnels. Le recours à des approches interdisciplinaires et interculturelles à tous les stades de la conception des systèmes d'IA contribuerait également à leur amélioration du point de vue du respect des principes d'égalité et de non-discrimination.

7. Enfin, des principes éthiques solides, clairs et universellement acceptés et applicables devraient sous-tendre le développement et le déploiement de tous les systèmes basés sur l'IA. L'Assemblée parlementaire considère que ces principes peuvent être regroupés dans les grandes catégories suivantes: la transparence, y compris l'accessibilité et l'explicabilité ; la justice et l'équité, y compris la non-discrimination ; la capacité

2. Projet de résolution adopté à l'unanimité par la commission le 11 septembre 2020.

d'imputer la responsabilité des décisions à une personne, y compris l'engagement de la responsabilité de l'intéressé et l'existence de voies de recours; la sûreté et la sécurité ; et le respect de la vie privée et la protection des données à caractère personnel.

8. L'Assemblée se félicite de ce que les acteurs publics et privés aient commencé à examiner et à élaborer des normes éthiques ainsi que des normes en matière de droits humains applicables à l'utilisation de l'IA. L'Assemblée se félicite en particulier de la Recommandation Rec/CM(2020)1 du Comité des Ministres sur les impacts des systèmes algorithmiques sur les droits de l'homme, complétée par ses lignes directrices sur le traitement des impacts sur les droits de l'homme des systèmes algorithmiques, ainsi que de la recommandation de la Commissaire aux droits de l'homme du Conseil de l'Europe intitulée «Décoder l'intelligence artificielle: 10 mesures pour protéger les droits de l'homme». L'Assemblée approuve les propositions générales énoncées dans ces textes qui s'appliquent également dans le domaine de l'égalité et de la non-discrimination.

9. L'Assemblée souligne que le législateur ne devrait pas invoquer la complexité de l'IA pour s'abstenir d'introduire des réglementations destinées à protéger et à promouvoir l'égalité et la non-discrimination dans ce domaine: les enjeux en matière de droits humains sont clairs et requièrent une action. Outre les principes éthiques, il faudrait élaborer des procédures, des outils et des méthodes de réglementation et d'audit des systèmes basés sur l'IA afin de garantir leur conformité aux normes internationales en matière de droits humains, et notamment aux droits à l'égalité et à la non-discrimination. Compte tenu de la forte dimension transnationale et internationale des technologies basées sur l'IA, des normes internationales apparaissent également nécessaires dans ce domaine.

10. Au vu de ces considérations, l'Assemblée invite les États membres:

10.1. à revoir leur législation anti-discrimination et à la modifier si nécessaire, de manière à ce qu'elle couvre tous les cas où une discrimination directe ou indirecte, y compris par association, pourrait être causée par l'utilisation de l'IA, et à ce que les plaignants aient pleinement accès à la justice ; à cet égard, à veiller tout particulièrement à garantir la présomption d'innocence et à éviter d'imposer aux victimes de discrimination une charge de la preuve disproportionnée ;

10.2. à élaborer une législation, des normes et des procédures nationales claires visant à garantir que tout système fondé sur l'IA respecte les droits à l'égalité et à la non-discrimination partout où son utilisation risquerait d'affecter la jouissance de ces droits ;

10.3. à veiller à ce que les organismes de promotion de l'égalité soient pleinement habilités à traiter les problèmes concernant les principes d'égalité et de non-discrimination résultant d'une utilisation de l'IA et soutenir les personnes qui portent plainte dans ce domaine, et à s'assurer qu'ils disposent de toutes les ressources nécessaires pour s'acquitter de ces tâches.

11. Afin que l'utilisation par les autorités publiques de technologies fondées sur l'IA soit soumise à un contrôle parlementaire et à un examen public adéquats, l'Assemblée invite les parlements nationaux:

11.1. à veiller à ce que l'utilisation de ces technologies fasse l'objet de débats parlementaires réguliers, et à ce qu'un cadre propice à ces débats soit mis en place;

11.2. à exiger du gouvernement d'informer à l'avance le parlement de l'utilisation de ces technologies;

11.3. à rendre obligatoire l'inscription systématique dans un registre public de chaque utilisation par les autorités de ces technologies.

12. Afin d'aborder les questions sous-jacentes de la diversité et de l'inclusion dans le domaine de l'IA, l'Assemblée invite en outre les États membres:

12.1. à promouvoir l'intégration des femmes, des jeunes filles et des personnes appartenant aux minorités dans les filières d'enseignement des STIM dès le plus jeune âge et jusqu'aux plus hauts niveaux, ainsi qu'à collaborer avec le secteur industriel pour faire en sorte que la diversité et l'intégration soient favorisées tout au long des parcours professionnels ;

12.2. à soutenir la recherche sur les biais concernant les données et les moyens de contrer efficacement leur impact dans les systèmes basés sur l'IA ;

12.3. à promouvoir la culture numérique et l'accès aux outils numériques par tous les membres de la société.

13. L'Assemblée invite toutes les entités, tant publiques que privées, travaillant sur et avec des systèmes basés sur l'IA, à veiller à ce que le respect de l'égalité et de la non-discrimination soit intégré dès le départ dans la conception des systèmes en cause et testé de manière adéquate avant leur déploiement, dès lors que ces systèmes pourraient avoir un impact sur l'exercice ou l'accès à des droits fondamentaux. À cette fin, elle invite ces entités à envisager de renforcer les capacités afin de mettre en place un cadre d'évaluation de l'impact sur les droits humains pour le développement et le déploiement de systèmes basés sur l'IA, tant par les entités privées que par les entités publiques. Elle encourage en outre le recours à des équipes interdisciplinaires et diversifiées à tous les stades du développement et du déploiement des systèmes basés sur l'IA.

14. Enfin, l'Assemblée invite les parlements nationaux à soutenir les travaux menés au niveau international, notamment par le biais du Comité ad hoc du Conseil de l'Europe sur l'intelligence artificielle (CAHAI), à veiller à ce que les normes en matière de protection des droits humains soient effectivement appliquées dans le domaine de l'IA, et à garantir le respect des principes d'égalité et de non-discrimination dans ce domaine.

B. Projet de recommandation[3]

1. L'Assemblée renvoie à sa Résolution … (2020) intitulée «Prévenir les discriminations résultant de l'utilisation de l'intelligence artificielle». Elle observe que cette résolution a été adoptée alors que des travaux étaient en cours au sein du Conseil de l'Europe, menés par le Comité ad hoc sur l'intelligence artificielle (CAHAI).

2. L'Assemblée rappelle que l'égalité et la non-discrimination sont des droits fondamentaux et que tous les États membres sont tenus de respecter ces droits conformément à la Convention européenne des droits de l'homme (STE n° 005), selon l'interprétation retenue par la jurisprudence de la Cour européenne des droits de l'homme, et à la Charte sociale européenne (STE n° 035), telle qu'interprétée par le Comité européen des droits sociaux.

3. L'Assemblée appelle par conséquent le Comité des Ministres à tenir compte de l'impact particulièrement grave que pourrait avoir le recours à l'intelligence artificielle sur la jouissance des droits à l'égalité et à la non-discrimination lorsqu'il évaluera la nécessité et la faisabilité d'un cadre juridique international applicable à l'intelligence artificielle.

3. Projet de recommandation adopté à l'unanimité par la commission le 11 septembre 2020.

C. Exposé des motifs par M. Christophe Lacroix, rapporteur

1. Introduction

1. L'intelligence artificielle (IA) est en train de transformer nos modes de vie. Des processus décisionnels automatisés sont employés dans le cadre de procédures de sélection de candidats à un emploi ou à un cursus d'enseignement supérieur ; ils servent à évaluer la capacité de remboursement d'une personne ou à déterminer son droit à des prestations sociales ; ils déterminent les informations mises à la disposition des internautes dans leurs fils d'actualité personnalisés ou dans les résultats de leurs moteurs de recherche ; ils définissent les personnes ciblées par les publicités à visée politique ou autre, ainsi que le contenu de ces messages.

2. On suppose fréquemment que toute décision prise par une machine est objective et exempte de préjugés. Cette présomption repose sur l'ignorance du rôle nécessairement joué par l'humain dans la conception des algorithmes en jeu, ainsi que des préjugés entachant les données utilisées pour les alimenter. Il est clairement établi aujourd'hui que l'utilisation de l'IA peut non seulement reproduire des résultats discriminatoires connus, mais aussi en générer de nouveaux.

3. L'émergence d'une IA non réglementée par un processus démocratique indépendant et souverain risque de conduire à une augmentation des cas de violation des droits humains, notamment en créant, perpétuant et même aggravant la discrimination et l'exclusion, que cela soit ou non son objectif explicite.

4. Les défis sont multiples et auront une incidence à la fois sur des individus et sur l'ensemble de nos sociétés. En outre, les technologies et les algorithmes utilisés ne connaissent pas de frontières. Par conséquent, les mesures prises au niveau national pour prévenir la discrimination dans ce domaine – bien qu'essentielles – ne sauraient suffire, d'où l'importance d'une réglementation internationale. Le moyen le plus sûr de garantir que les questions de droits humains en cause sont effectivement prises en considération passe par l'adoption d'une solide approche multilatérale commune.

5. Ce rapport vise à définir et à proposer un cadre international de base pour une IA axée sur l'être humain, le respect des principes éthiques et des droits humains, la non-discrimination, l'égalité et la solidarité. L'objectif général est d'assurer la garantie des droits de tous et en particulier ceux des personnes les plus à risque face aux effets potentiellement discriminatoires inhérents au recours à l'IA, qu'il s'agisse des femmes, des personnes appartenant aux minorités ethniques, linguistiques ou sexuelles, des travailleurs, des consommateurs, des enfants, des personnes âgées, des personnes handicapées ou d'autres personnes menacées d'exclusion.

2. Portée du rapport

6. Le déploiement de l'IA à grande échelle touche des aspects de plus en plus nombreux de la vie quotidienne des citoyens et peut avoir un impact important sur leur accès aux droits, notamment en introduisant un élément de discrimination. En notre qualité de décideurs politiques, nous devons donc assumer la responsabilité particulière de réfléchir à une éventuelle réglementation de ces systèmes, afin, entre autres, d'empêcher une telle discrimination.

7. Il est important de noter d'emblée que ce rapport a été rédigé en parallèle à l'élaboration par différentes commissions de l'Assemblée de plusieurs autres rapports ayant trait à l'IA. Ces rapports concernent: «Justice par algorithme – le rôle de l'intelligence artificielle dans les systèmes de police et de justice pénale»; «Les interfaces cerveau-machine: nouveaux droits ou nouveaux dangers pour les libertés fondamentales?»; «Aspects juridiques concernant les «'véhicules autonomes'»; «La nécessité d'une gouvernance démocratique de l'intelligence artificielle»; «Intelligence artificielle et marchés du travail: amis ou ennemis?»; «Intelligence artificielle et santé: défis médicaux, juridiques et éthiques à venir». De nombreuses questions concernant l'égalité et la non-discrimination se posent dans le cadre de ces rapports. Si j'évoque brièvement certaines de ces questions dans mon propre rapport, j'ai, par souci d'efficacité, volontairement concentré mon analyse sur d'autres problématiques liées à la prévention des discriminations résultant de l'utilisation de l'IA.

8. Pour les besoins du présent rapport, nous avons proposé une description du concept d'IA à l'annexe ci-jointe. J'aimerais faire remarquer à ce stade que le vocable « intelligence artificielle », dans son acception courante, est un terme nébuleux à la portée imprécise. On peut tenter de saisir son essence par analogie avec l'intelligence humaine considérée comme la somme de l'expérience et de l'acquis d'un individu alors que l'IA résulte d'une combinaison de « big data » (vastes ensembles de données préexistantes qui viennent remplacer l'expérience individuelle) et d'apprentissage automatique à partir de ces données[4]. Cet

apprentissage automatique ou « *machine learning* » consiste à élaborer un modèle mathématique basé sur des algorithmes et ayant recours à des techniques telles que celles des réseaux neuronaux, conçu pour faire apprendre, à partir d'un ensemble de données, un processus à une machine, afin que celle-ci devienne capable de faire des prédictions (justes) face à des situations inconnues[5].

9. Cette explication permet de mieux conceptualiser les mécanismes en jeu. Toutefois, l'élaboration d'un cadre de réglementation cohérent passe par la définition du seuil au-delà duquel tout système devrait être réglementé. D'un côté, définir l'IA comme embrassant tout code informatique reviendrait à considérer chaque logiciel de traitement de texte, voire chaque site web, comme relevant de l'IA et conférerait à cette notion une portée trop large (il serait alors difficile de concevoir un système réglementaire cohérent applicable à toutes les formes d'IA sur la base d'une définition aussi large). D'un autre côté, une définition juridique confinée aux techniques et applications déjà employées pourrait s'avérer impropre à couvrir de futurs développements, dans la mesure où les processus législatifs pêchent souvent par leur lenteur tandis que le secteur de l'IA évolue à une vitesse fulgurante[6].

10. S'agissant de prévenir les discriminations générées par l'IA, j'aimerais préciser que l'enjeu tient non pas tant à la nature de l'IA ou à son mode de fonctionnement qu'à la fonction pour laquelle ont été conçus des systèmes basés sur l'IA. Autrement dit, toute politique ou réglementation visant l'IA, notamment sous l'angle de la prévention de la discrimination générée par le recours à cette technologie, devrait englober les processus décisionnels automatisés et plus spécialement ceux découlant d'un apprentissage automatique[7].

11. L'objectif des processus décisionnels automatisés qui nous intéressent dans le présent rapport consiste à faire des choix et/ou à ordonner les choses d'une certaine manière. Quels candidats seront invités à un entretien de recrutement et quel niveau de rémunération sera proposé aux personnes sélectionnées ? Qui sera admis à étudier dans telle ou telle université ? À quel niveau sera fixée votre prestation sociale ? Quels articles seront proposés dans votre fil d'actualité, et dans quel ordre ?

12. Certes, tout processus de sélection, qu'il soit automatisé ou non, suppose des choix. L'objet du présent rapport est d'identifier les mesures que devraient adopter les pouvoirs publics et les autres acteurs compétents, afin de veiller à ce que les résultats obtenus par l'utilisation de l'IA dans des processus de décisions automatisés débouche sur des résultats équitables, et surtout, sans produire ni perpétuer des discriminations au sein de nos sociétés.

3. L'utilisation de l'IA engendre des résultats discriminatoires

> *« [L'attribution aux femmes, par le biais de l'IA, de plafonds de crédit inférieurs] revêt une importance pour celles souhaitant créer une entreprise dans un monde toujours convaincu, apparemment, que les personnes de sexe féminin sont moins vouées au succès ou solvables que leurs congénères masculins. Elle revêt une importance pour l'épouse qui essaie de se sortir d'une relation violente. Elle revêt une importance pour les minorités victimes de préjugés institutionnels. Elle revêt une importance pour une pléthore d'individus. Elle revêt donc une importance pour moi. »*
>
> *Jamie Heinemeier Hansson*[8]

13. Comme indiqué plus haut, les processus décisionnels automatisés sont déjà largement employés dans la vie quotidienne, dans des domaines aussi variés que l'administration de la justice, les processus d'admission dans l'enseignement supérieur, le recrutement, « l'optimisation » des heures de travail du personnel et l'évaluation de la capacité de remboursement ou du droit aux prestations sociales d'un individu.

14. De nombreux éléments (voir ci-dessous) montrent que ces processus sont souvent susceptibles de produire des résultats injustes et d'entraîner des discriminations fondées par exemple sur le genre, l'origine ethnique ou sociale ou la santé mentale. En notre qualité de législateurs, nous devons nous pencher sur ces violations des droits humains.

---

4. Autrement dit, si « expérience + apprentissage = intelligence humaine », alors « big data + apprentissage automatique = intelligence artificielle ». Evgeniou T., « AI Regulatory Challenges », présentation faite lors de la 1re réunion du Groupe parlementaire de l'OCDE sur l'intelligence artificielle (IA), Paris, 26 février 2020.
5. Frankle J., « Artificial Intelligence for Policymarkers », présentation faite lors de la 1re réunion du Groupe parlementaire de l'OCDE sur l'intelligence artificielle (IA), Paris, 26 février 2020.
6. À titre d'exemple, le nombre d'articles scientifiques dans ce domaine augmenterait de 30 % chaque année. Evgeniou T., « AI Regulatory Challenges », op. cit.
7. Frankle J., « Artificial Intelligence for Policymarkers », op. cit.
8. Heinemeier Hansson J., **«**About the Apple Card**»**, blogpost, 11 novembre 2019.

15. Dans les deux sections qui suivent consacrées à la mise en évidence de certains cas connus de discrimination déjà causés par l'utilisation de l'IA, je fais une distinction entre le recours à cette technologie dans le secteur privé et dans le secteur public. Différents droits peuvent en effet être en jeu dans ces domaines et, en cas de discrimination, les voies de recours potentielles des victimes varient également. J'examine ensuite dans une troisième section les flux d'informations gérés par l'IA, lesquels soulèvent des questions distinctes et peuvent notamment exacerber la discrimination en renforçant les stéréotypes et les préjugés.

3.1. Le secteur privé

16. Dans la mesure où elle peut accélérer de manière significative des processus qui prendraient plus de temps à l'homme, l'IA peut fournir des outils puissants en vue de la rationalisation des services fournis aux clients et de l'amélioration des performances commerciales d'une entreprise. Toutefois, l'utilisation de l'IA peut également avoir des effets très négatifs pour certains groupes de personnes.

17. Amazon, l'un des plus gros employeurs du secteur privé, a consacré des années à la mise au point d'un outil de recrutement basé sur l'IA. Cet outil a appris à examiner les candidatures sur la base des CV soumis à l'entreprise sur une période de 10 ans. En 2018, cependant, Amazon a décidé d'abandonner l'outil en cause en raison de la propension de celui-ci à discriminer au motif du genre[9]. Les grandes entreprises recourent également de plus en plus à des tests de personnalité automatisés lors des processus de recrutement, afin de présélectionner les candidats. Pourtant, d'aucuns ont constaté que ces tests étaient discriminatoires à l'égard des candidats souffrant de troubles mentaux et excluaient les intéressés sur la base d'évaluations laissant peu de place à la prédiction objective de leur potentiel professionnel[10].

18. La publicité ciblée en ligne, basée sur l'apprentissage automatique, est une autre source bien connue de discrimination dans le domaine de l'emploi. Des recherches indépendantes ont par exemple révélé que les femmes sont nettement moins nombreuses que les hommes à voir des publicités en ligne émanant d'entreprises proposant une aide à la recherche d'un emploi bien rémunéré[11]. Ce type de discrimination est difficile à détecter, à moins de procéder à des recherches approfondies, et presque impossible à contester pour les victimes (par ailleurs ignorantes des propositions faites aux membres de l'autre sexe).

19. Au-delà de l'emploi, le département des services financiers de New York a également été invité à enquêter sur les allégations selon lesquelles plusieurs grandes sociétés spécialisées dans l'aide à l'établissement de la déclaration d'impôts auraient eu recours aux fonctionnalités publicitaires de Google pour dissimuler des options supplémentaires de déclaration d'impôts aux personnes à faible revenu, lesquelles auraient pu notamment bénéficier d'un service d'aide gratuit[12]. Amnesty International a mis en lumière les pratiques de Facebook qui permettaient aux annonceurs dans le domaine du logement de cibler ou d'exclure certains groupes de personnes en fonction de leur origine ethnique ou de leur âge[13]. En mars 2019, le ministère américain du Logement et du Développement urbain a en outre accusé Facebook d'encourager, permettre et provoquer des discriminations – fondées sur la race, la couleur, la religion, le sexe, le statut familial, l'origine nationale et le handicap – par le biais de sa plateforme publicitaire. Cette dernière permettait aux annonceurs de dissimuler leurs publicités aux internautes de certains quartiers, ou de choisir de ne pas adresser de messages publicitaires en matière de logement aux internautes manifestant leur intérêt pour certains sujets comme « la mode hijab » ou la « culture hispanique »[14].

20. Des cas comme celui-ci montrent comment le comportement en ligne, par exemple l'entrée de certains critères de recherche par un internaute, permet d'inférer des informations privées sensibles sur l'intéressé comme son origine ethnique, ses croyances religieuses, son orientation sexuelle ou son identité de genre, ou bien ses centres d'intérêt, et de cibler la publicité d'une manière potentiellement discriminatoire. De telles pratiques soulèvent également de graves questions en matière de protection de la vie privée[15].

---

9. Dastin J., «Amazon scraps secret AI recruiting tool that showed bias against women», *Reuters*, 10 octobre 2018
10. O'Neil C., «Ineligible to Serve: Getting A Job», *Weapons of Math Destruction* [Algorithmes: la bombe à retardement], New York, Broadway Books, 2016.
11. Spice B., «Questioning the fairness of targeting ads online», Carnegie Mellon University, 7 juillet 2015.
12. NYDFS, «Governor Cuomo calls on DFS to investigate claims that advertisers use Facebook platform to engage in discrimination», communiqué de presse, 1er juillet 2019.
13. Amnesty International, «Surveillance Giants: How the business model of Google and Facebook threatens human rights», novembre 2019, p. 37-38.
14. Jee C., «Facebook has been charged over housing ads that discriminate on race, colour, and religion», *MIT Technology Review*, 29 mars 2019.
15. Wachter S., «Affinity profiling and discrimination by association in online behavioural advertising», *Berkeley Technology Law Journal*, 2020, 35(2) (à paraître).

21. Pour s'aligner sur les exigences de la législation anti-discrimination, Facebook a dû accepter de modifier ses algorithmes afin d'empêcher un tel ciblage à l'avenir[16]. Il a été toutefois démontré que les algorithmes basés sur l'apprentissage automatique provoquent une discrimination même lorsque les annonceurs ne ciblent pas délibérément la publicité en fonction de critères correspondant à des caractéristiques protégées par la législation anti-discrimination. Ainsi, selon certaines recherches, les offres d'emploi relatives à des emplois moins bien rémunérés (concierges, chauffeurs de taxi) sont majoritairement diffusées auprès des personnes appartenant aux minorités et celles d'enseignants et de secrétaires d'école maternelle auprès des femmes[17].

22. Il s'avère que des problèmes analogues surgissent dès lors que des systèmes basés sur l'IA sont utilisés pour évaluer la solvabilité d'un individu. Peu après le lancement de l'Apple Card en 2019, par exemple, des centaines de clients ont commencé à dénoncer des résultats sexistes. Les hommes se voyaient accorder des plafonds de crédit plusieurs fois supérieurs à ceux des femmes placées dans une situation identique, lesquelles ne disposaient en outre d'aucun moyen de contester la décision. Le personnel d'Apple, quand il était interpellé sur le sujet, ne pouvait que balbutier: « C'est la faute de l'algorithme.»[18].

23. Dans mon pays, la Belgique, certaines compagnies d'assurance-maladie cherchent actuellement à utiliser des « apps » pour smartphones dans le but de recueillir des informations sur l'état de santé des personnes couvertes par leurs polices d'assurance. Ce procédé soulève de graves problèmes sous l'angle du respect de la vie privée et pourrait également s'analyser en tant que source de discrimination fondée sur l'état de santé.

24. Ces exemples n'ont pas qu'une valeur anecdotique: ils démontrent clairement que l'utilisation de l'IA non seulement risque de générer des discriminations fondées sur différents motifs, mais en génère déjà dans de nombreux cas.

25. Qu'elles soient directes ou indirectes, ces discriminations violent les droits fondamentaux. À supposer qu'on leur permette de se produire ou de se prolonger sans contrôle, elles risquent de se perpétuer par le biais de l'utilisation de l'IA, voire, dans certains cas de s'aggraver. Les femmes, les personnes LGBTI, les membres de minorités ethniques ou religieuses, les personnes handicapées, entre autres, resteront enfermées dans des emplois moins bien rémunérés pour des motifs discriminatoires, avec moins de possibilités d'accès au crédit ou aux biens et services. Les personnes appartenant à des groupes considérés comme indésirables par les fournisseurs de logements, quant à elles, resteront exclues de certaines zones résidentielles, ce qui s'analyse non seulement en une discrimination individuelle, mais aussi en un facteur susceptible d'aggraver la ségrégation spatiale et sociale dans la société.

26. Si les exemples cités ci-dessus visent principalement les États-Unis, il est à noter que les entreprises en question revendiquent des millions voire des milliards de clients à travers l'ensemble des continents, dont l'Europe, et que leurs algorithmes sont susceptibles de produire des effets discriminatoires dans tous les États membres du Conseil de l'Europe. Les parlements ont le devoir d'aborder ces questions, en veillant à ce que les lois anti-discrimination soient suffisamment sévères pour protéger les individus et lutter contre la discrimination systématique, à ce que les entreprises qui recourent à des systèmes d'IA discriminatoires puissent être tenues de rendre des comptes et à ce que des recours efficaces soient mis en place.

3.2. Le secteur public[19]

27. Le secteur privé n'est pas seul à utiliser l'intelligence artificielle: celle-ci sert également aux pouvoirs publics, notamment dans le cadre de l'État-providence, puisque les services sociaux utilisent les technologies numériques pour déterminer si des individus sont éligibles à l'aide sociale, calculer le montant de leurs droits ou enquêter sur d'éventuelles fraudes ou erreurs. De nombreux exemples soulèvent de vives inquiétudes quant à l'utilisation de technologies numériques exploitant des données à caractère personnel d'une manière portant atteinte à la vie privée et/ou privant à tort des personnes de prestations d'aide sociale.

---

16. ACLU, «Facebook agrees to sweeping reforms to curb discriminatory ad targeting practices», 19 mars 2019.
17. Hao K., «Facebook's ad-serving algorithm discriminates by gender and race», *MIT Technology Review*, 5 avril 2019.
18. Voir notamment: Apple Newsroom, «Apple Card launches today for all US customers», 20 août 2019 ; la série de tweets publiés à ce sujet par David Heinemeier Hansson, https://twitter.com/dhh/status/1192540900393705474, et les nombreuses réponses reçues par ce dernier, y compris l'une émanant d'un des cofondateurs d'Apple (https://twitter.com/stevewoz/status/1193330241478901760); Natarajan S. et Nasiripour S., «Viral Tweet About Apple Card Leads to Goldman Sachs Probe», *bloomberg.com*, 9 novembre 2019.

28. Ainsi, aux Pays-Bas, le système SyRI est conçu pour détecter les risques de fraude en compilant et comparant des éléments d'information émanant de plusieurs bases de données gérées par l'administration. Dans le cadre d'un des projets pilotes lancés dans ce contexte, les données de 63 000 personnes percevant un complément de revenus et n'étant pas soupçonnées de la moindre irrégularité ont été comparées aux données relatives à l'utilisation de l'eau à usage domestique détenues par les entreprises publiques de distribution d'eau, afin d'établir automatiquement si tel ou tel abonné habite seul ou pas. Quarante-deux cas de fraude ont été détectés, soit un taux de réussite de seulement 0,07 %. Ce système soulève de sérieuses questions concernant le respect du droit à la vie privée et de la présomption d'innocence. Un tribunal néerlandais a récemment jugé que la législation régissant le SyRI ne prévoit pas de protection suffisante contre les ingérences dans la vie privée, car les mesures prises en l'occurrence pour prévenir et combattre la fraude dans l'intérêt du bien-être économique sont disproportionnées. En outre, le système manque de transparence et le fait qu'il cible les quartiers pauvres crée un risque de discrimination fondée sur le niveau socio-économique ou la qualité de migrant[20].

29. Toujours en Europe, la Cour suprême de Pologne a ordonné l'abrogation d'un système élaboré en 2014 pour classer automatiquement les bénéficiaires d'allocations de chômage en trois catégories – en fonction de données recueillies au moment de l'inscription, puis dans le cadre d'un entretien en ligne – ce qui n'a pas empêché l'Autriche d'adopter un système analogue en 2018. En Suède, un système créé pour collecter automatiquement en ligne des rapports d'activités fournis par des demandeurs d'emploi a été abandonné en 2018, en raison de la grande proportion d'erreurs (entre 10 et 15 %) détectée parmi les décisions prises automatiquement sur la base des informations ainsi recueillies.

30. Au Royaume-Uni, le système d'aide universel qui combine six prestations en une seule allocation est le premier service public à avoir été numérisé par défaut, un algorithme ayant été mis en place pour calculer chaque mois le montant de l'allocation sur la base d'informations recueillies en temps réel auprès des employeurs, des autorités fiscales et d'autres administrations. De nombreuses personnes ont perdu leur allocation pour la seule raison qu'elles ne disposent pas des compétences nécessaires pour remplir les nouveaux formulaires en ligne. Par ailleurs, l'allocation peut être diminuée d'office, sans explication ni avertissement, en fonction des résultats de l'algorithme. C'est en effet au système et non à l'allocataire que revient le bénéfice du doute. Pourtant, d'après les autorités elles-mêmes, environ 2 % des millions de calculs effectués (soit quelques dizaines de milliers de dossiers) produisent chaque mois des résultats erronés. La pandémie de covid-19 ne cessant de générer un nombre croissant de chômeurs dépendant de l'aide sociale, le risque est réel que davantage d'individus soient injustement privés de l'accès à la protection sociale[21].

31. Un système analogue (communément appelé « Robodebt ») mis en place en Australie produit des résultats particulièrement néfastes. Le couplage automatisé de données (remplaçant l'examen précédemment effectué manuellement par des fonctionnaires) a été introduit en 2016 pour identifier d'éventuelles divergences entre les données – relatives aux revenus des bénéficiaires de l'aide sociale – détenues par les autorités fiscales et les services sociaux, afin de détecter d'éventuels trop-payés ou des fraudes. Toute personne présentant une anomalie jugée suspecte par l'algorithme devait fournir la preuve du contraire via un formulaire en ligne, sous peine de voir ses prestations réduites ou supprimées. Or, l'algorithme compare les données annuelles des autorités fiscales aux revenus perçus toutes les deux semaines par les bénéficiaires de prestations sociales, alors que les revenus des intéressés varient souvent de manière considérable (contrats de travail de courte durée, travail saisonnier, etc.). Des milliers de personnes ont ainsi été privées de prestations à tort, et nombre d'entre elles placées dans l'incapacité de contester ces décisions (notification automatique envoyée à une ancienne adresse ; absence d'accès au portail permettant de transmettre les preuves exigées ; etc.). Dans de nombreux cas, des personnes se sont retrouvées du jour au lendemain gravement endettées et plusieurs cas de suicide ont été signalés. D'après certaines sources, les autorités auraient tenté de récupérer auprès des citoyens environ 600 millions de dollars australiens (soit 360 millions d'euros) en se fondant sur ce système qui génère souvent des erreurs (en plaçant malgré tout la charge de la preuve sur l'allocataire et dont les résultats sont très difficiles à contester).

---

19. Les exemples présentés dans le présent chapitre ont été exposés, lors de son audition par la commission le 4 décembre 2019, par M. Christiaan van Veen, directeur du projet sur l'État-providence numérique et les droits de l'homme, Center for Human Rights and Global Justice, New York University School of Law et Conseiller spécial sur les nouvelles technologies et les droits de l'homme auprès du Rapporteur Spécial de l'ONU sur les droits de l'homme et l'extrême pauvreté.

20. Voir également l'analyse de cette affaire dans Allen R., Q.C. et Masters D., Regulating for an equal AI: a new role for equality bodies, *Equinet*, Bruxelles, 2020, p. 107-109.

21. Lavelle D., «Coronavirus is shining a light on the wretched universal credit system», *The Guardian*, 3 avril 2020.

32. Ces exemples ne sont que la partie émergée de l'iceberg et les problèmes se multiplieront à mesure que les gouvernements chercheront à intensifier leur utilisation de la technologie au nom d'une plus grande efficacité. Le récent scandale survenu autour de l'ajustement automatisé des résultats de niveau A au Royaume-Uni dans le contexte de la pandémie de covid-19, ajustement ayant particulièrement touché les élèves des zones défavorisées, et d'autres problèmes du même type affectant le baccalauréat international ne sont que deux exemples supplémentaires parmi d'autres[22].

33. Trois préoccupations majeures se posent concernant la non-discrimination. On observe tout d'abord un manque d'examen préalable, de contrôle démocratique et de débat public sur ces questions. Ainsi, l'introduction du système SyRI a été fort peu débattue au Parlement néerlandais malgré les mises en garde de l'autorité pour la protection des données et d'autres acteurs. Par ailleurs, les demandes d'accès à l'information sont souvent rejetées en raison de la pléthore d'exceptions prévues par la loi ou de la méconnaissance par les autorités elles-mêmes de la technologie utilisée. Le fait que les personnes pauvres et marginalisées soient touchées de manière inégale passe dès lors souvent inaperçu. Deuxièmement, l'IA est souvent réputée forcément plus juste et plus précise que l'être humain. Si cette hypothèse peut se vérifier s'agissant de tâches très spécifiques, elle est beaucoup moins valable dès lors qu'il faut tenir compte d'un contexte plus large. Lorsque la technologie permet une généralisation massive des processus, mais entraîne parallèlement des erreurs à grande échelle, les personnes les moins à même de remettre le système en question (par exemple les personnes pauvres ou âgées, les migrants, les personnes handicapées) sont de nouveau touchées de manière disproportionnée. Cette situation revêt un caractère exceptionnellement grave lorsque l'IA est utilisée dans le contexte de l'État-providence. Enfin, les technologies numériques ciblent souvent à dessein les personnes pauvres et marginalisées, élargissant ainsi les possibilités d'une surveillance constante des intéressés par l'État au fil du temps.

34. Face à ces problématiques, les parlements doivent réaliser que les enjeux ne sont pas purement techniques, mais aussi profondément politiques. Le déploiement de systèmes basés sur l'IA est coûteux et leur utilisation – qui sert des objectifs politiques particuliers (décrits dans les exemples précités) au cœur des débats politisés sur l'État-providence – génère souvent un coût humain élevé. Par conséquent, il importe que la question du contrôle et de l'analyse des modalités d'utilisation de ces technologies soit régulièrement abordée lors des débats parlementaires. À cette fin, des règles pourraient être fixées, obligeant par exemple les gouvernements à informer à l'avance les parlements de l'utilisation de ces technologies, à inscrire chaque utilisation dans un registre public et à veiller à mettre en place un cadre propice à ces discussions. Les parlementaires n'ont pas besoin d'être des spécialistes de l'IA pour comprendre les problématiques politiques et sociétales sous-jacentes. En Australie et au Royaume-Uni par exemple, où les autorités s'appuient sur des calculs automatisés pour exiger des citoyens, bien souvent à tort, le remboursement de prestations sociales versées, la charge de la preuve a effectivement été renversée de sorte qu'il est devenu difficile, voire impossible, de contester les décisions prises. La procédure n'est pas équitable et les intéressés ne peuvent pas exercer leur droit à un recours. Aux Pays-Bas, la présomption d'innocence et le droit au respect de la vie privée de dizaines de milliers de personnes ont été bafoués alors que seulement quelques cas de fraude à l'aide sociale ont été avérés. En l'occurrence, force est également de constater une violation du droit à l'assistance sociale puisque, en raison du recours à des processus automatisés, des personnes ont eu plus de mal à bénéficier de prestations auxquelles elles ont droit. Dans tous ces cas, ce sont les personnes les plus défavorisées de la société qui sont le plus durement touchées.

3.3. Le cas particulier des flux d'informations

«Le problème avec internet... c'est qu'il favorise les extrêmes. Imaginons que vous conduisez sur la route et que vous êtes témoin d'un accident de voiture. Bien sûr que vous regardez. Tout le monde regarde. Internet déduit de ce comportement que tout le monde réclame des accidents de voiture, donc il essaie de répondre à cette demande.»

Evan Williams, fondateur de Twitter[23]

35. Je tiens par ailleurs à attirer l'attention sur un autre contexte dans lequel l'utilisation de l'IA risque d'exacerber les discriminations ; je veux parler de l'encouragement des extrémismes et de la haine. Il a été démontré que les algorithmes de certains sites web, notamment les sites de réseaux sociaux, recommandent automatiquement à leurs utilisateurs des points de vue de plus en plus radicalisés.

---

22. Voir notamment Hill A. et Davies C., «A-level results day 2020 live: 39.1% of pupils' grades in England downgraded – as it happened», *The Guardian*, 13 août 2020; Evgeniou T. et autres, «What happens when AI is used to set grades?», *Harvard Business Review*, 13 août 2020.
23. Lewis P., «Fiction is outperforming reality: how YouTube's algorithm distorts truth», *The Guardian*, 2 février 2018.

36. Il est fréquemment souligné que la diffusion en ligne d'opinions radicales ou extrêmes (ainsi que de fausses informations – sujet qui tombe toutefois en dehors du champ couvert par le présent rapport) pourrait avoir eu un effet décisif sur le résultat de l'élection présidentielle américaine de 2016. Ancien employé de Google, Guillaume Chaslot a conçu un algorithme permettant de déterminer si les vidéos automatiquement recommandées par l'algorithme de YouTube aux personnes ayant regardé des vidéos contenant les mots « Trump » ou « Clinton » présentaient un parti pris. Les résultats ont permis de constater non seulement qu'un grand nombre de vidéos recommandées exprimaient des avis extrêmes, mais aussi que, quel que soit le nom du candidat initialement entré dans la barre de recherche, l'algorithme de YouTube était en fin de compte beaucoup plus susceptible de recommander des vidéos favorables à Trump qu'à Clinton (quand il ne s'agissait pas tout simplement de vidéos complotistes foncièrement hostiles à Clinton)[24].

37. La tendance de l'algorithme de YouTube à recommander des vidéos extrêmes ne se limite pas à la politique. Ainsi, des expériences menées dans d'autres domaines ont abouti à des résultats analogues. À titre d'exemple, des vidéos relatives au régime végétarien mènent à des vidéos sur le véganisme, des vidéos relatives au jogging débouchent sur des vidéos consacrés aux ultramarathons, des vidéos relatives au vaccin contre la grippe mènent à des vidéos complotistes anti-vaccination, etc.[25]. Des préoccupations du même ordre ont été soulevées concernant les flux d'informations de Facebook[26].

38. Les algorithmes utilisés sur les réseaux sociaux et les sites web des journaux sont souvent optimisés en vue de la rentabilité. Ils favorisent donc par défaut les éléments susceptibles d'attirer un nombre élevé de clics ou « engagements ». Les internautes sont incités à s'attarder sur les sites (et par conséquent à visionner un plus grand nombre de publicités) qui flattent leur tendance naturelle à se diriger vers ce qui « titille ».

39. Je tiens à souligner que cet engrenage n'est pas inéluctable, pas plus que l'enfermement de chaque utilisateur des sites d'information en ligne ou des réseaux sociaux dans des « bulles » dont les « colocataires » partageraient tous un seul et même point de vue. Beaucoup dépend en effet des objectifs définis pour l'algorithme mis en place: autrement dit, que doit-il optimiser ? Des objectifs autres que l'accroissement direct de la rentabilité peuvent tout aussi bien être visés, comme par exemple présenter aux internautes la plus grande diversité de points de vue. C'est le choix opéré notamment par certains journaux scandinaves dans l'élaboration des algorithmes employés sur leurs éditions en ligne[27].

4. Données

« Ordures à l'entrée, ordures à la sortie »[28]

40. Les données jouent un rôle capital dans le domaine de l'IA. D'une part, de vastes ensembles de données (appelés « *big data* ») sont nécessaires pour « entraîner le système » et affiner l'apprentissage automatique dans le but d'élaborer des systèmes d'IA complexes. Ces données concernent généralement des individus et sont souvent générées par les intéressés eux-mêmes (lorsqu'ils choisissent par exemple de cliquer sur un lien dans un fil d'actualité ou qu'ils remplissent un formulaire en ligne).

41. Malgré les protections introduites, au moins dans les pays de l'Union européenne, par le règlement général sur la protection des données (RGPD)[29], les utilisateurs de systèmes d'information ne sont pas toujours conscients de la collecte de ces données ni de l'exploitation qui peut en être faite par la suite. Une telle ignorance est de nature à soulever de multiples problèmes au niveau du respect de la vie privée: un dénominateur commun à bon nombre d'utilisations de l'IA. Aux fins du présent rapport, je me contenterai de signaler que les utilisateurs de systèmes informatiques n'ont pas tous le même niveau de connaissances dans ce domaine. Ainsi, tout le monde n'est pas sur un pied d'égalité en ce qui concerne la compréhension des notions de collecte en ligne et de protection des données à caractère personnel.

---

24. Cité dans Streitfeld D., «'The Internet is broken': @ev is trying to salvage it», *New York Times*, 20 mai 2017.
25. Tufekci Z., «YouTube, the Great Radicalizer », *New York Times*, 10 mars 2018.
26. Evans J., «Facebook isn't free speech, it's algorithmic amplification optimised for outrage», *Techcrunch.com*, 20 octobre 2019.
27. Bucher T., *If…Then: Algorithmic Power and Politics*, New York: Oxford University Press, 2018, p. 142.
28. «Garbage in, garbage out» [GIGO], aphorisme couramment employé par les informaticiens, cité depuis au moins 1957: Mellin W.D., «Work with new electronic 'brains' opens field for army math experts», *The Hammond Times* 10 (1957), p. 66.
29. Règlement UE 2016/679 du Parlement européen et du Conseil du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant la directive 95/46/CE (règlement général sur la protection des données).

42. D'autre part, les données elles-mêmes ne sont jamais neutres, mais reflètent toujours le lieu et l'époque de leur collecte[30]. Les partis pris sont inhérents aux données générées par l'homme et sont à la fois source de stéréotypes et de préjugés et le résultat de ceux-ci. Les préjugés qui prévalent en un moment et en un endroit précis, ainsi que dans l'esprit de ceux qui conçoivent et mènent les exercices de collecte de données, se reflètent dans les données collectées.

43. Le recours à des ensembles de données empreints de parti pris – ou reflétant un parti pris, un préjugé ou une pratique discriminatoire du passé – constitue l'une des principales causes de la discrimination en IA. Si, dans le passé, certains secteurs ont employé moins de femmes et/ou de personnes issues de minorités ethniques (ou les ont employées pour un salaire moindre), si les personnes appartenant à certains groupes se sont vu refuser des emprunts ou si les personnes appartenant aux minorités ont davantage cliqué sur des publicités les incitant à louer plutôt qu'à acheter un logement, tout système d'IA qui fonde ses décisions d'optimisation sur la reconnaissance et la reproduction de modèles historiques ne fera qu'enraciner la discrimination[31]. Pour corriger ce défaut, il faut non seulement avoir conscience des schémas historiques, mais aussi prendre des décisions volontaristes au niveau de la conception, un domaine que j'explore plus en détail ci-dessous.

44. Dans certains cas, les préjugés peuvent être faciles à corriger. S'agissant par exemple d'un logiciel de reconnaissance faciale dont les performances sont moins précises en fonction de la couleur de la peau, l'utilisation d'un éventail plus large de photographies pour entraîner la machine peut remédier à certains problèmes. Dans ce cas précis, l'élargissement de l'ensemble de données peut s'avérer relativement simple, car une vaste gamme de photos de personnes d'origines ethniques différentes est disponible gratuitement sur internet. Néanmoins, la facilité avec laquelle il est possible de résoudre un tel problème dépend aussi de l'exploitation des données. Dans l'affaire tristement célèbre de Google Photos dont la version initiale étiquetait automatiquement les photos d'Afro-Américains comme « gorilles »[32], il a été facile de corriger ce préjugé raciste. Par contre, lorsque des techniques de reconnaissance faciale basées sur l'IA ne sont pas utilisées simplement pour catégoriser des images comme appartenant à certains groupes, mais sont exploitées dans le système de justice pénale, par exemple pour identifier des individus spécifiques dans des situations où leur liberté individuelle peut être en jeu, et lorsque ces systèmes fonctionnent de manière nettement moins précise pour les personnes ayant la peau foncée, des solutions beaucoup plus complexes peuvent s'avérer nécessaires pour remédier aux conséquences profondes de ces défauts sur les droits individuels fondamentaux.

45. Certaines données n'ayant jamais été mesurées historiquement, plusieurs groupes de personnes risquent de ne pas apparaître dans les ensembles de données disponibles. Les personnes n'étant identifiées ni comme homme ni comme femme (notamment certaines personnes intersexuées) doivent non seulement remplir des formulaires en ligne hostiles où elles sont tenues de cocher une case « homme » ou « femme », mais souffrent de l'impossibilité de voir leur situation spécifique mesurée et d'identifier ou d'empêcher les discriminations à leur encontre. Dans le domaine des tests médicaux, les femmes ont également été historiquement exclues des expériences médicales, de sorte que leur santé et les réactions de leur corps à certains médicaments sont moins bien comprises que celles des hommes. Certes, ces problèmes datent d'avant l'utilisation de systèmes basés sur l'IA. Ils signifient pourtant que l'utilisation desdits systèmes entraînés à l'aide de données historiques aura tendance à reproduire et à enraciner les discriminations existantes[33].

46. Par ailleurs, les données, si elles constituent une certaine représentation de la réalité, sont souvent réductrices et débouchent sur des approximations plus ou moins grossières de la réalité qu'elles sont censées représenter[34]. De nombreux États refusent par exemple d'autoriser la collecte de données ethniques (souvent au motif que l'utilisation abusive de ces données dans le passé montre qu'elles ne devraient plus jamais être collectées). Pour pallier cette situation, on a recours à des variables indirectes telles que le pays de naissance des individus, de leurs parents ou de leurs grands-parents. Ces indicateurs, s'ils permettent de

---

30. Evgeniou T., **«**AI Regulatory Challenges**»**, op. cit.
31. Hao K., «Facebook's ad-serving algorithm discriminates by gender and race», *MIT Technology Review*, 5 avril 2019.
32. CNIL, Comment permettre à l'homme de garder la main? Les enjeux éthiques des algorithmes et de l'intelligence artificielle, décembre 2017, pp. 31-32.
33. Wachter-Boettcher S., *Technically Wrong: Sexist Apps, Biased Algorithms, and Other Threats of Toxic Tech*, New York, Londres, W. W. Norton, 2017; Jackson G., «Centuries of exclusion has meant women's diseases are often missed, misdiagnosed or remain a total mystery», *The Guardian*, 14 novembre 2019. L'utilisation de l'IA dans le domaine médical fait l'objet d'un autre rapport de l'Assemblée.
34. O'Neil C., «Civilian casualties: Justice in the Age of Big Data», *Weapons of Math Destruction*, op. cit., p. 95; Bucher T., *If…Then: Algorithmic Power and Politics*, op. cit., pp. 8-12.

saisir de nombreuses personnes appartenant à des minorités ethniques, n'en omettent pas moins beaucoup d'autres, ne serait-ce que parce que le fait que leur famille vit depuis des générations dans le pays n'élimine pas les discriminations fondées sur la couleur de la peau ou sur la langue[35].

47. Dans d'autres cas, les informations pertinentes pour un algorithme peuvent se révéler extrêmement difficiles à calculer. Ceci est particulièrement vrai des concepts abstraits (l'équité, par exemple, laquelle revêt une importance cruciale dans le domaine judiciaire) à la différence d'éléments concrets et quantifiables comme le nombre d'agressions à l'arme blanche recensées dans une zone particulière au cours d'une période donnée. Je n'examinerai pas en détail dans le présent rapport la question de l'administration de la justice au moyen d'algorithmes, laquelle fait actuellement l'objet des travaux d'une autre commission. Je tiens toutefois à signaler les graves discriminations susceptibles de se produire lorsqu'un algorithme privilégie l'efficacité (en s'appuyant sur des éléments faciles à quantifier) par rapport à l'équité (en tenant moins compte des implications plus larges des effets d'un algorithme sur la société dans son ensemble). Le tristement célèbre programme COMPAS utilisé par certains États américains pour évaluer les risques de récidive (ceci afin d'aider les juges à décider d'imposer ou non des peines de prison) fournit un exemple regrettable en la matière[36].

48. Globalement, la capacité des systèmes basés sur l'IA à prévenir la discrimination dépend fortement des données utilisées pour les entraîner. Les préjugés et discriminations historiques, l'absence de données sur les questions clés, l'utilisation de variables indirectes et les difficultés inhérentes à la quantification de concepts abstraits sont autant de questions devant être abordées et résolues efficacement en cas de déploiement d'un système basé sur l'IA, dès lors que son utilisation est susceptible d'avoir un impact sur les droits humains. En notre qualité de législateurs, nous devons prendre conscience des enjeux cruciaux en matière de droits humains dans ce domaine et concevoir des moyens de garantir la protection des citoyens contre ce type de discriminations.

5. Conception et finalité

> *« Le roi Midas déclara 'Je veux que tout ce que je touche se transforme en or' et son vœu fut pleinement exaucé. C'était le but qu'il mit dans la machine, pour ainsi dire, et sa nourriture, sa boisson et ses proches s'étant effectivement transformés en or, il finit ses jours misérable et affamé. »*
>
> *Stuart Russell, chercheur en IA*[37]

49. Les algorithmes sont conçus pour travailler de manière ciblée à l'atteinte d'un objectif spécifique identifié par les programmeurs. Un modèle mathématique est élaboré afin de permettre d'apprendre un processus à une machine à partir d'un ensemble de données, afin qu'elle soit en mesure de faire des prédictions justes face à des situations inconnues. Il est donc crucial d'affecter à l'algorithme un objectif idoine, car toutes les décisions à venir en matière de conception, ainsi que les résultats produits au final par le système basé sur l'IA, dépendront de ce choix initial. Des objectifs ou des choix politiques inconsidérés entraîneront des résultats indésirables et inéquitables.

50. A titre d'exemple, les processus automatisés de recrutement en usage dans de grandes entreprises sont fréquemment optimisés afin d'accroître leur « efficience », c'est-à-dire de réduire autant que faire se peut le nombre de candidats à interviewer en présentiel ou dont le dossier de candidature devra être examiné par un être humain. Or, un algorithme conçu pour identifier des candidats qui semblent correspondre à la culture ambiante dans l'entreprise ne contribuera probablement pas à renforcer la diversité au sein de cette organisation ou à améliorer ladite culture[38]. La recherche de « l'efficience » est également au cœur de nombreuses utilisations problématiques de l'IA dans le secteur public, notamment dans le contexte de l'État-providence tel que décrit plus haut.

51. Lorsque l'objectif d'un modèle d'apprentissage automatique s'accorde mal avec la nécessité d'éviter la discrimination, les résultats qu'il produit perpétueront ou exacerberont à nouveau celle-ci. L'outil publicitaire de Facebook, mentionné précédemment, proposait aux annonceurs une série d'objectifs d'optimisation: le

---

35. Les problèmes associés à la collecte de données ethniques sont régulièrement étudiés par la Commission européenne contre le racisme et l'intolérance (ECRI) ; pour un exemple récent, voir Cahuc P. et Valfort M.-A., «La pandémie de covid-19 va renforcer les disparités raciales et ethniques», *Le Monde*, 19 août 2020.
36. O'Neil C., «Civilian casualties: Justice in the Age of Big Data», *Weapons of Math Destruction*, op. cit., pp. 146-150 et 164-165.
37. Russell S., «Three principles for creating safer AI», *TED Talk*, 2017.
38. Lauret J., «Amazon's sexist AI recruiting tool: how did it go so wrong? », *becominghuman.ai*, 16 août 2019.

nombre de fois qu'une publicité a été vue, le taux d'engagement qu'elle a généré, le montant des ventes auxquelles elle a conduit, etc. Si le fait de cibler davantage la population blanche concernant les offres d'achat d'habitation suscitait des engagements plus nombreux, l'algorithme agissait dans ce sens, discriminant ainsi les utilisateurs noirs. Ceci, pour la simple et unique raison que l'algorithme avait été optimisé en vue d'atteindre des objectifs commerciaux sans tenir compte de la nécessité de respecter les droits individuels fondamentaux tels que l'égalité d'accès au logement[39].

52. S'agissant des flux d'informations, le fait que des algorithmes soient conçus pour « optimiser l'engagement » semble également poser un problème de taille. Les plateformes « gratuites » telles que Facebook et YouTube, mais aussi les médias en ligne financés par la publicité ont intérêt à augmenter le temps que les utilisateurs passent sur leurs sites de manière à les exposer à davantage de messages publicitaires. Si les données montrent que les utilisateurs ont tendance à être plus attirés par les contenus extrêmes, les algorithmes favoriseront lesdits contenus. Même si ces plateformes font généralement valoir qu'elles « se contentent de montrer aux utilisateurs ce qu'ils veulent voir », il n'en demeure pas moins que l'objectif des algorithmes est de générer le plus de recettes publicitaires possibles pour les entreprises concernées[40]. Dans ce contexte, la question de savoir ce qui est réellement informatif ou digne d'intérêt est laissée de côté, tandis que les contenus sexistes, racistes, antisémites, islamophobes, vecteurs d'antitsiganisme, homophobes, transphobes, et autres propos incitant à la haine sont souvent promus.

53. Avant même de déterminer l'objectif spécifique de l'optimisation d'un algorithme, il convient de se poser des questions fondamentales sur la finalité d'un système basé sur l'IA. Pour tirer un exemple du monde « réel », les programmes d'interpellation et de perquisition conçus pour détecter les délits mineurs exacerbent les préjugés du système de justice pénale lorsqu'ils se contentent de cibler les quartiers pauvres à la recherche de certains types de comportements criminels, tout en laissant de côté les quartiers riches (où pourtant la criminalité en col blanc sévit et où la « violence domestique » est aussi susceptible de se produire qu'ailleurs). Les données collectées dans le cadre de stratégies de surveillance policière biaisées dans le but d'alimenter des systèmes basés sur l'AI révèlent automatiquement un taux d'activité criminelle plus élevé dans les zones ciblées, ce qui entraîne ensuite une surveillance policière accrue dans les mêmes zones et crée une « boucle de rétroaction » ou cercle vicieux profondément discriminatoire qui nuit aux habitants des quartiers pauvres (souvent issus de minorités ethniques) et ne rend pas compte des activités méritant tout autant l'attention de la police dans d'autres quartiers. Ces « boucles de rétroaction » se produisent parce que les résultats algorithmiques basés sur des ensembles de données faussés tendent à confirmer des pratiques discriminatoires et parce qu'on ne procède à aucune évaluation de ce qui est tout simplement absent desdits ensembles[41].

54. Il convient de souligner qu'à l'instar de n'importe quelle procédure de sélection un processus décisionnel automatisé ne saurait être totalement neutre, car il sera toujours le résultat de choix (au stade de la conception) reflétant le point de vue des développeurs sur la manière dont les choses devraient être ordonnées. Toutefois, en raison de la capacité de ces processus à prendre un grand nombre de décisions en un temps très court, les conséquences du déploiement de systèmes discriminatoires basés sur l'IA peuvent s'avérer dramatiques. En notre qualité de législateurs, il nous appartient de trouver des moyens de garantir que les choix en matière de conception opérés dans le cadre de l'élaboration et du déploiement de processus décisionnels automatisés intègrent systématiquement la nécessité de protéger les droits humains et de prévenir la discrimination.

6. Diversité

> *« L'initiative 'Black girls need to learn how to code' [Les filles noires doivent apprendre à coder] est un prétexte pour ne pas aborder le problème de la marginalisation persistante des femmes noires dans la Silicon Valley. »*
>
> *Safiya Umoja Noble*[42]

39. Hao K., «Facebook's ad-serving algorithm discriminates by gender and race», *MIT Technology Review*, 5 avril 2019.
40. Tufekci Z., «YouTube, the Great Radicalizer », *New York Times*, 10 mars 2018 ; Evans J., «Facebook isn't free speech, it's algorithmic amplification optimised for outrage», *Techcrunch.com*, 20 octobre 2019.
41. O'Neil C., *Weapons of Math Destruction*, op. cit., pp. 146-150 et 164-165.
42. Noble S. U., *Algorithms of Oppression: How Search Engines Reinforce Racism*, New York, New York University Press, 2018, p. 66.

55. Les algorithmes expriment les valeurs, croyances et convictions de leurs concepteurs. De même, la capacité des processus décisionnels automatisés à inclure et à refléter la diversité présente dans nos sociétés est tributaire de ces facteurs[43].

56. L'absence de diversité que l'on constate habituellement au sein des entreprises de haute technologie se voyant confier la conception d'algorithmes soulève de sérieux problèmes à cet égard. La sous-représentation des femmes dans les études et professions relevant de la science, de la technologie, de l'ingénierie et des mathématiques (STIM) a déjà été reconnue par l'Assemblée, qui a invité les États à prendre des mesures pour encourager les femmes et les jeunes filles à suivre des études secondaires et supérieures dans ces disciplines[44]. Toutefois, les problèmes dépassent les simples « inégalités » entre sexes. La chercheuse Kate Crawford a mis en évidence l'endogamie sociale, raciale et de genre caractéristique des milieux où se recrutent ceux qui entraînent aujourd'hui les systèmes d'intelligence artificielle[45].

57. La diversité de la main-d'œuvre n'est cependant pas un simple problème de « filière ». La discrimination sur le lieu de travail entraîne également des taux de rotation élevés chez les femmes et les membres de minorités. La culture de l'entre-soi masculin, du harcèlement sexuel et des préjugés sexistes sur le lieu de travail contribue à ce que, dans les entreprises de haute technologie, les femmes quittent leur carrière à mi-parcours deux fois plus souvent que les hommes. La persistance de la perception raciste des personnes de couleur comme étant « non techniques » conduit aussi à la marginalisation des intéressés dans le domaine de la haute technologie[46].

58. Le manque de diversité dans les entreprises de haute technologie n'est pas seulement discriminatoire en soi, mais débouche directement sur une IA discriminatoire. Safiya Umoja Noble, dans son travail sur les moteurs de recherche, a étudié de manière approfondie la façon dont les groupes dominants sont capables de classer et d'organiser les représentations des autres de manière à reproduire et à exacerber les préjugés racistes[47]. La conception des formulaires en ligne tend également à refléter uniquement les biographies les plus banales, excluant ainsi toute personne échappant à la norme[48]. L'utilisation massive de personnages féminins pour incarner les agents conversationnels automatisés («*chatbots*») lesquels sont le plus souvent cantonnés dans un rôle subalterne, perpétue également des stéréotypes sexistes néfastes[49].

59. Pour résoudre ces problèmes, il est essentiel – mais pas suffisant – d'accroître la diversité de la main-d'œuvre travaillant au développement des systèmes d'IA ; il conviendrait notamment de prendre des mesures décisives pour améliorer l'accès des femmes et des minorités aux professions STIM. Il faudrait aussi impérativement veiller à ce que tous les étudiants de ces disciplines, lesquels conçoivent des technologies pour les gens, soient formés et instruits sur l'histoire des personnes marginalisées. L'adoption d'une approche interdisciplinaire – impliquant dès le départ le recours non seulement à des spécialistes de la technologie, mais aussi à des spécialistes des sciences sociales, des sciences humaines et du droit – en matière de conception de systèmes d'IA contribuerait aussi grandement à prévenir les discriminations causées par l'utilisation de ces systèmes.

7. Le syndrome de la « boîte noire »: Transparence, explicabilité et imputabilité

> *« On ne m'a donné aucune explication. Pas moyen de plaider ma cause. »*
>
> *Jamie Heinemeier Hansson*[50]

60. Les victimes de discrimination due à l'utilisation de l'IA vivent souvent une expérience semblable, à savoir que même lorsqu'elles sont en mesure de prouver qu'une discrimination a bien eu lieu, elles ne parviennent pas à obtenir la moindre explication sur la raison pour laquelle cette discrimination s'est produite. Les étudiants français par exemple ont été confrontés à des situations de ce type dès l'entrée en vigueur en

---

43. Bucher T., *If…Then: Algorithmic Power and Politics*, op. cit., p. 158; Wachter-Boettcher S., *Technically Wrong: Sexist Apps, Biased Algorithms, and Other Threats of Toxic Tech*, op. cit.
44. Résolution 2235 (2018) «L'autonomisation des femmes dans l'économie».
45. Crawford K., «Artificial intelligence's white guy problem», *New York Times*, 25 juin 2016.
46. Presentation during the Committee's hearing on 12 September 2019 by Alice Coucke, Snips AI; Noble S. U., *Algorithms of Oppression: How Search Engines Reinforce Racism*, op. cit.
47. Noble S. U., *ibid.,* p. 86.
48. Wachter-Boettcher S., *Technically Wrong: Sexist Apps, Biased Algorithms, and Other Threats of Toxic Tech*, New York, London, W. W. Norton, 2017.
49. «Think Piece 2: The rise of gendered AI and its troubling repercussions» in I'd Blush If I Could: Closing gender divides in digital skills through education, EQUALS and UNESCO, 2019.
50. Heinemeier Hansson J., «About the Apple Card», blogpost, 11 novembre 2019.

2018 du nouveau système en ligne Parcoursup censé gérer les demandes d'entrée à l'université[51]. Comme dans le cas de l'Apple Card évoqué plus haut, il est fréquent que les concepteurs d'un algorithme eux-mêmes soient incapables d'expliquer exactement pourquoi celui-ci a produit un résultat ou un ensemble de résultats donné. Les éléments spécifiques pris en compte par un algorithme donné, ainsi que le poids accordé à chacun d'entre eux, sont rarement connus et les entreprises privées qui ont investi dans le développement de tels algorithmes répugnent généralement à les révéler, car elles considèrent ces derniers comme un actif de grande valeur protégé par le droit de propriété intellectuelle.

61. Cette absence de transparence et d'explicabilité est souvent désignée sous le terme de « syndrome de la boîte noire ». Elle peut rendre les décisions prises en recourant à l'IA extrêmement difficiles à contester pour les individus. Elle crée également des obstacles pour les organismes nationaux de promotion de l'égalité qui cherchent à aider les plaignants à porter des affaires devant les tribunaux dans ce domaine. Pourtant, l'utilisation de l'IA peut avoir de graves conséquences pour certains groupes ou individus en les discriminant directement ou indirectement et peut aussi exacerber les préjugés et la marginalisation. Il est crucial de prévoir la mise en place de recours efficaces et accessibles permettant de contrer les discriminations lorsqu'elles se produisent et de veiller à ce que leurs auteurs puissent en être tenus responsables.

62. Le chapitre suivant décrit certains éléments qui pourraient utilement être incorporés aux normes juridiques, de manière à élaborer un cadre solide à la fois pour prévenir les discriminations causées par l'utilisation de systèmes d'IA et pour traiter les cas lorsqu'ils se présentent. Sur ce point, j'aimerais souligner que lorsque les entreprises ou les pouvoirs publics sont confrontés à des preuves de discrimination, ils trouvent des moyens de résoudre ces problèmes, que ce soit sous la pression de l'opinion publique ou pour donner effet à une décision de justice. Il nous appartient cependant de veiller à ce que de telles mesures soient prises le plus vite possible.

8. Nécessité d'un ensemble commun de normes juridiques fondées sur les principes éthiques et les droits fondamentaux reconnus

63. Les discriminations algorithmiques ne sauraient s'analyser comme un problème purement technique concernant les seuls chercheurs et praticiens de la technologie. Elles soulèvent en effet des questions en matière de droits humains, lesquelles nous concernent tous et vont au-delà des chartes éthiques (non contraignantes) déjà élaborées par certaines grandes entreprises. Nous ne pouvons pas nous permettre de nous retrancher derrière les complexités de l'IA en invoquant systématiquement une forme d'impuissance qui empêcherait nos sociétés, et en particulier les législateurs, d'adopter des réglementations de nature à protéger et à promouvoir l'égalité et la non-discrimination dans ce domaine[52]. Nous avons un besoin urgent de procédures, d'outils et de méthodes pour réglementer et contrôler ces systèmes afin de garantir leur conformité aux normes internationales en matière de droits humains. S'il ne fait aucun doute qu'une législation nationale efficace s'impose, nous devons aussi fixer des normes internationales compte tenu de la dimension transnationale et internationale marquée des technologies reposant sur l'IA.

64. Sans chercher à être exhaustif, je tiens à signaler ici certains textes phares qui ont déjà été adoptés, ainsi que des travaux actuellement en cours au niveau international et qui contribuent à baliser le chemin. A cet égard, je note avec intérêt le lancement, le 18 novembre 2019, des travaux du Comité ad hoc sur l'intelligence artificielle (CAHAI). Ce comité intergouvernemental du Conseil de l'Europe est notamment chargé d'examiner, sur la base de larges consultations multipartites, la faisabilité et les éléments potentiels d'un cadre juridique pour la conception, le développement et l'application de l'IA, fondés sur les normes du Conseil de l'Europe dans le domaine des droits de l'homme, de la démocratie et de l'État de droit[53]. Je note également avec intérêt la recommandation de la Commissaire aux droits de l'homme du Conseil de l'Europe, «Décoder l'intelligence artificielle: 10 mesures pour protéger les droits de l'homme», qui relève clairement l'importance des questions de non-discrimination et d'égalité dans ce domaine, ainsi que l'étude «Discrimination, intelligence artificielle et décisions algorithmiques» qui porte un éclairage précieux sur les risques de discrimination causés par la prise de décision algorithmique et d'autres types d'IA[54]. Enfin, je tiens à saluer également l'apport de la Recommandation Rec/CM(2020)1 du Comité des Ministres sur les impacts des systèmes algorithmiques sur les droits de l'homme, complétée par des lignes directrices sur cette question[55].

---

51. Défenseur des droits (France), Décision 2019-021 du 18 janvier 2019 relative au fonctionnement de la plateforme nationale de préinscription en première année de l'enseignement supérieur (Parcoursup).
52. Zimmermann A. et al., «Technology can't fix algorithmic injustice», *Boston Review*, 9 janvier 2020.
53. Mandat du CAHAI approuvé par le Comité des Ministres, lors de sa 1353ème réunion, le 11 septembre 2019.
54. Zuideveen Borgesius F., Discrimination, intelligence artificielle et décisions algorithmiques, étude publiée par la Direction Générale de la Démocratie, Conseil de l'Europe, 2018.

65. Au-delà des travaux menés au sein du Conseil de l'Europe, on peut citer entre autres initiatives européennes et internationales importantes dans ce domaine, les lignes directrices en matière d'éthique pour une IA digne de confiance publiées en avril 2019 par la Commission européenne[56]; une étude publiée par l'Agence des droits fondamentaux sur la qualité des données et l'intelligence artificielle: atténuer les distorsions et les erreurs dans les algorithmes[57]; les principes de l'OCDE sur l'IA, adoptées en mai 2019[58]; la déclaration du G20 sur une intelligence artificielle centrée sur l'humain, adoptée en juin 2019[59]; ainsi que les travaux de plusieurs agences des Nations Unies, dont ceux de l'UNESCO, encore en cours au moment de la rédaction du présent rapport.

66. Sur la base d'une analyse de ces travaux incontournables au niveau international, l'annexe au présent rapport expose cinq principes éthiques fondamentaux qui doivent sous-tendre tout travail de réglementation dans le domaine de l'IA: la transparence, la justice et l'équité, la responsabilisation (ou l'obligation de rendre compte), la sûreté et la sécurité, ainsi que le respect de la vie privée. Il est souligné dans l'annexe que le degré d'intégration du respect de ces principes fondamentaux dans un système d'IA donné dépend des utilisations prévues et prévisibles de celui-ci. En substance, plus le préjudice potentiel est important, plus les exigences à respecter sont strictes. Je tiens à préciser ici que le droit à l'égalité et à la non-discrimination est fondamental et intrinsèque aux démocraties fondées sur les droits humains et l'État de droit. Il doit être strictement respecté et protégé à tout moment et l'atteinte à ces principes résultant de l'utilisation d'un système d'IA ne devrait jamais être tolérée, quel que soit le contexte.

67. S'agissant spécifiquement de l'égalité et de la non-discrimination, certains points méritent une vigilance accrue. Premièrement, la législation anti-discrimination doit couvrir effectivement tous les motifs de discrimination, y compris, mais pas uniquement les discriminations réelles ou perçues fondées sur le sexe, le genre, l'âge, l'origine nationale ou ethnique, la couleur, la langue, les convictions religieuses, l'orientation sexuelle, l'identité de genre, les caractéristiques sexuelles, l'origine sociale, l'état civil, le handicap ou l'état de santé. La liste des motifs interdits par la loi devrait être ouverte, à savoir aussi complète que possible sans prétendre à l'exhaustivité.

68. Deuxièmement, aussi bien les formes directes qu'indirectes de discrimination doivent être efficacement couvertes. Cette règle revêt une importance particulière lorsqu'un pays n'autorise pas la collecte de certaines données (par exemple les données ethniques), mais que des algorithmes déduisent ces caractéristiques à partir de variables de substitution (pays de naissance d'une personne ou de ses parents, code postal, objet des interrogations entrées dans les moteurs de recherche, etc.) Pour des raisons analogues, la loi devrait également proscrire la discrimination fondée sur l'association réelle ou supposée d'une personne à un certain groupe[60].

69. Troisièmement, compte tenu des difficultés particulières inhérentes à la preuve des discriminations causées par l'utilisation de l'IA, il est crucial de ne pas imposer aux victimes de discriminations un fardeau de la preuve excessif. Exiger d'un individu qu'il établisse son « innocence » face à une prise de décision automatisée pourrait avoir – comme nous l'avons indiqué plus haut dans le rapport – de lourdes conséquences en matière de droits humains. Le partage de la charge de la preuve mise en place par les directives de l'Union européenne relatives à l'égalité constitue un modèle utile à cet égard[61].

70. Quatrièmement, il est indispensable de se doter de mécanismes efficaces d'application et le soutien des organismes nationaux de promotion de l'égalité peut s'avérer déterminant dans ce domaine. Beaucoup ont déjà commencé à réfléchir de manière constructive au rôle que ces organismes pourraient jouer pour garantir et promouvoir l'égalité dans un monde où l'utilisation de l'IA est en constante augmentation[62]. Il nous appartient de soutenir cette action et de veiller à ce que les organismes de promotion de l'égalité disposent de ressources suffisantes pour la mener à bien.

---

55. Recommandation CM/Rec(2020)1 du Comité des Ministres sur les impacts des systèmes algorithmiques sur les droits de l'homme (adoptée le 8 avril 2020, lors de la 1373ème réunion des Délégués des Ministres).
56. Groupe d'experts de haut niveau sur l'intelligence artificielle, Lignes directrices en matière d'éthique pour une IA digne de confiance, Bruxelles, Commission européenne, 8 avril 2019.
57. «Data quality and artificial intelligence – mitigating bias and error to protect fundamental rights», FRA, juin 2019.
58. Recommandation du Conseil sur l'intelligence artificielle, OCDE, 22 mai 2019.
59. www.mofa.go.jp/files/000486596.pdf.
60. Wachter S., «Affinity profiling and discrimination by association in online behavioural advertising», op. cit.
61. Directives 2000/43/CE et 2000/78/CE.
62. Allen R., Q.C. et Masters D., «Regulating for an equal AI: a new role for equality bodies», *Equinet*, Bruxelles, 2020.

71. Enfin, la transparence des modèles commerciaux revêt une importance particulière lorsque les systèmes basés sur l'IA affectent les choix proposés aux individus en ligne et lorsqu'il est difficile pour les intéressés de comparer les résultats algorithmiques. Tout système d'IA devrait également être systématiquement soumis à des tests rigoureux – visant à déceler les préjugés et discriminations qu'il peut véhiculer – avant son déploiement et périodiquement par la suite[63]. Les tests périodiques sont encore plus importants lorsque l'on a recours à l'apprentissage automatique pour permettre à l'algorithme d'évoluer après son déploiement, au risque de voir celui-ci arriver à se comporter d'une manière qui n'était pas prévue par les développeurs au départ.

72. Les considérations précédentes concernent très spécifiquement la problématique des discriminations résultant de l'utilisation de l'IA. Plus largement, j'insisterais sur l'importance d'associer dès le début de la réflexion, la société civile et les citoyens pour une participation critique et une acceptation construite paritairement au sujet des algorithmes. D'autre part, la régulation des algorithmes pourrait aller de pair avec la mise en place d'alternatives algorithmiques proposées par les pouvoirs publics (avec une véritable politique de proactivité en la matière) pour contrer la logique purement commerciale des principales entreprises technologiques; cette politique technologique d'initiative publique et éthique s'ajouterait à la politique de régulation.

73. En ce qui concerne la reconnaissance des droits humains, la jurisprudence de la Cour européenne doit comme toujours guider les États dans l'élaboration des normes de droit interne et servir de ligne rouge à leur action. Enfin, le domaine de l'IA évoluant très rapidement, il semblerait utile de poser, par anticipation, la question de la reconnaissance éventuelle de nouveaux droits humains tels que le droit à l'autonomie humaine, le droit à la transparence et à la justification, le droit du contrôle de l'IA, ou le droit à l'intégrité morale.

9. Conclusions

74. Comme la plupart des innovations techniques, l'IA n'est intrinsèquement ni bonne ni mauvaise. Comme un couteau tranchant, elle peut servir des objectifs utiles ou hautement préjudiciables. Le problème ne tient pas à l'outil, mais à sa conception et à (certains de) ses usages potentiels. Ce rapport vise résolument à faire en sorte que l'IA, en plus d'être intelligente, soit bénéfique aux individus et à nos sociétés.

75. L'utilisation de l'IA peut avoir un impact sur de nombreux droits (protection des données personnelles, respect de la vie privée, accès au travail, accès aux services publics et aux prestations sociales, coût des assurances, accès à la justice, procédure équitable, charge de la preuve, etc.) et toucher certains groupes plus que d'autres. Si elle ne crée pas forcément d'inégalités, elle risque d'exacerber celles prévalant déjà dans la société et de conduire à des résultats injustes dans de nombreux aspects de la vie.

76. En effet, l'apprentissage automatique repose sur l'utilisation de vastes ensembles de données ; or, les données sont toujours biaisées, car elles reflètent les discriminations existant déjà dans nos sociétés ainsi que les préjugés de ceux qui les collectent et les analysent.

77. Les algorithmes sont créés par des êtres humains travaillant le plus souvent au sein d'équipes homogènes. L'absence de diversité dans le milieu de l'IA crée un contexte favorable à la reproduction de stéréotypes préjudiciables. La diversification des équipes travaillant dans ce domaine est donc un enjeu majeur, lequel nécessite non seulement un travail en amont, au niveau de l'éducation, mais également sur le lieu de travail au prix d'une plus grande inclusivité à long terme.

78. L'absence de transparence des systèmes d'IA masque souvent des effets discriminatoires qui ne sont constatés qu'une fois le système déployé et ses effets négatifs subis par de nombreuses personnes. Pour éviter ces conséquences, les concepteurs doivent intégrer dès la phase initiale de leurs travaux une vision pluraliste, multidisciplinaire et inclusive en se posant plusieurs questions: Qui utilisera (ou sera affecté par) ce système ? Est-ce que le système produira des résultats justes pour tout le monde ? Dans le cas contraire, que dois-je faire pour corriger les injustices ? Comment définir les objectifs de l'outil (les résultats pour lesquels il doit être optimisé) de manière à ce que celui-ci produise des résultats non discriminatoires ?

79. Évaluer le respect des principes d'égalité et de non-discrimination avant qu'un système d'IA soit déployé revêt une importance d'autant plus grande que de nombreux droits individuels fondamentaux peuvent être en jeu. De plus, il peut s'avérer extrêmement difficile de contester les résultats produits, en raison de

63. Spinks S., «Algorithmic bias within online behavioural advertising means public could be missing out, says Associate Professor Sandra Wachter», Oxford Internet Institute blogpost, 26 novembre 2019.

l'effet « de boîte noire » des algorithmes. Or, ce sont souvent les personnes les plus marginalisées et les moins à même de contester les résultats qui sont les plus touchées, avec parfois des conséquences dévastatrices.

80. Les États doivent prendre ces questions à bras-le-corps et comprendre les défis qu'elles posent à nos sociétés aujourd'hui et qu'elles risquent de poser à l'avenir. Les questions en jeu dépassent les frontières et nécessitent également des réponses transnationales. Les acteurs privés – qui sont les principaux concepteurs de systèmes basés sur l'IA – doivent eux aussi être directement associés à la recherche de solutions. Les parlements également doivent s'intéresser de près aux implications de l'utilisation croissante de l'IA et à son impact sur les citoyens. Il s'agit, ensemble, d'encadrer ces systèmes afin de limiter leurs effets discriminatoires et de fournir un recours efficace lorsque des discriminations se produisent.

81. L'IA est souvent liée dans l'esprit collectif à l'innovation ; or, et cela est encore plus important, elle doit également répondre à l'impératif d'inclusivité. Il ne s'agit pas de freiner les innovations, mais de les encadrer de manière proportionnée aux enjeux afin que les systèmes basés sur l'IA intègrent pleinement le respect du droit à l'égalité et de l'interdiction de toute discrimination.

Annexe

Intelligence artificielle – Description et principes éthiques

On a tenté à plusieurs reprises de définir le terme «intelligence artificielle» depuis sa première utilisation en 1955. Ces initiatives s'intensifient aujourd'hui, car les organes normatifs, notamment le Conseil de l'Europe, réagissent aux capacités et à l'omniprésence croissantes de l'intelligence artificielle en œuvrant en faveur de son encadrement juridique. Il n'existe cependant toujours pas de définition «technique» ou «juridique» unique qui soit universellement admise[64]*. Aux fins du présent rapport, il est toutefois indispensable de décrire cette notion.*

À l'heure actuelle, le terme «intelligence artificielle» (IA) désigne en général les systèmes informatiques capables de percevoir et d'extraire des données de leur environnement, puis d'utiliser des algorithmes statistiques pour traiter ces données, afin d'obtenir des résultats qui correspondent à des objectifs prédéterminés. Les algorithmes se composent de règles définies par l'homme ou par l'ordinateur lui-même, qui «forme» l'algorithme en analysant des ensembles de données considérables et continue à affiner ces règles à mesure qu'il reçoit de nouvelles données. Cette méthode, connue sous le nom «d'apprentissage automatique» ou «d'apprentissage statistique», est actuellement la technique la plus utilisée pour les applications complexes; elle est uniquement devenue possible ces dernières années grâce à l'augmentation de la puissance de traitement des ordinateurs et à la disponibilité de données suffisantes. «L'apprentissage en profondeur» représente une forme particulièrement avancée d'apprentissage automatique, qui utilise plusieurs couches de «réseaux neuronaux artificiels» pour traiter les données. L'analyse ou la compréhension intégrale par l'homme des algorithmes développés par ces systèmes n'est pas toujours possible; aussi sont-ils parfois qualifiés de «boîtes noires» (il arrive que ce terme désigne également, mais pour une raison différente, les systèmes d'IA propriétaires protégés par les droits de propriété intellectuelle).

Les formes actuelles d'IA sont toutes «restreintes», c'est-à-dire affectées à une tâche unique définie. L'IA «restreinte» est également qualifiée parfois de «faible», même si les systèmes modernes de reconnaissance faciale, de traitement du langage naturel, de conduite autonome et de diagnostic médical, par exemple, sont incroyablement sophistiqués et effectuent certaines tâches complexes avec une rapidité et une précision étonnantes. «L'intelligence artificielle générale», parfois qualifiée d'IA «forte», qui est capable d'exécuter toutes les fonctions du cerveau humain, est encore à réaliser. La «super-intelligence artificielle» désigne un système dont les capacités dépassent celles du cerveau humain.

Comme le nombre de domaines dans lesquels les systèmes d'intelligence artificielle sont appliqués est en augmentation, puisqu'ils se propagent dans des domaines qui peuvent avoir un impact important sur les droits individuels et les libertés, ainsi que sur les systèmes démocratiques et l'État de droit, la dimension éthique de ce phénomène a fait l'objet d'une attention croissante et de plus en plus urgente.

Un large éventail d'acteurs a formulé de nombreuses propositions d'ensemble de principes éthiques qui devraient être appliqués aux systèmes d'IA. Ces propositions sont rarement identiques et diffèrent à la fois dans les principes qu'elles énoncent et par la manière dont elles définissent ces principes. Les études montrent que la teneur essentielle des principes éthiques qui devraient être appliqués aux systèmes d'IA fait néanmoins l'objet d'un large consensus; c'est notamment le cas des principes suivants[65]:

– *Transparence.* Le principe de transparence peut faire l'objet d'une interprétation élargie, de manière à englober l'accessibilité et l'explicabilité d'un système d'IA, en d'autres termes la possibilité donnée à un individu de comprendre le fonctionnement de ce système et le mode de production de ses résultats.

– *Justice et équité.* Ce principe englobe la non-discrimination, l'impartialité, la cohérence et le respect de la diversité et du pluralisme. Il suppose également que la personne à laquelle est appliqué un système d'IA puisse en contester les résultats, disposer d'une voie de recours et obtenir réparation.

– *Responsabilité.* Ce principe englobe le fait d'exiger qu'un être humain soit responsable de toute décision qui a des conséquences sur les droits et libertés individuels, et que l'obligation de rendre des comptes et la responsabilité juridique de ces décisions soient définies. Il est donc étroitement lié au principe de justice et d'équité.

---

64. Pour une vue d'ensemble élargie des tentatives de définition de «l'intelligence artificielle», voir *AI Watch: Defining Artificial Intelligence – Towards an operational definition and taxonomy of artificial intelligence*, Samoili S., López Cobo M., Gómez E., De Prato G., Martínez-Plumed F. et Delipetrev B., European Commission Joint Research Centre, 2020.

65. Voir *Lignes directrices sur l'éthique en matière d'IA: situation en Europe et dans le monde*, rapport provisoire commandé par le Comité ad hoc sur l'intelligence artificielle (CAHAI) du Conseil de l'Europe, Ienca M. et Vayena E., mars 2020.

– *Sûreté et sécurité.* Ce principe suppose que les systèmes d'IA fassent preuve de solidité, de sécurité contre toute ingérence extérieure et de sûreté contre la commission d'actes involontaires, conformément au principe de précaution.

– *Respect de la vie privée.* Si le respect des droits de l'homme en général peut être considéré comme inhérent aux principes de justice et d'équité, de sûreté et de sécurité, le droit au respect de la vie privée est particulièrement important chaque fois qu'un système d'IA procède au traitement de données à caractère personnel ou privé. Les systèmes d'IA doivent par conséquent respecter les normes contraignantes du Règlement général sur la protection des données (RGPD) de l'Union Européenne et de la Convention pour la protection des personnes à l'égard du traitement automatisé des données à caractère personnel du Conseil de l'Europe (STE n° 108, et sa version actualisée, la Convention 108+, STCE n° 223), le cas échéant.

La mise en œuvre effective des principes éthiques applicables aux systèmes d'IA exige une approche intégrée de l'éthique, notamment une évaluation de leur impact sur les droits de l'homme, de manière à garantir le respect des normes établies. Il ne suffit pas que ces systèmes soient conçus uniquement sur la base de normes techniques et que des éléments soient ajoutés à un stade ultérieur pour tenter de faire respecter les principes éthiques.

Dans quelle mesure le respect de ces principes doit-il être intégré dans des systèmes particuliers d'IA? Cela dépend des utilisations prévues et prévisibles de ces systèmes: plus leur impact sur l'intérêt général et les droits et libertés individuels est important, plus les garanties doivent être strictes. La réglementation éthique peut donc être mise en œuvre de différentes manières, depuis les chartes internes volontaires pour les domaines les moins sensibles jusqu'aux normes juridiques contraignantes pour les plus délicats. Dans tous les cas, il importe qu'elle prévoie des mécanismes de contrôle indépendants selon le niveau de réglementation.

Ces principes essentiels portent sur les systèmes d'IA et leur environnement immédiat. Ils n'ont pas vocation à être exhaustifs ni à exclure des préoccupations éthiques plus générales, telles que la démocratie (participation pluraliste des citoyens à l'élaboration de normes éthiques et réglementaires), la solidarité (qui admet les différences de point de vue des divers groupes) ou la durabilité (préservation de l'environnement de la planète).