

Étude sur l'impact des systèmes d'intelligence artificielle, leur potentiel de promotion de l'égalité, y compris l'égalité de genre, et les risques qu'ils peuvent entraîner en matière de non-discrimination



COMMISSION POUR L'ÉGALITÉ
DE GENRE (GEC) ET
LE COMITÉ DIRECTEUR
SUR L'ANTI-DISCRIMINATION,
LA DIVERSITÉ ET L'INCLUSION (CDADI)

COUNCIL OF EUROPE



CONSEIL DE L'EUROPE

Étude sur l'impact des systèmes d'intelligence artificielle, leur potentiel de promotion de l'égalité, y compris l'égalité de genre, et les risques qu'ils peuvent entraîner en matière de non-discrimination

COMMISSION POUR L'ÉGALITÉ DE GENRE (GEC)
ET LE COMITÉ DIRECTEUR
SUR L'ANTI-DISCRIMINATION,
LA DIVERSITÉ ET L'INCLUSION (CDADI)

Préparée par Ivana Bartoletti
and Raphaële Xenidis

Edition anglaise :

Study on the impact of artificial intelligence systems, their potential for promoting equality, including gender equality, and the risks they may cause in relation to non-discrimination

Les points de vue exprimés dans cet ouvrage n'engagent que le ou les auteurs et ne reflètent pas nécessairement la ligne officielle du Conseil de l'Europe.

La reproduction d'extraits (jusqu'à 500 mots) est autorisée, sauf à des fins commerciales, tant que l'intégrité du texte est préservée, que l'extrait n'est pas utilisé hors contexte, ne donne pas d'informations incomplètes ou n'induit pas le lecteur en erreur quant à la nature, à la portée et au contenu de ce texte. Le texte source doit toujours être cité comme suit :

« © Conseil de l'Europe, année de publication ». Pour toute autre demande relative à la reproduction ou à la traduction de tout ou partie de ce document, veuillez vous adresser à la Direction de la communication, Conseil de l'Europe (F-67075 Strasbourg Cedex), ou à publishing@coe.int.

Toute autre correspondance relative à ce document doit être adressée au Service Comité directeur sur l'anti-discrimination, la diversité et l'inclusion (CDADI)

Conseil de l'Europe
F-67075 Strasbourg Cedex France
E-mail: cdadi@coe.int

Conception de la couverture et mise en page : Documents and Publications Production Department (SPDP), Council of Europe
Cover photo: IStock

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

© Conseil de l'Europe, août 2023
Imprimé aux ateliers du Conseil de l'Europe

Préparée par :

Ivana Bartoletti,

Responsable mondiale de la protection de la vie privée chez Wipro, chercheuse invitée à l'Oxford Internet Institute de l'université d'Oxford et cofondatrice du réseau Women Leading in AI Network et **Raphaële Xenidis**, *Assistant professor* en droit européen, École de droit de Sciences Po Paris.

Table des matières

RÉSUMÉ	5
INTRODUCTION: LE CONTEXTE	9
SECTION 1	15
PRÉSENTATION DU « BIAIS AUTOMATISÉ »: COMMENT LES TECHNOLOGIES ALGORITHMIQUES PEUVENT-ELLES CONDUIRE À LA DISCRIMINATION ?	
1) Qu'est-ce que l'IA ?	15
2) Qu'est-ce que le biais algorithmique ?	17
3) L'impact discriminatoire de l'IA: quelques exemples concrets	23
4) En quoi la discrimination algorithmique est-elle différente d'autres formes de discrimination ?	33
5) La lutte contre la discrimination algorithmique: les meilleures pratiques et leurs limites	35
6) Questions de représentation et de participation: le manque de diversité et d'inclusion dans le secteur de l'IA	42
SECTION 2	47
LE PAYSAGE JURIDIQUE ET POLITIQUE EN EUROPE: FORCES ET FAIBLESSES	
I. La discrimination et l'égalité: les instruments juridiques et politiques et leurs limites	49
II. Le droit relatif à la vie privée et à la protection des données: équité et exactitude	68
III. Les réglementations sectorielles de l'IA: forces et limites pour la promotion de l'égalité et la lutte contre la discrimination	72
SECTION 3	77
LA PROMOTION DE L'ÉGALITÉ DANS ET PAR L'UTILISATION DE L'IA: LE RÔLE DES ACTIONS ET OBLIGATIONS POSITIVES	
I. La révision des règles existantes à la lumière des nouvelles asymétries de pouvoir	77
II. Une obligation de promouvoir l'égalité dans et par l'utilisation de systèmes algorithmiques: le rôle des actions et obligations positives	82
SECTION 4	91
RECOMMANDATIONS	
Recommandations relatives à la politique des pouvoirs publics: vers une approche de l'IA fondée sur les droits humains	91

Résumé

A lors que le déploiement des systèmes algorithmiques et des applications d'intelligence artificielle (IA) ne cesse de prendre de l'ampleur et de l'importance, la discrimination algorithmique est un sujet de préoccupation croissante pour les citoyens et les citoyennes. Des mesures réglementaires sont en train d'être élaborées partout dans le monde, y compris au sein de l'Union européenne. Le Conseil de l'Europe a commencé à travailler sur un cadre juridique pour le développement, la conception et l'application de l'intelligence artificielle, basé sur les normes en matière de droits humains, de démocratie et d'État de droit qu'il a établies. S'appuyant sur une « Étude de faisabilité pour un cadre juridique relatif à la conception, au développement et à l'application de l'IA, fondé sur les normes du Conseil de l'Europe », publiée en 2020, ainsi que sur les « Éléments potentiels d'un cadre juridique sur l'intelligence artificielle, fondés sur les normes du Conseil de l'Europe en matière de droits humains, de démocratie et d'État de droit », le Comité sur l'intelligence artificielle (CAI), créé en 2022, travaille actuellement à la rédaction d'une convention-cadre « sur le développement, la conception et l'application de l'intelligence artificielle, qui se fonde sur les normes du Conseil de l'Europe en matière de droits de l'homme, de démocratie et d'État de droit et est propice à l'innovation ». ¹ Un tel instrument juridiquement contraignant du Conseil de l'Europe aurait la capacité de favoriser une **approche** de l'utilisation de l'IA et des technologies algorithmiques **fondée sur les droits humains** dans et au-delà de la communauté internationale des États parties à la Convention européenne des droits de l'homme (CEDH). La présente étude enquête dans cette perspective les risques de discrimination liés aux technologies algorithmiques, les réponses juridiques spécifiques à la discrimination algorithmique que peut proposer le Conseil de l'Europe, et le potentiel de ces technologies pour promouvoir l'égalité, y compris l'égalité de genre. L'étude est structurée en trois sections suivies de recommandations, dont un résumé est présenté ci-dessous. La première section aborde les questions liées au biais automatisé et examine comment les technologies algorithmiques peuvent conduire à la discrimination. La deuxième section étudie les forces et les faiblesses du cadre juridique dont on peut se servir pour lutter contre la discrimination

1. Voir le mandat du Comité sur l'intelligence artificielle CM(2021)131 : <https://rm.coe.int/cai-mandat/1680a7b90a>.

algorithmique au niveau du Conseil de l'Europe. La troisième section analyse la manière dont l'action positive et les obligations positives peuvent être utilisées pour combattre la discrimination algorithmique et s'attaquer, dans une démarche de transformation, tant à ses racines sociales qu'à ses manifestations dans les applications technologiques.

Résumé des recommandations

Comme la présente étude le met en évidence, la résolution du problème de la discrimination algorithmique nécessite une approche multidimensionnelle. Il est proposé que le Conseil de l'Europe élabore une approche solide et ancrée dans les droits humains pour les questions relatives à l'IA dans le domaine de l'égalité, en préparant une recommandation spécifique du Comité des Ministres sur l'IA, l'égalité, y compris l'égalité de genre, et la discrimination. Cet instrument devrait être rédigé par un comité d'experts sous l'égide de la Commission pour l'égalité de genre (GEC) et du Comité directeur sur l'anti-discrimination, la diversité et l'inclusion (CDADI), et s'appuierait sur les principes généraux des droits humains que sont l'égalité, y compris l'égalité de genre, et la non-discrimination, y compris tels qu'ils figureront dans la future convention-cadre sur l'intelligence artificielle qui est en cours de préparation. Cette approche devrait s'organiser autour de quatre axes complémentaires d'intervention réglementaire et politique :

1. Prévention, transparence et responsabilité

Les États membres sont encouragés à :

- ▶ étendre le recours aux **actions positives** pour s'attaquer à la discrimination algorithmique, et mettre à profit la notion d'obligation positive ancrée dans la jurisprudence de la Cour européenne des droits de l'homme pour soumettre les fournisseurs et les utilisateurs à une obligation de prévention raisonnable la discrimination algorithmique.
- ▶ Introduire une **obligation de mener des évaluations des risques de discrimination et des analyses d'impact sur l'égalité** tout au long du cycle de vie des systèmes algorithmiques, en fonction des usages spécifiques qui sont les leurs.
- ▶ Examiner comment des **mécanismes de certification** pourraient être utilisés pour s'assurer que les biais ont été atténués et les risques de discrimination supprimés dans la mesure du possible pour des usages bien définis.
- ▶ Étudier le **rapport entre responsabilité, transparence et droit du secret des affaires** dans le domaine de l'IA, ainsi que les risques

associés concernant l'égalité, y compris l'égalité de genre, et la non-discrimination.

- ▶ Examiner la possibilité de mettre en place des obligations juridiques imposant aux utilisateurs de systèmes d'IA de **publier des données statistiques** pour permettre aux parties intéressées d'évaluer les effets discriminatoires d'un système donné dans le contexte de plaintes pour discrimination.
- ▶ Mettre en place des **mécanismes de transparence** afin que les personnes intéressées puissent évaluer les effets discriminatoires potentiels d'un système donné.
- ▶ Regrouper les mesures de prévention, de transparence et de responsabilité dans un **plan d'action global sur l'IA et l'égalité**.

2. Accès à la justice et mécanismes de recours

Les États membres sont encouragés à :

- ▶ faciliter l'accès à la justice en mettant en place des **mécanismes de contrôle public** et des **moyens d'action collective** permettant d'obtenir réparation en cas de discrimination algorithmique.
- ▶ Examiner la possibilité d'ajuster, compléter et consolider les **règles de preuve** et d'en renforcer l'efficacité pour une **répartition plus juste et plus équitable de la charge de la preuve**.
- ▶ Favoriser la **coopération entre les autorités de régulation**.
- ▶ Étudier les **nouvelles formes de vulnérabilité « algorithmique »** liées à l'utilisation des systèmes d'IA, et examiner la possibilité de mettre en place une **protection juridique** contre cette vulnérabilité.
- ▶ Affirmer sans ambiguïté que l'interdiction de la discrimination énoncée à l'article 14 de la CEDH concerne la **discrimination intersectionnelle et la discrimination par procuration**, deux formes d'atteintes discriminatoires que les systèmes algorithmiques sont particulièrement susceptibles de provoquer.
- ▶ Étudier les possibilités d'utiliser le **droit de la protection du consommateur pour compléter la lutte contre la discrimination**, par exemple en facilitant l'accès à l'information, en interdisant certaines caractéristiques dans les systèmes algorithmiques conformément à la notion de clause abusive, etc.

3. Diversité, inclusion, représentation et participation

- ▶ Les États membres devraient **identifier, soutenir et appliquer de façon active des mesures d'action positive**, notamment de diversification des communautés professionnelles.
- ▶ **Les obligations positives de promouvoir l'égalité** devraient donner une base juridique pour faire en sorte que la promotion de l'égalité soit au cœur de l'élaboration des systèmes d'IA et des systèmes algorithmiques.
- ▶ Les obligations positives de promouvoir l'égalité pourraient également se traduire par l'obligation pour les entreprises du secteur de l'IA de **développer et de mettre en œuvre une stratégie d'égalité** couvrant les groupes protégés par l'article 14 de la CEDH et l'article 1 du Protocole 12 à la CEDH.

4. Participation démocratique, sensibilisation de l'opinion publique et renforcement des capacités

Les États membres sont encouragés à :

- ▶ instituer un **droit à l'information sur la médiation algorithmique** dans le contexte de plaintes ou de réclamations pour discrimination.
- ▶ promouvoir la **mise en place de programmes d'éducation au numérique**, pour renforcer parmi les citoyen-nes la connaissance de leurs droits numériques en matière d'égalité, y compris l'égalité de genre, et de non-discrimination.
- ▶ renforcer les exigences légales applicables à la **participation démocratique dans l'élaboration des normes**, compte tenu du rôle important que joue la normalisation dans le domaine de l'IA en ce qui concerne l'égalité, y compris l'égalité de genre, et la non-discrimination.
- ▶ Investir dans le **renforcement des capacités**, notamment la recherche interdisciplinaire sur les algorithmes non discriminatoires et les stratégies visant à protéger l'égalité dans l'utilisation des systèmes algorithmiques.

Introduction : le contexte

L'intelligence artificielle (IA), qui est omniprésente, est souvent saluée pour sa capacité à réduire les frictions et à simplifier des processus qui étaient auparavant manuels et fastidieux. Les recherches menées dans ce domaine continuent d'accélérer leur trajectoire, franchissant les frontières et changeant le mode de vie de chacun et chacune d'entre nous.

Dans le domaine des soins de santé, l'automatisation du diagnostic médical pourrait donner à une patiente la possibilité d'utiliser sans rendez-vous des services aussi complexes que le dépistage du cancer du sein et les examens IRM. Le processus permettrait de diagnostiquer les maladies dangereuses en plus grand nombre et à un stade beaucoup plus précoce. Les villes intelligentes peuvent améliorer la gestion du trafic et l'allocation des ressources, et l'analyse de données massives peut optimiser les ressources de notre environnement. L'IA est également de plus en plus utilisée comme outil d'information et de prise de décision dans divers domaines, notamment l'administration et les politiques publiques, le logement et les soins de santé, l'éducation et la justice pénale. Ces temps derniers, il a beaucoup été question dans les médias de l'IA générative (un type spécifique d'IA utilisée pour générer de nouveaux contenus – du texte, des images, de la musique, etc. – en utilisant des algorithmes d'apprentissage profond comme GAN ou Transformers), en raison de l'accessibilité et de l'utilisation de ChatGPT3, une évolution de l'IA générative qui produit des contenus se rapprochant de ceux produits par un être humain et peut être utilisée pour générer des codes informatiques, des exposés de niveau universitaire, des poèmes, etc.

Au cours des dernières années, ces potentialités et les avantages considérables que cette technologie peut apporter aux citoyen·nes ont été quelque peu éclipsés par la prise de conscience d'un inconvénient : la « *logicisation* »² des discriminations et des inégalités existantes. Par exemple, dans ce que les Néerlandais·es ont appelé le « *toeslagenaffaire* », ou le scandale des allocations familiales, des milliers de personnes ont subi les conséquences d'un algorithme d'auto-apprentissage biaisé qui créait des profils de risque dans le but de détecter des fraudes aux allocations familiales. Les victimes

2. La « *logicisation* » (en anglais : « *softwarisation* ») des biais signifie que les inégalités existantes finissent par être codées et perpétuées dans des machines obscures et protégées par la propriété intellectuelle (voir page 10 pour de plus amples explications).

de ce cas de profilage algorithmique ont été plongées dans un désarroi et une pauvreté aggravée, au point que l'une d'entre elles a fait une tentative de suicide³. Un rapport parlementaire sur le scandale des allocations familiales a relevé plusieurs lacunes graves, notamment des biais institutionnels et la dissimulation ou la déformation, par les autorités, d'informations visant à tromper le parlement sur les faits⁴.

En 2018, Reuters a rapporté qu'Amazon a essayé d'utiliser l'IA pour mettre au point un système de sélection de CV en utilisant les CV que l'entreprise avait collectés au cours de la décennie précédente⁵. Comme ces CV provenaient principalement d'hommes et que les conséquences de ce déséquilibre n'avaient pas été sérieusement prises en compte, le nouveau système est donc apparu discriminatoire à l'égard des femmes et a dû être abandonné. En 2019, la carte de crédit commercialisée par Apple a été visée par une enquête parce que les femmes se voyaient attribuer une limite de crédit plus basse que leurs conjoints masculins qui avaient le même revenu et la même solvabilité⁶.

Ces exemples ne sont ni des scénarios marginaux ni des scénarios extrêmes. Les systèmes algorithmiques sont trop souvent conçus et alimentés par des données et des modèles historiques qui reproduisent des stéréotypes et de fausses hypothèses sur le genre, la race, l'orientation sexuelle, les capacités, la classe sociale, l'âge, la religion ou les convictions, la géographie et d'autres facteurs socioculturels et démographiques. **Il en résulte que, en l'absence d'effort particulier, l'utilisation des technologies algorithmiques perpétue et amplifie les inégalités sociétales et les stéréotypes nuisibles.**

La prise de conscience des risques de discrimination algorithmique s'est cristallisée autour des discussions sur le « biais », qui est désormais devenu un enjeu public considérable. Une enquête menée en 2022 a montré que plus de 36 % des entreprises « rencontrent des difficultés ou subissent un impact commercial direct en raison d'un biais lié à l'IA dans leurs algorithmes, tels

-
3. Melissa Heikkilä, Dutch scandal serves as a warning for Europe over risks of using algorithms, Politico, 29 mars 2022, voir : <https://www.politico.eu/article/dutch-scandal-serves-as-a-warning-for-europe-over-risks-of-using-algorithms/> (dernière consultation le 30 août 2022)
 4. Voir Tweede Kamer der Staten-Generaal, Parlementaire ondervraging kinderopvangtoeslag (2020) : <https://zoek.officielebekendmakingen.nl/kst-35510-1.pdf>.
 5. Amazon scraps secret AI recruiting tool that showed bias against women, Reuters, 11 octobre 2018 : <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G> (dernière consultation le 25 juillet 2022).
 6. Alisha Haridasani Gupta, "Are Algorithms Sexist?" *The New York Times* (15 November 2019) available at: <https://www.nytimes.com/2019/11/15/us/apple-card-goldman-sachs.html> (last accessed: 25 juillet 2022).

que [...] [l]a perte de revenus, [l]a perte de client-es, [l]a perte de salarié-es, [i] des frais juridiques occasionnés par des poursuites ou actions en justice [et] [d] l'atteinte à la réputation de la marque/la réaction des médias⁷ ». Les pouvoirs législatifs et réglementaires du monde entier sont également aux prises avec ces risques et avec les insuffisances de la législation existante pour y faire face. Le questionnaire auquel ont répondu les membres et observateurs de la GEC et du CDADI aux fins de la présente étude montre une large sensibilisation aux questions juridiques liées au biais algorithmique⁸. Dans presque tous les États parties, des initiatives politiques ou législatives sont en cours ou des consultations publiques ont lieu à cette fin.

Le Conseil de l'Europe a entrepris des travaux dans ce domaine. Le Comité ad hoc sur l'intelligence artificielle (CAHA) a été chargé en 2019-2021 de consulter les parties prenantes et d'examiner la faisabilité et les éléments potentiels d'un cadre juridique pour le développement, la conception et l'application de l'intelligence artificielle, fondée sur les normes du Conseil de l'Europe en matière de droits humains, de démocratie et d'État de droit. Le Comité a publié en 2020 une « Étude de faisabilité pour un cadre juridique relatif à la conception, au développement et à l'application de l'IA, fondé sur les normes du Conseil de l'Europe », ainsi que des « Éléments potentiels d'un cadre juridique sur l'intelligence artificielle, fondés sur les normes du Conseil de l'Europe en matière de droits humains, de démocratie et d'État de droit ». À la suite de ces développements, un nouveau Comité sur l'intelligence artificielle (CAI) a été créé en 2022 et chargé de rédiger une convention-cadre « sur le développement, la conception et l'application de systèmes d'intelligence artificielle, fondée sur les normes du Conseil de l'Europe en matière de droits humains, de démocratie et d'État de droit, et propice à l'innovation⁹ ». Un instrument juridiquement contraignant du Conseil de l'Europe aurait une capacité appréciable de favoriser une **approche de** l'utilisation de l'IA et des technologies algorithmiques **fondée sur les droits humains** dans et au-delà de la communauté internationale des États parties à la CEDH. En outre, le Comité des Ministres a demandé à la Commission pour l'égalité de genre et au Comité directeur sur l'anti-discrimination, la diversité et l'inclusion de contribuer aux travaux sur un éventuel cadre juridique pour les systèmes d'intelligence artificielle, de préparer une étude sur l'impact des systèmes

7. Voir DataRobot, "DataRobot's State of AI Bias Report Reveals 81% of Technology Leaders Want Government Regulation of AI Bias" (2022) : <https://www.datarobot.com/newsroom/press/datarobots-state-of-ai-bias-report-reveals-81-of-technology-leaders-want-government-regulation-of-ai-bias/>.

8. Voir la section II.

9. Voir le mandat du Comité sur l'intelligence artificielle CM(2021)131 : <https://rm.coe.int/cai-terms-of-reference/1680a7b90b>.

d'intelligence artificielle, leur potentiel pour promouvoir l'égalité, y compris l'égalité de genre, et les risques qu'ils peuvent entraîner en matière de discrimination, et, sous réserve des résultats de l'étude, d'élaborer en étroite coopération avec le CAI un éventuel instrument juridique spécifique.

La présente étude a trois objectifs. Premièrement, elle explique comment les biais apparaissent dans l'IA et les technologies algorithmiques, et comment ils peuvent entraîner une discrimination. Elle montre que les biais ne sont pas uniquement liés aux données, mais aussi aux fondements humains et sociaux plus larges de ces outils technologiques. Deuxièmement, l'étude examine les moyens qu'ont les responsables politiques, les parlementaires et les entreprises de faire face aux risques discriminatoires des technologies algorithmiques et évalue quels instruments juridiques existants pourraient être utilisés à cette fin à l'avenir. Elle recense également les lacunes des outils juridiques existants et propose des ajustements réglementaires pour promouvoir l'égalité et empêcher que des éléments de discrimination apparaissent pendant le développement et le déploiement des systèmes algorithmiques. Troisièmement, l'étude examine les conditions sociopolitiques nécessaires pour que les technologies algorithmiques soient utilisées pour promouvoir l'égalité. Elle présente les possibilités de tirer parti de ces technologies en utilisant les voies juridiques de l'action positive et des obligations positives. Enfin, l'étude recommande plusieurs pistes pour que l'utilisation des technologies algorithmiques n'automatise pas les inégalités existantes mais contribue à une société meilleure et plus équitable. Dans l'ensemble, cette étude vise à contribuer aux travaux d'un futur comité d'expert-es chargé, sous l'égide de la GEC et du CDADI, de rédiger un éventuel instrument juridique sectoriel spécifique sur l'impact des systèmes d'intelligence artificielle sur l'égalité, notamment l'égalité de genre, et la non-discrimination en 2024 et 2025.

L'étude se concentre principalement sur l'Europe et présente les opportunités et les problèmes que le déploiement des technologies algorithmiques dans la société pose en matière d'égalité et de discrimination. Elle analyse les réponses qui ont été données et qui sont en cours de discussion dans plusieurs pays membres du Conseil de l'Europe ou ayant le statut d'observateur à la GEC ou au CDADI. L'étude s'appuie sur l'étude de Borgesius relative à « la discrimination, l'intelligence artificielle et la prise de décision algorithmique » commandée par le Conseil de l'Europe en 2018, ainsi que sur le corpus de recherche interdisciplinaire sur la discrimination algorithmique et les

biais associés à l'IA¹⁰, qui se développe rapidement. Le document aborde les questions relatives à la discrimination algorithmique pour tous les motifs protégés par l'article 14 de la Convention européenne des droits de l'homme (CEDH), mais en mettant l'accent sur les trois groupes de motifs protégés que sont le genre et le sexe, l'identité de genre et les caractéristiques sexuelles ainsi que la race, l'origine ethnique et nationale, la couleur, la citoyenneté, la religion et la langue. L'étude examine les conséquences nuisibles des biais véhiculés par l'IA dans un large éventail de secteurs publics et privés, mais en mettant l'accent sur l'emploi et l'éducation. Enfin, l'étude se concentre sur le contexte juridique et les instruments du Conseil de l'Europe, mais s'aligne et complète l'approche basée sur le risque adoptée par l'Union européenne dans sa proposition de loi sur l'IA.

10. Voir Frederik Borgesius, *Discrimination, intelligence artificielle et prise de décision algorithmique* (2018) Conseil de l'Europe : <https://rm.coe.int/etude-sur-discrimination-intelligence-artificielle-et-decisions-algori/1680925d84>. Voir aussi Agence des droits fondamentaux de l'Union européenne, *Bias in Algorithms – Artificial Intelligence and Discrimination* (2022) t : https://fra.europa.eu/sites/default/files/fra_uploads/fra-2022-bias-in-algorithms_en.pdf, et Carsten Orwat, « Diskriminierungsrisiken durch Verwendung von Algorithmen » (Antidiskriminierungsstelle des Bundes, 2019).

Section 1

Présentation du « biais automatisé » : comment les technologies algorithmiques peuvent- elles conduire à la discrimination ?

Note terminologique: Par souci de clarté, le terme « **utilisateur** » d’algorithmes désigne les entreprises, les organismes publics ou toute autre partie prenante qui déploie un algorithme pour soutenir ou automatiser un processus décisionnel. En revanche, les « **utilisateurs finaux/ utilisatrices finales** » sont les personnes qui sont soumises à des décisions algorithmiques ou soutenues par des algorithmes, par exemple les client-es, les candidat-es à un emploi, les contribuables, etc. Les « **fournisseurs** » de systèmes algorithmiques et d’IA sont les entreprises qui conçoivent et commercialisent de tels systèmes sans les mettre en œuvre dans des conditions réelles. Parfois, lorsque les systèmes algorithmiques ou d’IA sont développés en interne, le fournisseur et l’utilisateur constituent la même entité.

1) Qu’est-ce que l’IA ?

Aux fins de cette analyse, nous utilisons la **définition large de l’IA** proposée par le comité ad hoc sur l’intelligence artificielle (CAHAI) du Conseil de l’Europe, qui indique que l’IA est un « terme générique » désignant diverses applications informatiques qui s’appuient sur différentes techniques et qui présentent des capacités communément et actuellement associées à l’intelligence humaine¹¹. Le CAHAI reconnaît que « [c]es techniques peuvent consister en des modèles formels (ou systèmes symboliques) ainsi qu’en

11. Comité ad hoc sur l’intelligence artificielle, étude de faisabilité CAHAI(2020)23 (Conseil de l’Europe, 2020), [8].

des modèles fondés sur des données (systèmes basés sur l'apprentissage) qui reposent généralement sur des approches statistiques, y compris par exemple l'apprentissage supervisé, l'apprentissage non supervisé et l'apprentissage de renforcement» et que «les systèmes d'IA agissent dans le monde réel ou numérique en percevant leur environnement par l'acquisition de données, en analysant certaines données structurées ou non structurées, en raisonnant sur les connaissances récoltées ou en traitant les informations dérivées des données, et sur cette base, décident des meilleures actions à prendre pour atteindre un certain objectif¹².» Un autre aspect de la définition est que «[ces systèmes] peuvent également être conçus pour **adapter leur comportement dans le temps en fonction de nouvelles données** et améliorer leurs performances en vue d'atteindre un certain objectif¹³.»

Cette définition élargie de l'IA s'explique par le fait qu'à ce jour, il n'existe pas de définition unique de l'IA acceptée par la communauté scientifique. Par exemple, la proposition de règlement de l'UE sur l'IA considère que l'IA est «un logiciel développé au moyen d'une ou plusieurs des techniques et approches énumérées [...] et qui peut, pour un ensemble donné d'objectifs définis par l'homme, générer des résultats tels que des contenus, des prédictions, des recommandations ou des décisions influençant les environnements avec lesquels il interagit¹⁴».

Selon la définition de l'UE, les techniques et les approches permettant de considérer qu'un logiciel est un système d'IA sont les suivantes :

- ▶ « Approches d'apprentissage automatique (notamment d'apprentissage supervisé, non supervisé et par renforcement, utilisant une grande variété de méthodes, dont l'apprentissage profond) ;
- ▶ Approches fondées sur la logique et les connaissances (notamment la représentation des connaissances, la programmation inductive (logique), les bases de connaissances, les moteurs d'inférence et de déduction, le raisonnement (symbolique) et les systèmes experts) ;
- ▶ Approches statistiques, estimation bayésienne, méthodes de recherche et d'optimisation¹⁵».

Cette diversité de techniques relevant de la définition de l'IA comprend les logiciels qui font fonctionner, par exemple, les moteurs de recherche, les systèmes de reconnaissance d'images, la synthèse vocale, les sites web de

12. Ibid.

13. Ibid.

14. Acte AI de l'UE, Art. 3(1).

15. Voir l'annexe 1 de la réglementation de l'UE relative à l'IA : «Techniques et approches d'intelligence artificielle visées à l'article 3, point 1 ».

traduction automatique, les assistants virtuels, les filtres antispams, les programmes d'aide au diagnostic médical, ainsi que des machines telles que les voitures à conduite autonome, les robots et une myriade d'objets relevant de la vaste catégorie de l'internet des objets¹⁶. Dans cette étude, nous estimons qu'il est important de souligner que **le sujet de la réglementation n'est pas l'IA prise isolément, mais plutôt le dispositif sociotechnique plus large** constitué par l'interaction des éléments sociaux avec les technologies algorithmiques.

2) Qu'est-ce que le biais algorithmique ?

Les algorithmes sont capables de traiter un éventail beaucoup plus large de données et de variables pour prendre des décisions, avec une rapidité et, dans bien des domaines, une fiabilité, qui dépassent de loin les capacités humaines. Qu'il s'agisse des publicités qui nous sont diffusées, des produits qui nous sont proposés ou des résultats qui nous sont présentés après une recherche en ligne, les algorithmes jouent un rôle de plus en plus important dans ces décisions.

Pendant, étant donné qu'ils ne font que présenter les résultats de calculs **définis par des humains** à partir de mégadonnées collectées auprès d'humains, de machines ou une combinaison des deux (à un moment donné du processus), les algorithmes reflètent et traitent les biais humains qui sont incorporés lorsqu'ils sont programmés, lorsqu'ils traitent des données et lorsque les humains interagissent avec eux¹⁷.

En bref, « le [biais algorithmique] **se produit lorsqu'un programme apparemment inoffensif intègre les préjugés de ses créateurs ou des données qui l'alimentent**¹⁸ ». Par conséquent, les femmes par exemple (en particulier celles qui appartiennent à un groupe minoritaire) peuvent se voir refuser des prêts et des crédits, et l'identification des mots prononcés par des personnes noires par les programmes de reconnaissance vocale présentera des taux d'erreur beaucoup plus élevés que pour les personnes blanches¹⁹.

16. Parlement européen, « Intelligence artificielle: définition et utilisation » (2021) voir: <https://www.europarl.europa.eu/news/fr/headlines/society/20200827STO85804/intelligence-artificielle-definition-et-utilisation>.

17. Voir par exemple Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Broadway Books 2017).

18. Garcia, Megan. « Racist in the Machine: The Disturbing Implications of Algorithmic Bias ». » *World Policy Journal* 33 (2016): 111 - 117.

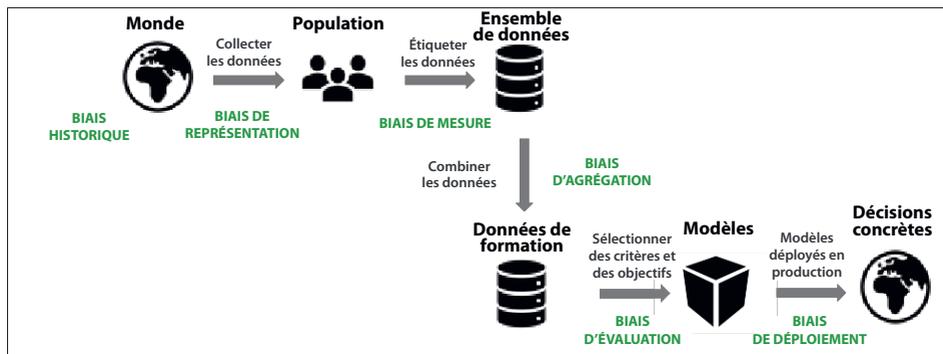
19. Allison Koenecke, et al., PNAS, 23 mars 2020

Comme le précise la notion d'« oppression algorithmique » de Sofiya Noble, le biais n'est pas un « bug » dans des systèmes par ailleurs impartiaux ; il est au contraire **systémique et propre au fonctionnement des systèmes d'information** qui alimentent les moteurs de recherche et autres applications web²⁰.

Contrairement à une idée largement répandue, les ensembles de données ne sont pas les seuls relais des biais dans les algorithmes d'apprentissage. En effet, les biais proviennent de différentes sources tout au long du cycle de vie des applications algorithmiques, de leur conception à leur déploiement et à leur utilisation. **La complexité de l'émergence et de l'impact des biais est la raison pour laquelle une attention particulière doit être accordée à l'ensemble du cycle de vie de l'IA et des systèmes algorithmiques**²¹. Plusieurs taxonomies répertoriant les sources de biais et leur canalisation dans les systèmes et les résultats de l'IA ont été élaborées par les chercheurs. Par exemple, le diagramme ci-dessous, réalisé par Suresh et Guttag, montre les différents points d'entrée du biais, et ce qu'ils impliquent.

-
20. Voir Safiya Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (New York: New York University Press, 2018) ; Vanessa Ceia, Benji Nothwehr, et Liz Wagner, *Gender and Technology: A rights-based and intersectional analysis of key trends* (Oxfam Research Backgrounder, 2021), 40.
 21. Ivana Bartoletti, *The Complex Issue of Algorithmic Fairness*, The Yuan, septembre 2021 : <https://www.the-yuan.com/129/The-Complex-Issue-of-Fairness-in-AI-Part-1.html> (dernière consultation le 28 juillet 2022)

Le tableau et les définitions ci-dessous ont été empruntés à l'ouvrage « A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle²² »



Suresh et Guttat distinguent cinq sources et types de biais dans les systèmes d'IA. Premièrement, ce qu'ils dénomment le « **biais historique** » décrit comment les hiérarchies sociales et les désavantages institutionnalisés façonnent les données sociales²³. Les données ne sont donc pas neutres car elles sont le reflet de la société inégalitaire dans laquelle nous vivons. Les femmes, par exemple, qui gagnent historiquement moins que les hommes, peuvent se voir désavantagées en ce qui concerne l'octroi d'un crédit²⁴ ou, dans le contexte de la publicité, se voir proposer des annonces avec des postes moins bien rémunérés²⁵.

Le « **biais de représentation** », quant à lui, apparaît dans la collecte des données²⁶. Si, par exemple, l'équipe de marketing d'une organisation fait de la publicité dans des quartiers à prédominance blanche, la base de client-es

22. Harini Suresh et John Guttat. 2021. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. Actes de la Conférence ACM sur l'équité et l'accès dans les algorithmes, les mécanismes et en Optimisation (EAAMO'21). ACM, New York, NY, USA, 9 pages: <https://doi.org/10.1145/3465416.3483305>.

23. Voir *ibid.*

24. Apple's 'sexist' credit card investigated by US regulator, BBC, 11 novembre 2019: <https://www.bbc.com/news/business-50365609> (dernière consultation : 15 juin 2022).

25. Samuel Gibbs, Women less likely to be shown ads for high-paid jobs on Google, study shows, The Guardian, 8 juillet 2015: <https://www.theguardian.com/technology/2015/jul/08/women-less-likely-ads-high-paid-jobs-google-study> (dernière consultation : 15 juin 2022).

26. Actes de la Conférence ACM sur l'équité et l'accès dans les algorithmes, les mécanismes et en optimisation (EAAMO'21). ACM, New York, NY (États-Unis d'Amérique), 9 pages: <https://doi.org/10.1145/3465416.3483305>.

qui en résulte ne sera pas représentative de l'ensemble de la population. Cet ensemble de données introduirait un biais s'il était utilisé par la suite pour créer un algorithme visant à analyser des groupes de population plus larges²⁷.

Les chercheurs et chercheuses ont également mis en lumière le « **biais de mesure** », qui « survient lors du choix, de la collecte ou du calcul des caractéristiques et des étiquettes à utiliser dans un problème de prédiction²⁸ ». De nombreuses caractéristiques et étiquettes ne posent pas de problème, comme l'étiquetage d'une image identifiant un chat ou un chien, mais des problèmes peuvent apparaître lorsque certains facteurs sont utilisés comme substituts. Le code postal, par exemple, peut être un indicateur de la race ou de l'orientation sexuelle, la profession peut être un indicateur de genre et le prénom est souvent utilisé comme substitut de l'âge²⁹. Par ailleurs, si le substitut simplifie excessivement la caractéristique à mesurer ou reflète des variations dans la qualité des mesures entre les groupes, un biais de mesure peut apparaître³⁰.

Le « **biais d'agrégation** » se rapporte à la façon dont les données sont combinées. Il se produit lorsque des groupes de données combinés de manière inappropriée produisent un modèle qui n'est performant pour aucun groupe ou qui ne l'est que pour le groupe majoritaire³¹. Les chercheurs et chercheuses mentionnent l'exemple des significations locales attribuées par des communautés spécifiques aux émojis, hashtags et phrases sur les réseaux sociaux, qui diffèrent des significations qui leur sont attribuées dans la population plus large des utilisateurs et utilisatrices de ces réseaux³². Cet écart pourrait conduire, par exemple, les modérateur·ices de contenus à appliquer à des groupes minoritaires des filtres sémantiques inadéquats modélisés sur les groupes majoritaires, ce qui aurait pour effet de les réduire au

27. Voir en outre les exemples p. XX sur le fait que la recherche en matière de soins de santé s'appuie sur des modèles masculins, in Criado Perez C, *Invisible women: Exposing data bias in a world designed for men* (Random House 2019).

28. Ibid.

29. Voir divers outils conçus pour prédire l'âge à partir de données sur les noms : <https://cebus.net/de/age.php>: <https://agify.io/> or <https://github.com/JasonKessler/agefromname>.

30. Criado Perez, op. cit.

31. Ibid.

32. Ibid, citant une étude de Desmond U. Patton, William R. Frey, Kyle A. McGregor, Fei-Tzin Lee, Kathleen McKeown et Emanuel Moss. 2020. Contextual Analysis of Social Media: The Promise and Challenge of Eliciting Context in Social Media Posts with Natural Language Processing. Actes de la conférence AAAI/ACM sur l'IA, l'éthique et la société (New York, NY, États-Unis) (AIES '20). Association for Computing Machinery, New York, NY, USA, 337-342. <https://doi.org/10.1145/3375627.3375841>.

silence et donc de restreindre injustement leur capacité à communiquer sur les réseaux sociaux.

Les chercheurs et chercheuses considèrent également qu'il existe un « **biais d'évaluation** », qui se produit lors de l'évaluation d'un modèle, si les données de référence (utilisées pour comparer le modèle à d'autres modèles qui effectuent des tâches similaires) ne représentent pas la population que le modèle analysera³³. Par exemple, les autrices de l'article « Gender Shades » ont découvert que deux ensembles de données de référence largement utilisés pour l'analyse faciale (IJB-A et Adience) étaient principalement composés de sujets à la peau claire (79,6 % et 86,2 %, respectivement)³⁴.

Enfin, le « **biais de déploiement** » concerne l'utilisation de modèles dans le monde réel, en particulier lorsqu'un modèle conçu pour résoudre un problème est utilisé pour une autre tâche³⁵. Cela peut se produire, par exemple, en raison d'un changement de stratégie marketing. En outre, un modèle fait souvent partie d'un système sociotechnique complexe où l'homme et les machines interagissent. Dans un environnement « réel », d'autres biais peuvent donc être introduits lorsque les êtres humains interprètent les sorties algorithmiques pour les utiliser comme données d'entrée plus loin dans la chaîne de décision prise en charge par les algorithmes³⁶.

Les **biais dits d'automatisation et de confirmation** peuvent également renforcer ces biais. Le biais d'automatisation se produit lorsque les êtres humains font davantage confiance aux machines et aux outils technologiques qu'à leur propre jugement ou à celui d'autres êtres humains, qui peut être contradictoire, et ont donc tendance à valider les résultats des algorithmes sans les remettre en question. Dans le contexte des machines prédictives, par exemple, un tel biais peut conduire à des évaluations de risques biaisées qui ne sont pas remises en question par les soi-disant « humains dans la boucle » [qui coopèrent avec l'IA] et donc à des comportements d'approbation automatique. Le biais de confirmation se produit lorsque des croyances

-
33. Voir Harini Suresh et John Guttag. 2021. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. Actes de la Conférence ACM sur l'équité et l'accès dans les algorithmes, les mécanismes et en optimisation (EAAMO'21). ACM, New York, NY (États-Unis d'Amérique), 9 pages: <https://doi.org/10.1145/3465416.3483305>.
 34. Buolamwini J and Gebru T, *Gender shades: Intersectional accuracy disparities in commercial gender classification* Conference on Fairness, Accountability and Transparency 2018).
 35. Voir *ibid.*
 36. Harini Suresh et John Guttag. 2021. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. Actes de la Conférence ACM sur l'équité et l'accès dans les algorithmes, les mécanismes et en optimisation (EAAMO'21). ACM, New York, NY (États-Unis d'Amérique), 9 pages: <https://doi.org/10.1145/3465416.3483305>.

préexistantes influencent le traitement de nouvelles informations, lesquelles sont d'autant mieux prises en compte qu'elles sont cohérentes avec ces croyances ou interprétées en cohérence avec ces croyances. Dans le contexte de l'IA, les stéréotypes de genre pourraient ainsi servir de prisme de renforcement aux décideurs/décideuses humain-es lorsqu'ils/elles interprètent des résultats algorithmiques biaisés. Dans une expérience, Green et Chen montrent également que les interprètes humains des évaluations automatisées des risques fournies par un algorithme produisent des « **interactions disparates** », c'est-à-dire que les interprétations d'évaluations des risques algorithmiques similaires sont plus clémentes envers les prévenu-esblanc-hes que les prévenu-es noir-es³⁷.

D'autres taxonomies de biais ont été proposées. Par exemple, Barocas et Selbst repèrent les moments et les situations clés où le biais est canalisé dans les systèmes d'IA : la **définition des « variables cibles »** (la caractéristique à mesurer ou à prédire par un modèle, par exemple la performance professionnelle) et des « **étiquettes de classe** » (les variations possibles dans l'occurrence de la variable cible, par exemple exceptionnelle, très bonne, bonne, insatisfaisante) ; l'utilisation des « **données de formation** » (les biais qui se produisent lors de l'étiquetage et de la collecte des données) ; la « **sélection des caractéristiques** » (les attributs qui doivent être considérés pertinents par un modèle, par exemple le revenu annuel) ; et l'utilisation de « **substituts** » [indicateurs indirects] (lorsque les attributs pertinents correspondent à des groupes protégés, par exemple le revenu annuel et le genre en raison de l'écart de rémunération entre les genres)³⁸.

L'exemple récent de ChatGPT3 illustre bien toute la diversité des biais qui existent. Tout d'abord, ces modèles sont entraînés sur Wikipédia, une plateforme largement dominée par les hommes. Il faut savoir par exemple que la version anglaise de Wikipédia contient plus de 1,5 million de biographies d'écrivain-es, d'inventeurs/inventrices et d'universitaires important-es, mais que moins de 19 % de ces biographies sont celles de femmes³⁹. On peut également s'interroger sur la diversité du personnel chargé d'étiqueter les données. Enfin, les idées qui circulent sur les modèles personnalisés pouvant se substituer au modèle unique ChatGPT afin que les valeurs que nous

37. Voir Green B et Chen Y, « Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments » (2019) Actes de la conférence sur l'équité, la responsabilité et la transparence 90.

38. Barocas S et Selbst AD, "Big Data's Disparate Impact" (2016) 104 California law review, 677-693.

39. Tripodi, F. (2021). Ms. Categorized: Gender, notability, and inequality on Wikipedia. *New Media & Society*, 0(0). <https://doi.org/10.1177/14614448211023772>.

défendons soient mieux respectées soulèvent de sérieuses questions quant à l'universalité des droits humains⁴⁰.

Ces taxonomies contribuent à démystifier le mythe selon lequel les biais ne sont issus que des données, et montrent le rôle complexe des interactions sociotechniques dans la (re)production de biais discriminatoires.

3) L'impact discriminatoire de l'IA : quelques exemples concrets

Cette section illustre comment les biais peuvent donner lieu à des discriminations dans différents secteurs.

Le recrutement : Reuters a rapporté en 2018 qu'Amazon avait développé un programme s'appuyant sur l'apprentissage automatique pour repérer les meilleurs CV de candidat·es. Le programme désavantageait systématiquement le CV des femmes car il reflétait l'écart entre les sexes parmi le personnel recruté au cours des dix dernières années. La neutralisation de termes comme « femmes » n'a pas permis de corriger le résultat discriminatoire, car le système a été en mesure de déduire le sexe à partir d'autres données⁴¹.

Des chercheurs et chercheuses de l'université d'Utrecht se sont associées à une plateforme de recherche d'emploi pour étudier comment l'utilisation d'un langage genré dans la barre de recherche donne des résultats différents, avec des attributions discriminatoires d'informations sur les offres d'emploi⁴². Le moteur de recherche a pour effet non seulement de renforcer les stéréotypes sur les professions typiques des hommes et des femmes, mais aussi de causer des préjudices en termes d'allocation et de distribution.

La diffusion ciblée en ligne d'offres d'emploi grâce aux services d'optimisation proposés par des plateformes de réseaux sociaux telles que Facebook vient aussi renforcer les stéréotypes de genre ainsi que la ségrégation de genre

40. Eric Hal Schwartz, OpenAI Promises Customizable ChatGPT After Bias Complaints, 20 février 2023 : <https://voicebot.ai/2023/02/20/openai-promises-customizable-chatgpt-after-bias-complaints/>.

41. Voir Dastin J, 'Amazon scraps secret AI recruiting tool that showed bias against women' Reuters (2018) : <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G> (dernière consultation le 22 juillet 2022).

42. Voir van Es K, Everts D et Muis I, "Gendered language and employment Web sites: How search algorithms can cause allocative harm" (2021) 26 First Monday : <https://journals.uic.edu/ojs/index.php/fm/article/view/11717/10200>.

sur le lieu de travail⁴³. Une expérience menée par AlgorithmWatch en 2020 a montré qu'en demandant à Facebook de diffuser des annonces « de manière neutre » (sans cibler un public spécifique), une annonce pour un poste de conduite des camions a été présentée à un public composé de 93 % d'hommes et 7 % de femmes⁴⁴. À l'inverse, une annonce pour un poste dans l'éducation a été diffusée à un public composé de 96 % de femmes et 4 % d'hommes⁴⁵.

Les systèmes de reconnaissance des visages et d'analyse des émotions intégrant une IA peuvent également donner lieu à une discrimination raciale ou désavantager les candidat-es avec un handicap⁴⁶. Cette discrimination s'explique par les taux de performance plus faibles de ces appareils sur les teintes de peau plus foncées, en particulier chez les femmes⁴⁷. En outre, les logiciels d'analyse des émotions qui sont formés à partir de sujets neurotypiques pourraient ne pas être en mesure de fonctionner correctement sur des sujets neurodivers. L'analyse des émotions par l'IA étant de plus en plus utilisée dans le secteur du recrutement, par exemple pour analyser les enregistrements vidéo des présentations des candidat-es, cette différenciation pourrait poser des problèmes d'accessibilité et d'inclusion.

L'accès aux biens et services, aux banques et aux assurances: en Finlande, le tribunal national pour l'égalité et la non-discrimination a conclu à une discrimination multiple directe dans une affaire où le requérant s'était vu refuser un prêt en ligne. Après avoir enquêté sur cette affaire, l'organisme de promotion de l'égalité de traitement (le médiateur chargé de la non-discrimination) a constaté que la société utilisait des modèles statistiques pour évaluer la solvabilité de la personne faisant la demande en fonction

43. Voir Ali M et autres, « Discrimination through optimization : How Facebook's Ad delivery can lead to partial outcomes » (2019) 3. Actes de l'ACM sur l'interaction homme-ordinateur 1.

44. 4 864 hommes, mais seulement 386 femmes. Voir Wulf J, Automated Decision-Making Systems and Discrimination : Understanding causes, recognition cases, supporting those affected (AlgorithmWatch 2022), 7 : https://algorithmwatch.org/en/wp-content/uploads/2022/07/AutoCheck-Guidebook_ADM_Discrimination_EN-AlgorithmWatch_June_2022_b.pdf et Kayser-Bril N, 'Automated Discrimination : Facebook uses gross stereotypes to optimize ad delivery' AlgorithmWatch : <https://algorithmwatch.org/en/automated-discrimination-face-book-google/> (dernière consultation le 22 juillet 2022).

45. Ibid. 6 456 femmes, mais seulement 258 hommes.

46. Voir Buolamwini J et Gebru T, Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification (Proceedings of Machine Learning Research 2018) ; Hannah Devlin, « AI systems claiming to 'read' emotions pose discrimination risks » (16 février 2020), *The Guardian* : <https://www.theguardian.com/technology/2020/feb/16/ai-systems-claiming-to-read-emotions-pose-discrimination-risks> (dernière consultation le 22 juillet 2022).

47. Voir Buolamwini J et Gebru T, Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification (Proceedings of Machine Learning Research 2018).

de son âge, de son sexe, de sa langue et de son lieu de résidence, sans tenir compte de ses antécédents réels en matière de crédit. Dans cette affaire, le fait que le demandeur soit un homme, qu'il parle finnois et qu'il soit originaire d'une zone rurale a été considéré comme un facteur de désavantage dans l'évaluation effectuée par l'institution financière⁴⁸.

Une histoire similaire a été rapportée en Allemagne, où une cliente s'est vu refuser un crédit lors d'un achat en ligne. En enquêtant sur les raisons du rejet auprès de l'établissement de crédit, la cliente a appris qu'une combinaison de son âge et de son sexe semblait avoir motivé le rejet automatisé, sur la base de stéréotypes intersectionnels préjudiciables selon lesquels les femmes autour de 40 ans sont souvent divorcées et ont donc moins de pouvoir d'achat⁴⁹.

Dans le secteur des assurances, une étude menée par les universités de Padoue, d'Udine et de Carnegie Mellon a montré que des facteurs tels que le lieu de naissance et la citoyenneté influencent le prix des polices d'assurance automobile payé par les client-es⁵⁰. Dans une étude de cas, les auteu-rices ont montré que le fait d'indiquer le Ghana comme lieu de naissance dans une demande pouvait entraîner une augmentation de prix de 1 000 EUR par rapport à une personne déclarant être née en Italie.

Une autre étude d'AlgorithmWatch a montré que la discrimination numérique s'étend bien au-delà de l'IA⁵¹. Les formulaires en ligne peuvent entraîner une discrimination fondée sur la race, l'origine ethnique ou la nationalité, par exemple lorsqu'ils n'autorisent l'enregistrement que de patronymes contenant trois lettres ou plus. Les personnes dont le nom est plus court se verront refuser l'inscription ou ne pourront pas ouvrir de compte, ce qui est souvent une condition préalable à l'achat de biens et de services en ligne.

48. Voir : Lorenz Matzat and Minna Ruckenstein, "Finnish Credit Score Ruling raises Questions about Discrimination and how to avoid it" (21 November 2018) *AlgorithmWatch* : <https://algorithmwatch.org/en/finnish-credit-score-ruling-raises-questions-about-discrimination-and-how-to-avoid-it/> (dernière consultation le 22 juillet 2022); Rainer Hiltunen, "Multiple discrimination in assessing creditworthiness" (1 August 2018), European network of legal experts in gender equality and non-discrimination: <https://www.equalitylaw.eu/downloads/4658-finland-multiple-discrimination-in-assessing-creditworthiness-pdf-120-kb> (dernière consultation le 22 July 2022).

49. Voir Wulf J, Automated Decision-Making Systems and Discrimination: Understanding causes, recognition cases, supporting those affected (AlgorithmWatch 2022), 6-7

50. L'étude a été citée par AlgorithmWatch, voir *ibid*.

51. Lulamae, Josephine, «Fixing Online Forms Shouldn't Wait Until Retirement», AlgorithmWatch (13 janvier 2022): <https://algorithmwatch.org/en/unding-online-forms/> (dernière consultation le 22 juillet 2022).

L'évaluation des risques dans le domaine de la sécurité, de la prévention du crime, du maintien de l'ordre et du système judiciaire: en Espagne, le logiciel VioGén a été utilisé pour évaluer les risques de violence sexiste et de féminicide par des partenaires intimes. Malgré une évaluation globalement favorable, les critiques portent sur plusieurs cas de faux négatifs où des évaluations indiquant des risques faibles ont conduit à la mise en œuvre de moyens de prévention insuffisants ayant débouché sur des conséquences tragiques⁵².

Les Pays-Bas ont déployé plusieurs systèmes prédictifs à des fins de prévention de la criminalité, des systèmes qui ont été sévèrement critiqués car ils ont créé une discrimination fondée sur la race, l'ethnicité et la nationalité. Par exemple, une enquête menée en 2020 par Amnesty International a révélé que le « Sensing Project », qui visait à prévenir localement le vol à l'étalage et le vol à la tire, aboutissait à un profilage ethnique discriminatoire de personnes d'origine est-européenne, et en particulier de membres de la communauté rom⁵³. En observant le trafic automobile dans et autour de la zone de déploiement, le système a utilisé l'origine est-européenne des passager-es comme facteur de risque prédictif de criminalité. D'autres systèmes d'anticipation de la criminalité, par exemple à Amsterdam, utiliseraient des facteurs tels que « le nombre de foyers monoparentaux », « le nombre de bénéficiaires des prestations sociales » et « le nombre d'immigrant-es non occidentaux » pour déterminer les « points chauds » de la criminalité dans tout le pays⁵⁴.

Dans les aéroports, les technologies de contrôle de la sécurité et de contrôle aux frontières utilisant des systèmes automatisés de reconnaissance du genre se sont avérées discriminatoires à l'égard des personnes transgenres, intersexes, non binaires et non-conformes aux normes de genre, car elles reposent sur un système de classification binaire du genre qui ne rend pas compte de la complexité réelle de l'identité et de l'expression de genre⁵⁵.

-
52. Michele Catanzaro, "In Spain, the VioGén algorithm attempts to forecast gender violence", AlgorithmWatch (27 April 2020) : <https://algorithmwatch.org/en/viogen-algorithm-gender-violence/> ((dernière consultation le 22 juillet 2022).
 53. Amnesty International « We Sense Trouble: Automated Discrimination and Mass Surveillance in Predictive Policing in the Netherlands » (2020), 5 : https://www.amnesty.nl/content/uploads/2020/09/Report-Predictive-Policing-RM-7.0-FINAL-TEXT_CK-2.pdf (dernière consultation le 22 juillet 2022).
 54. <https://www.vice.com/en/article/5dpmd/the-netherlands-is-becoming-a-predictive-policing-hot-spot>
 55. Voir JD Shadel, « #TravelingWhileTrans : The trauma of returning to 'normal' » (The Washington Post, 2021) : [https://www.washingtonpost.com/travel/2021/06/16/trans-travel-tsa-lgbtq-and-quinan-c-l-and-mina-hunt-«Biometric Bordering and Automatic Gender Recognition: Challenging Binary Gender Norms in Everyday Biometric Technologies.»Communication, Culture and Critique 15.2 \(2022\) : 211-226.](https://www.washingtonpost.com/travel/2021/06/16/trans-travel-tsa-lgbtq-and-quinan-c-l-and-mina-hunt-«Biometric Bordering and Automatic Gender Recognition: Challenging Binary Gender Norms in Everyday Biometric Technologies.»Communication, Culture and Critique 15.2 (2022) : 211-226.)

La reconnaissance faciale est de plus en plus utilisée pour la détection et la prévention des crimes. Par exemple, les services de répression peuvent l'utiliser pour comparer les photos de suspects aux photos d'identité judiciaire et de permis de conduire. Si « les algorithmes de reconnaissance des visages se targuent d'une grande précision en matière de classification (plus de 90 %) », ces résultats ne sont pas universels⁵⁶. En 2018, le projet Gender Shades a révélé des divergences dans la précision de cette technologie et sa capacité à reconnaître les différentes teintes de peau et les différents sexes. Les algorithmes ont systématiquement démontré que la précision la plus faible concernait les femmes à la peau foncée et que la plus élevée s'appliquait aux hommes à la peau claire⁵⁷. Dans un contexte de justice pénale, les technologies de reconnaissance faciale, dont l'exactitude est intrinsèquement biaisée, peuvent mal identifier des personnes suspectes et même conduire à l'incarcération de personnes de couleur innocentes, comme cela s'est produit aux États-Unis⁵⁸. Il est donc inquiétant de constater que, même si elle est précise, la reconnaissance faciale donne des moyens importants aux systèmes répressifs qui ont une longue histoire de surveillance raciste et anti-militante et peut aggraver les inégalités préexistantes⁵⁹ ».

L'accès aux services publics et administratifs : l'utilisation de technologies de reconnaissance faciale dans les services publics ou en association avec eux peut conduire à exclure ou à refuser des prestations aux utilisateur·rices finaux/finales. Par exemple, un photomaton de l'Office d'État des transports de Hambourg, en Allemagne, n'a pas reconnu le visage d'une requérante aux fins de prendre une photo biométrique, ce qui était nécessaire pour sa demande administrative. Bien que l'office public ait nié que l'échec provenait du logiciel de reconnaissance faciale utilisé, un employé local a indiqué

-
56. Alex Najibi, Racial Discrimination in Face Recognition Technology, Université de Harvard, octobre 2020 : [https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/#:~:text=Face recognition algorithms boast high,and 18-30 years old.](https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/#:~:text=Face%20recognition%20algorithms%20boast%20high,and%2018-30%20years%20old.)
 57. Gender Shades Project, voir : <http://gendershades.org/overview.html> (dernière consultation le 31 août 2022)
 58. RACE AND WRONGFUL CONVICTIONS IN THE UNITED STATES: https://www.law.umich.edu/special/exoneration/Documents/Race_and_Wrongful_Convictions.pdf (dernière consultation le 31 août 2022).
 59. Alex Najibi, Racial Discrimination in Face Recognition Technology, Harvard University, October 2020.

que les échecs ont souvent lieu en relation avec la couleur de peau des candidat-es⁶⁰.

Aux Pays-Bas, le déploiement du système SyRi (System Risk Indication) utilisé pour détecter les fraudes à l'aide sociale, s'est avéré à l'origine de discriminations fondées sur le revenu et l'origine ethnique avant d'être interrompu par une décision de justice en 2020⁶¹. En 2021, un scandale lié à l'aide sociale a contraint le gouvernement néerlandais à démissionner après que plus de 20 000 parents, considérés par un système d'IA comme ayant fraudé aux allocations familiales, ont fait l'objet d'une enquête par les autorités fiscales néerlandaises⁶². Le système AI a traité la double nationalité comme un facteur de risque élevé, ce qui a entraîné un nombre disproportionné d'enquêtes et de procédures judiciaires à l'encontre de familles issues de l'immigration, dont les allocations familiales ont été suspendues et dont certaines ont dû rembourser les prestations perçues⁶³. L'affaire montre également comment le manque de responsabilité et de transparence autour de l'utilisation de ces systèmes peut conduire à priver ceux qui subissent des décisions de l'IA d'une explication ou de la possibilité de faire appel de ces décisions.

L'éducation : on sait que les logiciels de reconnaissance faciale peuvent être biaisés et entraîner une discrimination intersectionnelle fondée sur la race et le genre⁶⁴. Utilisés dans des logiciels de surveillance installés dans des établissements d'enseignement, ils peuvent avoir une incidence négative sur les conditions dans lesquelles les étudiant-es racialisé-es passent les examens et même leur capacité à le faire. Par exemple, le logiciel de surveillance

-
60. Voir Wulf J, Automated Decision-Making Systems and Discrimination : Understanding causes, recognition cases, supporting t'ose affected (AlgorithmWatch 2022), p. 8. Cette hypothèse est corroborée par des études indiquant qu'il existe une discrimination intersectionnelle fondée sur le genre et la couleur de la peau dans les logiciels de reconnaissance faciale, par exemple Buolamwini J et Gebru T, Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification (Proceedings of Machine Learning Research 2018).
 61. Koen Vervloesem, «How Dutch activists got an invasive fraud detection algorithm banned», AlgorithmWatch (6 avril 2020): <https://algorithmwatch.org/en/syri-netherlands-algorithm/> (dernière consultation le 22 juillet 2022).
 62. Nadia Benaissahet systeem doet precies wat het wordt opgedragen" (29 janvier 2021) *Bits of Freedom*: <https://www.bitsoffreedom.nl/2021/01/29/het-systeem-doeet-precies-wat-het-wordt-opgedragen/>.
 63. Jon Henley, "Dutch government faces collapse over child benefits scandal " (14 January 2021) *The Guardian*: <https://www.theguardian.com/world/2021/jan/14/dutch-government-faces-collapse-over-child-benefits-scandal> and Björn ten Seldam & Alex Brenninkmeijer, "The Dutch benefits scandal: a cautionary tale for algorithmic enforcement" (30 April 2021) *EU Law Enforcement*: <https://eulawenforcement.com/?p=7941>.
 64. Buolamwini J and Gebru T, Gender shades: Intersectional accuracy disparities in commercial gender classification (Conference on Fairness, Accountability and Transparency 2018).

utilisé par plusieurs universités aux Pays-Bas avait du mal à reconnaître les étudiant-es à la peau foncée⁶⁵. L'université n'ayant pas pris sa plainte au sérieux, une étudiante soutenue par le Centre Racisme et Technologie a déposé une plainte officielle auprès de l'Institut des droits humains, l'autorité nationale chargée de la non-discrimination dans le pays⁶⁶. Les logiciels de surveillance peuvent également avoir un impact négatif sur les étudiant-es en situation de handicap, par exemple en créant de l'anxiété, en n'autorisant pas la présence d'une personne accompagnante ou en ne permettant pas aux étudiant-es de faire des pauses en s'éloignant de l'ordinateur⁶⁷. Pour les familles à faible revenu qui partagent des chambres en raison du manque d'espace, l'utilisation d'un logiciel de surveillance peut créer un désavantage en signalant un « comportement aberrant » lorsque les membres de la famille sont identifiés en train de passer derrière l'écran⁶⁸.

Les soins de santé : Criado Perez a révélé comment les secteurs de la recherche et des soins de santé s'appuient sur des modèles masculins pour évaluer les risques et l'efficacité des médicaments et finissent par produire des données sur la santé qui sont de moindre qualité pour les femmes et les personnes de genre divers. Ce manque de données sur le genre dans le secteur des soins de santé conduit à des systèmes prédictifs moins fiables lorsqu'il s'agit de diagnostiquer les femmes et les patient-es de genre divers⁶⁹. Des études montrent que le manque de données dans le domaine de la santé touche également d'autres groupes minoritaires⁷⁰.

-
65. Racism and Technology Centre, « Student stapt naar College voor de Rechten van de Mens vanwege gebruik racistische software door de VU » (15 juillet 2022) : <https://racismandtechnology.center/2022/07/15/student-stapt-naar-college-voor-de-rechten-van-de-mens-vanwege-gebruik-racistische-software-door-de-vu/#more-1691> (dernière consultation le 28 juillet 2022).
 66. Fleur Damen, « De antispieksoftware herkende haar niet als mens omdat ze zwart is maar bij de vu vond ze geen gehoor » *De Volkskrant* (15 July 2022) : <https://www.volkskrant.nl/nieuws-achtergrond/de-antispieksoftware-herkende-haar-niet-als-mens-omdat-ze-zwart-is-maar-bij-de-vu-vond-ze-geen-gehoor~b6810279/> (dernière consultation le 27 juillet 2022). Pour la plainte, voir : <https://racismandtechnology.center/2022/07/15/student-stapt-naar-college-voor-de-rechten-van-de-mens-vanwege-gebruik-racistische-software-door-de-vu/#more-1691>
 67. Lydia X. Z. Brown, « How Automated Test Proctoring Software Discriminates Against Disabled Students » (16 novembre 2020), Centre for Democracy and Technology <https://cdt.org/insights/how-automated-test-proctoring-software-discriminates-against-disabled-students/> (dernière consultation le 28 juillet 2022).
 68. Ibid.
 69. Voir Criado Perez C, *Invisible women: Exposing data bias in a world designed for men* (Random House 2019).
 70. Ibid.

Une étude américaine réalisée par Obermeyer et al. montre comment un système utilisé pour prédire les risques sanitaires aux fins d'allocation de ressources désavantageait systématiquement les patient·es issus de minorités ethniques. En effet, le système utilisait des données sur l'accès antérieur des groupes aux soins de santé, ce qui a progressivement intégré la discrimination structurelle existante⁷¹.

Une étude publiée par l'Organisation mondiale de la santé en 2022 montre que les systèmes algorithmiques utilisés dans le secteur des soins de santé sont entraînés sur des données concernant des populations majoritairement jeunes, non représentatives des sujets plus âgés⁷². Cela réduit la qualité des prédictions pour les populations âgées et pourrait se traduire par une sous-performance disproportionnée de ces systèmes, y compris par des diagnostics erronés.

Les médias et les moteurs de recherche: des études révèlent que les représentations des femmes dans les images renvoyées par les moteurs de recherche en ligne sont biaisées et reflètent des stéréotypes sexistes, racistes et intersectionnellement discriminatoires. Par exemple, Noble montre dans une expérience avec le moteur de recherche Google comment les images de filles et de femmes noires sont sexualisées⁷³. D'autres groupes de femmes minoritaires sont aussi en butte à des stéréotypes sexualisés dans les résultats des moteurs de recherche, par exemple dans les recherches liées au mot « lesbienne »⁷⁴. Même si l'on a essayé de corriger ces biais, une étude récente portant sur les principaux moteurs de recherche indique qu'il existe

71. Voir Obermeyer Z et autres, 'Dissecting racial bias in an algorithm used to manage the health of populations' (2019) 366 Science 447.

72. J Stypinska, « AI ageism: a critical roadmap for studying age discrimination and exclusion in digitalized societies » (2022) AI & Soc, <https://doi.org/10.1007/s00146-022-01553-5>, et OMS, Ageism in artificial intelligence for health (2022), <https://www.who.int/publications/item/9789240040793>.

73. Voir, par exemple, Safiya Noble, *Algorithms of oppression: how search engines reinforce racism* (New York University Press 2018).

74. Du fait de la prédominance des conceptions misogynes concernant les femmes non hétérosexuelles, il se crée dans de nombreuses langues une association entre le mot « lesbienne » et les contenus pornographiques. Les algorithmes des moteurs de recherche reproduisent cette association et influencent le type de résultats obtenus en recherchant ces mots. En 2019, Google a ainsi dû modifier ses algorithmes pour éviter que les résultats de recherche associés au mot « lesbienne » ne génèrent que des résultats liés à des contenus pornographiques, alors que d'autres mots associés à la communauté LGBTI n'aboutissaient pas aux mêmes résultats. Voir Marie Turcan, Pourquoi le mot « lesbienne » sur Google ne renvoie-t-il que vers des sites pornographiques? (Numerama, 2019), <https://www.numerama.com/politique/478663-pourquoi-le-mot-lesbienne-sur-google-ne-renvoie-t-il-que-vers-des-sites-pornographiques.html>.

des «biais de représentation» ainsi que des «biais de face-isme» [proéminence du visage] dans la façon dont les femmes sont représentées, de sorte que «les femmes sont moins susceptibles d'être représentées dans des contenus médiatiques neutres en termes de genre [...] et leur rapport visage/corps dans les images est souvent plus faible» que pour les hommes⁷⁵. Les solutions techniques de «débaisage» peuvent traiter certains des symptômes du problème, par exemple en rééquilibrant le nombre de photos de femmes dans une recherche d'images de «PDG», mais pas ses racines, en l'occurrence les stéréotypes nuisibles, les préjugés de représentation et d'attribution ainsi que l'inégalité structurelle qui sont profondément ancrés dans notre réalité culturelle et matérielle. Par exemple, des tests récents semblent indiquer que DALLE2, qui est un outil de création d'images reposant sur l'IA, ajoute des «invites de diversité» à des requêtes non spécifiques, par exemple en ajoutant les étiquettes «noir» ou «femme» à une invite demandant au logiciel de générer une image d'un «PDG⁷⁶». Cette approche est analogue à une forme d'action positive comme les quotas. On peut lui reprocher de ne pas s'attaquer aux causes profondes de cette discrimination, à savoir le manque de diversité dans les kits de formation, mais s'ils sont utilisés à grande échelle, ces correctifs ont au moins le mérite de diffuser des représentations plus diversifiées qui, à long terme, peuvent contribuer à atténuer les stéréotypes préjudiciables.

La violence fondée sur le genre en ligne, le discours de haine et le harcèlement: la discrimination numérique prend également la forme de violences fondées sur le genre, par exemple lorsque des vidéos manipulées (deepfake) sont utilisées pour harceler les femmes dans le contexte dit du «porno de vengeance». La diffusion non consentie de contenus à caractère sexuel, souvent sous forme d'images, a également été considérée comme une forme de violence fondée sur le genre qui touche particulièrement les femmes et les filles jeunes ou les femmes qui sont des personnalités publiques telles que les journalistes, les défenseuses des droits humains ou

75. Ulloa R. et al., «Representativeness and face-ism: Gender bias in image search» (2022), *New Media & Society*, vol.

76. Matthew Sparkes, "AI art tool DALL-E 2 adds 'black' or 'female' to some image prompts" (22 juillet 2022) *New Scientist*, voir: <https://www.newscientist.com/article/2329690-ai-art-tool-dall-e-2-adds-black-or-female-to-some-image-prompts/> (dernière consultation le 28 juillet 2022); voir également OpenAI, "Reducing Bias and Improving Safety in DALL-E 2" (18 juillet 2022): <https://openai.com/blog/reducing-bias-and-improving-safety-in-dall-e-2/> (dernière consultation le 28 juillet 2022).

les femmes politiques⁷⁷. En outre, il a été mis en évidence que le discours de haine sexiste et d'autres formes de haine en ligne sont subordonnés à l'utilisation croissante des plateformes de médias sociaux⁷⁸. Dans le même temps, la modération des contenus touche particulièrement les groupes minoritaires, qui risquent d'être réduits au silence⁷⁹ tout en faisant l'objet de campagnes de haine. Par exemple, le fait d'associer de manière systématique des mots liés à la communauté lesbienne (comme « lesbienne ») à des contenus pornographiques entraîne souvent un « bannissement furtif » (*shadow ban* en anglais) qui restreint la visibilité des messages publiés sur les médias sociaux, ou rend purement et simplement impossible l'utilisation de certains mots dans les noms de compte et les pseudonymes. L'effet de « musellement » de la modération de contenus a de fortes répercussions négatives sur la visibilité des organisations qui s'attachent à contrer les récits haineux et discriminatoires ciblant ces communautés minoritaires, ainsi que la portée des activités et des événements qu'elles mettent en place à cette fin⁸⁰.

Les stéréotypes de genre dans tous les domaines : un récent rapport de l'ONU, intitulé « I'd blush if I could: closing gender divides in digital skills through education » (Je rougirais si je pouvais : s'appuyer sur l'éducation pour combler les écarts de genre en matière de compétences numériques), a révélé que les assistants numériques intégrant une IA et dotés de voix féminines peuvent renforcer les préjugés sexistes existants. Cette tendance vers les assistantes virtuelles à voix féminine « semble plutôt associée à la notion d'assistance qu'aux questions de son, ton, syntaxe et cadence⁸¹ ». Peut-être choisit-on une voix féminine pour séduire l'utilisateur-riche et lui faire croire que l'IA est malléable et inoffensive. Mais, finalement, l'effet produit relève

77. Voir Sara De Vido et Lorena Sosa, *Criminalisation of gender-based violence against women in European States, including ICT-facilitated violence* (Réseau européen des experts juridiques en matière d'égalité des sexes et de non-discrimination 2021) : <https://www.equalitylaw.eu/downloads/5535-criminalisation-of-gender-based-violence-against-women-in-european-states-including-ict-facilitated-violence-1-97-mb> (dernière consultation le 23 juillet 2022).

78. Voir Bartoletti, Ivana. Chapter 3: Algorithms and the Rise of Populism in *An artificial revolution: On power, politics and AI*. Black Spot Books, 2020.

79. Voir Rachel Griffin, "The Sanitized Platform" (2022) 13 J Intell Prop Info Tech & Elec Com L 36.

80. L'Eurolasian Lesbian Community (EL**C*), une organisation créée en 2017 pour défendre les droits des femmes LGBTI, indique ainsi qu'elle n'a pas pu utiliser le mot « lesbian » dans son nom d'utilisateur sur Facebook, alors que d'autres mots associés à la communauté LGBTI (tels que « gay » ou « queer ») pouvaient être employés. Voir EL**C*, « Lesbophobia: An intersectional form of violence » (2021), <https://europeanlesbianconference.org/wp-content/uploads/2021/10/Lesbophobia-3.pdf>.

81. UNESCO, *I'd blush if I could: closing gender divides in digital skills through education*, 100 : <https://unesdoc.unesco.org/ark:/48223/pf0000367416.page=1>

de la « normalisation de cette nouvelle servitude numérique qui a pris place dans nos maisons et nos vies quotidiennes à travers Alexa, Siri et Cortana⁸² ».

4) En quoi la discrimination algorithmique est-elle différente d'autres formes de discrimination ?

La discrimination relayée par des technologies algorithmiques présente des **problèmes spécifiques**.

Premièrement, le déploiement de systèmes algorithmiques dans les processus de prise de décision a des **répercussions de grande ampleur sur la société**. À cet égard, il existe un écart considérable, par exemple, entre un-e employé-e de banque qui attribue inconsciemment un taux d'intérêt hypothécaire plus élevé à une personne issue d'une minorité et un logiciel qui traite des milliers de fichiers par jour et qui peut généraliser ce biais à toute personne dont le nom a une consonance africaine.

Deuxièmement, le comportement humain est contrôlé par des mécanismes sociaux et juridiques qui, sans être parfaits, sont néanmoins censés corriger les mauvais comportements à court et à long terme. En revanche, le **déploiement des technologies algorithmiques compromet souvent les principes de responsabilité et de transparence, ainsi que le contrôle des processus décisionnels**. Alors que les décisions humaines erronées peuvent en principe faire l'objet d'un recours, il est ainsi très difficile avec les systèmes d'IA de mettre en œuvre l'obligation de rendre des comptes, du fait de l'absence d'information sur le déploiement de l'IA, de l'opacité des systèmes utilisés et de la réticence des fournisseurs à soumettre à un examen public ces processus de prise de décision assistés par algorithmes⁸³. Ces obstacles rendent par ailleurs difficile la collecte de preuves de discrimination algorithmique.

Troisièmement, les **sources de la discrimination algorithmique sont difficiles à détecter**. En raison de la complexité de ces systèmes sociotechniques et du fait qu'ils sont en constante évolution, le biais peut se produire à n'importe quelle étape du pipeline algorithmique. En outre, les **algorithmes peuvent être brevetés, complexes et difficiles à comprendre**. Parfois, il s'agit en fait d'une boîte noire contenant des procédures qui peuvent être

82. Ivana Bartoletti, *An Artificial Revolution : on Power, Politics and AI* (Indigo Press).

83. Voir Gabriele Spina Ali et Ronald Yu, *Artificial Intelligence between Transparency and Secrecy: From the EC Whitepaper to the AIA and Beyond*, *European Journal of Law and Technology*: <https://www.ejlt.org/index.php/ejlt/article/download/754/1044/3716> (dernière consultation le 16 septembre 2022).

inexplicables pour un-e chercheur/chercheuse humain-e. Cette « logicisation » (softwarisation) des biais signifie que les inégalités existantes finissent par être codées et perpétuées dans des machines obscures et protégées par la propriété intellectuelle. Cette situation est extrêmement problématique car les biais deviennent plus difficiles à identifier et à contester.

En résumé, au moins **six problèmes** se posent en matière de discrimination algorithmique fondée sur des données⁸⁴. Les décisions qui sont prises par les machines le sont à une **échelle** beaucoup **plus grande**, mais l'interaction entre les humains et les machines rend les **sources de discrimination difficiles à détecter et à traiter**. Le « **nettoyage** » des données biaisées est un **problème technique** et un exercice **dépendant du contexte**, et l'existence de substituts et de corrélations avec des groupes protégés complique encore la tâche. Le **déterminisme algorithmique** est particulièrement problématique en ce qui concerne la discrimination, car les systèmes prédictifs utilisent les **corrélations découlant d'une discrimination historique** (par exemple, l'écart de rémunération entre les hommes et les femmes) **comme bases quasi-« causales » pour la prise de décision**, créant ainsi des boucles de rétroaction. Dans le même temps, l'IA et les systèmes algorithmiques sont souvent **opaques** et difficiles à expliquer, et l'**attribution de la responsabilité de la discrimination n'est pas claire**.

Ces biais, qui ne sont pas *in fine* technologiques, ne peuvent pas être corrigés par les seuls moyens technologiques. Au contraire, la lutte contre la discrimination algorithmique et les désavantages résultant de données biaisées exige un degré d'analyse beaucoup plus élevé et **des décisions politiques positives visant à empêcher activement le renforcement des inégalités structurelles ancrées dans les données sociales**. Par exemple, pour éviter d'« automatiser » les stéréotypes de genre et l'écart de rémunération entre les hommes et les femmes - le fait que les femmes ont historiquement gagné moins que les hommes - l'employeur doit prendre consciemment la décision de cibler les femmes lorsqu'il fait de la publicité pour des emplois mieux rémunérés, qui sont généralement des emplois « masculins » ou de gestion en ligne. En effet, le risque de confier leur distribution à des algorithmes d'optimisation risque de reproduire les stéréotypes de genre et

84. Voir Gerards J et Xenidis R, *Algorithmic discrimination in Europe: Challenges and Opportunities for EU Gender Equality and Non-Discrimination Law* (Réseau européen d'experts juridiques dans le domaine de l'égalité des genres et de la non-discrimination / Commission européenne, 2021).

les inégalités salariales⁸⁵. Pour comprendre les biais algorithmiques, il faut commencer par prendre conscience de la façon dont les technologies algorithmiques aggravent, renforcent et perpétuent les inégalités existantes lorsqu'aucune protection n'est mise en place. Pour ces raisons, la **lutte contre la discrimination algorithmique nécessite une approche multiforme englobant diverses disciplines** telles que les sciences sociales, l'éthique et le droit, ainsi que des **domaines réglementaires**, notamment la législation sur la non-discrimination, la protection des consommateur·ices, la protection des données, le commerce, etc.

5) La lutte contre la discrimination algorithmique : les meilleures pratiques et leurs limites

Pour faire face aux risques discriminatoires que représentent les technologies algorithmiques, le secteur a pris diverses initiatives qui vont des **solutions techniques pour « débiaiser » et « auditer » les systèmes algorithmiques** à l'adoption de **codes de conduite volontaires, d'instruments d'IA éthique** et d'autres formes d'**autorégulation**. Cette section présente quelques **exemples des pratiques de bonne gouvernance** adoptées et évalue leurs **limites**.

Les entreprises ont intensifié les étapes de la gouvernance en prévision de la réglementation à venir, d'autant que les mesures de gouvernance ex ante et ex post gagnent en popularité et en importance. Les grandes entreprises technologiques (souvent elles-mêmes touchées par des controverses sur les biais) ont mis en place des comités d'éthique, intégré la gouvernance de l'IA dans les structures de gouvernance existantes et/ou déployé des techniques de « débiaisage » pour résoudre certains problèmes.

Par exemple, Microsoft a élaboré six principes d'IA pour accélérer ce changement culturel et améliorer la sensibilisation du personnel aux questions éthiques⁸⁶. Il s'agit notamment de l'équité, de la fiabilité et de la sécurité, du respect de la vie privée et de la sécurité, de l'inclusion, de la transparence et de la responsabilité. La gouvernance est constituée de trois équipes centrales chargées de mettre en œuvre les principes fondamentaux, de gérer les politiques, la gouvernance, l'habilitation et l'utilisation sensible, et de diriger

85. Voir Ali M et autres, « Discrimination through optimization: How Facebook's Ad delivery can lead to partial outcomes » (2019) 3 Actes de l'ACM sur l'interaction homme-ordinateur 1 et Imana B, Korolova A et Heidemann J, *Audit for discrimination in algorithms delivering job ads* (2021).

86. Principes de Microsoft pour une IA responsable : <https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1%3aprimar6> (dernière consultation le 4 octobre 2022).

la mise en œuvre des processus d'IA responsable dans l'adoption des systèmes et des outils.

IBM a développé et mis en œuvre AI Fairness 360⁸⁷, une boîte à outils open-source utilisée pour examiner, signaler et atténuer la discrimination et les biais dans les modèles d'apprentissage automatisé. Les principaux objectifs de cette boîte à outils sont de faire en sorte que les algorithmes de recherche sur l'équité soient progressivement utilisés dans un cadre industriel et de fournir un cadre commun aux chercheurs et chercheuses sur l'équité qui pourront ainsi partager et évaluer les algorithmes.

Amazon a intégré de nouveaux outils d'aide à la détection des discriminations dans les technologies d'intelligence artificielle et d'apprentissage automatisé. Dans le cadre de l'offre d'informatique en nuage (cloud) Amazon Web Services, un nouveau test a été mis en place ainsi qu'un ensemble plus large d'éléments destinés aux client-es qui cherchent à développer une IA équitable et non biaisée sur la plateforme. Le test a été mis au point par Wachter, Brent Mittelstadt et Russell de l'Oxford Internet Institute de l'Université d'Oxford. Il s'agit du test dénommé « Disparité démographique conditionnelle », qui est un nouveau test visant à « garantir l'équité dans la modélisation algorithmique et les décisions fondées sur des données⁸⁸ ».

Les développeurs/développeuses du programme de création d'images « DALLE-2 », fondé sur l'IA, ont mis en œuvre une technique d'atténuation des biais après avoir constaté que les images produites présentaient des défauts nuisibles de représentation. Tandis que des messages génériques tels que « PDG » et « maçons » génèrent principalement des images d'hommes, des messages génériques tels qu'hôtesse de l'air » et « infirmière » produisent des images représentant presque exclusivement des femmes⁸⁹. Les développeurs/développeuses reconnaissent que ces stéréotypes peuvent être nuisibles, par exemple lorsqu'ils portent atteinte à la dignité de groupes protégés, les excluent de situations socialement valorisées et renforcent les représentations mentales de rôles sociaux ségrégués⁹⁰. Les images

87. IBM, introducing AI Fairness 360: <https://www.ibm.com/blogs/research/2018/09/ai-fairness-360/> (dernière consultation le 4 octobre 2022).

88. *AI modelling tool developed by Oxford academic incorporated into Amazon anti-bias software*, Oxford Internet Institute, 21 April 2021: <https://www.oii.ox.ac.uk/news/releases/ai-modelling-tool-developed-by-oxford-academics-incorporated-into-amazon-anti-bias-software-2/> (dernière consultation le 29 septembre 2022)

89. Voir OpenAI, "Reducing Bias and Improving Safety in DALL-E 2" (18 July 2022): <https://openai.com/blog/reducing-bias-and-improving-safety-in-dall-e-2/>.

90. Voir Pamela Mishkin et al, "DALL-E 2 Preview - Risks and Limitations" (2022): <https://github.com/openai/dalle-2-preview/blob/main/system-card.md#bias-and-representation>.

stéréotypées contribuent à leur tour à confirmer les préjugés sociétaux et à alimenter les préjugés en matière d'allocation, en influençant la distribution de biens sociaux essentiels. La technique d'atténuation mise en œuvre par les développeurs/développeuses de DALL-2 semble augmenter la diversité des groupes de population représentés dans les images produites. Cette technique a néanmoins été critiquée car elle ne ferait qu'ajouter des termes liés à la diversité aux messages génériques, notamment « femmes » ou « noires » afin d'augmenter la représentativité. Certains des symptômes du biais algorithmique ont pu ainsi être traités mais pas leurs causes profondes⁹¹.

Bien qu'il s'agisse d'exemples positifs d'efforts de gouvernance existants pour lutter contre la discrimination algorithmique dans le secteur, il est important de souligner leurs limites.

Les limites des solutions techniques : débiaisage et atténuation des biais

Premièrement, les **solutions techniques de débiaisage et d'atténuation des biais** ne peuvent pas résoudre à elles seules le problème de la discrimination algorithmique. À ce sujet, Balayn et Gürses ont souligné avec force que « [I] e biais repose sur des conceptualisations du biais qui ne rendent pas compte de la complexité de la discrimination en raison des limites de la configuration de l'apprentissage automatisé⁹² ». Le débiaisage ne peut pas corriger la discrimination algorithmique de manière complète ou efficace pour deux raisons principales. D'une part, ces techniques se concentrent exclusivement sur les entrées et les sorties des systèmes d'IA **sans tenir compte du contexte dans lequel ils sont utilisés**⁹³. Elles sont essentiellement centrées sur les algorithmes et **ne tiennent pas compte des points d'interaction entre la machine et l'humain**, qui sont également une source de biais⁹⁴.

D'autre part, les **techniques de débiaisage n'ont pas encore atteint un stade de développement permettant un déploiement généralisé** : « [I]es cas d'utilisation sont limités, les conceptualisations proposées des biais peuvent simplifier à l'excès les questions de discrimination, et l'efficacité et la

91. Voir Matthew Sparkes, "AI art tool DALL-E 2 adds 'black' or 'female' to some image prompts", New Scientist (22 July 2022): <https://www.newscientist.com/article/2329690-ai-art-tool-dall-e-2-adds-black-or-female-to-some-image-prompts/>.

92. Balayn A et Gürses S, Beyond Debiasing : Regulating AI and its inequalities (European Digital Rights 2021), 51 : https://edri.org/wp-content/uploads/2021/09/EDRI_Beyond-Debiasing-Report_Online.pdf.

93. Voir *ibid*, 12, 64.

94. Voir *ibid*, 50.

facilité d'utilisation des méthodes de débiaisage et des outils d'audit restent à établir⁹⁵». L'application pratique des techniques de débiaisage pose également un problème en raison des difficultés entourant l'accès aux données sensibles ainsi que des variations contextuelles pouvant survenir en fonction des cas d'utilisation⁹⁶. Par exemple, la loi anti-discrimination peut exiger que différentes conceptions de l'équité interviennent dans différents cas d'utilisation ou à différentes étapes d'un même cas d'utilisation, ce qui est difficile à traduire en termes techniques et à mettre en cohérence.

D'où la **question de savoir ce que signifie être «équitable» pour un algorithme**. Un grand nombre d'études en informatique sont consacrées à l'équité algorithmique. Les approches de l'équité sont parfois présentées comme pouvant garantir la conformité éthique et juridique des systèmes algorithmiques. Pourtant, les **notions de «biais» et d'«équité» sont des notions techniques qui ne se recoupent pas parfaitement avec leurs équivalents éthiques et juridiques**. Dans la législation sur la discrimination, en particulier, l'interdiction des biais sera limitée à ceux qui ciblent ou ont un impact négatif sur les groupes protégés. L'élimination de ces biais à un moment donné du cycle de vie de l'IA pourrait permettre d'atteindre l'équité d'un point de vue technique sans pour autant satisfaire aux obligations légales existantes en matière d'égalité tout au long du cycle en question.

En outre, les informaticien·nes ont élaboré un **large éventail de définitions de l'équité**, dont certaines sont contradictoires. Dès lors, **en fonction de la définition, un algorithme peut être techniquement équitable sans nécessairement respecter la législation anti-discrimination**⁹⁷. D'un point de vue mathématique, il existe plusieurs façons d'obtenir un résultat équitable, qui sont toutes liées à différentes perceptions et interprétations de l'équité elle-même. Différentes conceptions de l'équité existent ainsi, allant de celle qui consiste à donner la « même chance » à toutes les personnes sans tenir compte des points de départ extrêmement différents jusqu'à celle consistant à reconnaître qu'il existe des différences entre les personnes et à donner

95. Ibid, 12, 50.

96. Voir ibid.

97. Voir la discussion sur les différentes façons de mesurer les biais et les définitions divergentes de l'équité dans l'exemple du système de prédiction du risque de récidive COMPAS: Angwin, Julia, et al. « Machine bias ». *Ethics of Data and Analytics*. Auerbach Publications, 2016. 254-264 and Jeff Larson, Surya Mattu, Lauren Kirchner and Julia Angwin, « How We Analyzed the COMPAS Recidivism Algorithm » (2016) ProPublica: <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>.

à certaines un avantage temporaire pour compenser un désavantage⁹⁸. On pourrait faire valoir, par exemple, que le fait de traiter « comme les autres » une personne issue d'une minorité qui sollicite un prêt est équitable. Or, si la personne en question présente, en raison d'un racisme historique tenace, un risque plus élevé de perdre son emploi et de se retrouver insolvable sans qu'il y ait faute de sa part, l'application de l'équité comme simple moyen d'égaliser des résultats peut conduire à un renforcement de l'inégalité car cette catégorie de personnes peut voir sa solvabilité continuer de se dégrader.

Les définitions de « l'équité en tant que critère d'exactitude » et les techniques de « débiaisage » visant à acquérir *davantage* de données et à concevoir des systèmes algorithmiques *plus* précis présentent également des limites importantes. Si les « **injustices qui influent sur l'exactitude** » en raison de problèmes liés à la représentativité des données, à la collecte des données et aux pratiques de traitement des données peuvent être résolues par des modifications des politiques en matière de données visant à accroître la précision des décisions algorithmiques⁹⁹, les biais découlant d'injustices passées requièrent des solutions différentes. Les « **injustices qui n'influencent pas sur l'exactitude** » donnent lieu à des biais qui ne peuvent pas être corrigés par des améliorations des pratiques de collecte de données¹⁰⁰. Ils expriment des faits qui sont exacts mais problématiques car résultant d'une discrimination et d'une exclusion historiques. **Seules des politiques ciblant les causes profondes et les effets de ces inégalités peuvent remédier à ce type de biais.** Par exemple, si un service du personnel souhaite automatiser le recrutement en prédisant quel-les candidat-es seront les plus performant-es, l'intégration d'un plus grand nombre de données sur les recrutements passés ne permettra pas de s'attaquer aux causes des préjugés sexistes, qui se trouvent dans la ségrégation de genre sur le marché de l'emploi, les problèmes de plafond de verre, l'écart de rémunération entre les sexes, les stéréotypes de genre, etc.

En raison de ces limites, les récits solutionnistes de débiaisage doivent être démystifiés, d'autant que le **débiaisage ne peut être qu'un élément d'une stratégie anti-discrimination plus large** par rapport aux systèmes algorithmiques. **Une telle stratégie devrait être centrée sur les droits humains et les interventions sociojuridiques, et tenir compte de l'ensemble du cycle**

98. Dans l'optique d'une égalité réelle et transformatrice, les mesures spéciales temporaires ou les actions positives fournissent un soutien spécial ou un avantage provisoire à un groupe défavorisé afin de transformer un *statu quo* inégal à long terme. Voir la discussion dans la section 3 de la présente étude.

99. Hellman, Deborah. "Big Data and Compounding Injustice." *Journal of Moral Philosophy, à paraître, Virginia Public Law and Legal Theory Research Paper 2021-27* (2021).

100. Ibid.

de déploiement des systèmes de prise de décisions algorithmiques allant de la formulation du problème à traiter, au contexte de mise en œuvre du système, de ses performances réelles et de son impact pratique. En outre, comme l'ont souligné Balayn et Gürses, les fournisseurs de services d'IA ne devraient pas bénéficier d'une grande latitude dans le choix des stratégies visant à prévenir l'impact discriminatoire de leurs systèmes¹⁰¹. Au contraire, **le contrôle démocratique et les garanties réglementaires devraient établir un cadre autour des approches acceptées en matière d'équité et de lutte contre la discrimination, en tenant pleinement compte des limites techniques et de la nécessité de s'attaquer aux causes profondes de la discrimination algorithmique.** La participation des utilisateur·ices finaux/finales directement touchés·es par ces systèmes, en particulier des groupes minoritaires, doit également être assurée. À cet égard, nous soulignons dans nos recommandations ci-dessous que cela devrait également s'appliquer aux activités de normalisation.

Les limites des audits de biais : accès aux données et normes divergentes

Deuxièmement, **l'audit de biais** a été présenté comme une autre solution pouvant traiter la discrimination algorithmique. Pourtant, des problèmes se posent en ce qui concerne l'accès aux données et les normes divergentes.

L'audit est « une série d'approches permettant de réexaminer les systèmes de traitement algorithmique ». Ces approches « peuvent prendre différentes formes, notamment la vérification des documents de gouvernance, le test des résultats d'un algorithme, voire l'inspection de son fonctionnement interne¹⁰² ». Il a été suggéré que l'audit pourrait être utilisé comme un moyen de prévention contre la mise sur le marché de systèmes algorithmiques discriminatoires¹⁰³. Toutefois, le **manque d'accès aux données sur l'égalité**, les **incertitudes liées au RGDP** sur le traitement autorisé des catégories de données sensibles et l'**absence de normes uniformément acceptées** rendent difficile l'audit des algorithmes de discrimination.

D'une part, les juristes **ne sont pas certains que le Règlement Général sur la Protection des Données (RGPD) autorise le traitement de catégories**

101. Voir *ibid*, 11.

102. Digital Regulation Cooperation Forum, "Auditing algorithms: the existing landscape, role of regulators and future outlook" (2022): <https://www.gov.uk/government/publications/findings-from-the-drcf-algorithmic-processing-workstream-spring-2022/auditing-algorithms-the-existing-landscape-role-of-regulators-and-future-outlook>.

103. Voir Kim PT, 'Auditing Algorithms for Discrimination' (2017) 166 *University of Pennsylvania Law Review Online* 189.

sensibles de données à caractère personnel à des fins de débiaisage ou, plus largement, à des **fins d’anti-discrimination**¹⁰⁴. D’autre part, le **manque de données sur l’égalité**, qui découle de pratiques souvent restrictives en matière de collecte de données sur l’égalité en Europe, pose des problèmes lorsqu’il s’agit d’identifier les inégalités dans des domaines spécifiques tels que l’accès au logement, à l’éducation, aux soins de santé, à l’emploi, etc. pour divers groupes de population protégés¹⁰⁵. Il limite l’accès à des informations précises sur la vérité du terrain et l’ampleur des inégalités structurelles dans la société.

Ce problème d’accès à des données à caractère sensible doit également être considéré dans le contexte plus large de **l’extraction et de l’exploitation des données** par les grandes entreprises technologiques. **L’accès à ces données à des fins d’audit de la discrimination et de l’anti-discrimination en général devrait donc être confié à d’autres entités, notamment des organismes de promotion de l’égalité, les inspections du travail, les OSC ayant un intérêt légitime** au sens des articles 11, 12, 13 et 14 des directives européennes sur l’égalité 2000/43/CE et 2000/78/CE, etc. La mise en place d’une **collecte de données sur l’égalité plus systématique, éthique et réglementée** dans toute l’Europe, qu’appellent de leurs vœux depuis bien longtemps de nombreux acteurs du secteur, y compris des organismes de promotion de l’égalité, serait également un progrès à des fins d’audit algorithmique. L’inspiration pourrait venir du Royaume-Uni où le gouvernement a annoncé, au titre de la politique intitulée « Data : a new direction strategy¹⁰⁶ », qu’une nouvelle condition sera introduite dans le cadre de la loi sur la protection

104. Van Bekkum, Marvin et Zuiderveen Borgesius, Frederik, Using Sensitive Data to Prevent Discrimination by AI: Does the GDPR Need a New Exception? (2022): <http://dx.doi.org/10.2139/ssrn.4104823> (dernière consultation le 28 juillet 2022).

105. Voir Commission européenne, *Analyse et examen comparatif des pratiques de collecte de données sur l’égalité dans l’Union européenne : cadre juridique et pratique dans les États membres de l’UE* (Office des publications, 2017) [en anglais uniquement]: <https://data.europa.eu/doi/10.2838/6934> (dernière consultation le 28 juillet 2022); Lilla Farkas, *Analyse et examen comparatif des pratiques de collecte de données sur l’égalité dans l’Union européenne : collecte de données dans le domaine de l’ethnicité* (Office des publications, 2020) [en anglais uniquement]: <https://data.europa.eu/doi/10.2838/447194> (dernière consultation le 28 juillet 2022); Ringelheim, Julie, «Processing Data on Racial or Ethnic Origin for Antidiscrimination Policies: How to Reconcile the Promotion of Equality with the Right to Privacy?» (2007) NYU School of Law Jean Monnet Working Paper No. 08/06: <http://dx.doi.org/10.2139/ssrn.983685> (dernière consultation le 28 juillet 2022).

106. Data: a new direction - government response to consultation, 22 June 2022: <https://www.gov.uk/government/consultations/data-a-new-direction/outcome/data-a-new-direction-government-response-to-consultation> (dernière consultation le 28 juillet 2022)

des données (2018) pour autoriser le traitement de données de catégorie spéciale pour le suivi et l'atténuation des biais algorithmiques.

En outre, il n'existe pas encore d'obligations légales ni de normes uniformes pour l'audit algorithmique. Diverses méthodologies ont été proposées¹⁰⁷. Certaines des boîtes à outils développées par les chercheurs ont été adoptées par de grandes entreprises, par exemple l'instrument « Aequitas » conçu par l'Oxford Internet Institute et adopté par Amazon¹⁰⁸. Néanmoins, l'élaboration de normes réglementaires uniformes pour l'audit algorithmique dans le domaine de la non-discrimination augmenterait considérablement la sécurité juridique pour les fournisseurs. Elle permettrait également de renforcer la confiance du public dans les systèmes algorithmiques. Enfin, des normes réglementaires uniformes en matière d'audit algorithmique inciteraient les entreprises à réaliser des audits sur la discrimination, ce qui permettrait aux victimes potentielles de disposer d'informations utiles pour évaluer l'opportunité d'intenter une action (en justice) et de fournir aux juges des éléments de preuve compréhensibles.

6) Questions de représentation et de participation : le manque de diversité et d'inclusion dans le secteur de l'IA

La sous-représentation des groupes défavorisés dans les communautés professionnelles qui contribuent au développement de l'IA est un aspect important du problème de la discrimination algorithmique. Le manque de diversité et d'inclusion dans ces communautés indique que les femmes et les groupes sous-représentés ne participent pas (suffisamment) à l'élaboration des technologies algorithmiques qui **ne prennent donc pas suffisamment en compte les besoins de ces groupes, les désavantagent ou les excluent complètement.** Une enquête publiée par le Conseil de l'Europe aux fins de la présente étude montre que **la plupart des États parties à la CEDH qui ont répondu sont conscients de la question de la diversité dans le secteur de l'IA.** Les États parties soulignent la nécessité d'orienter davantage de femmes et de personnes issues de groupes minoritaires vers les disciplines

107. Pour une analyse, voir par exemple Jack Bandy, (2021) 'Problematic Machine Behaviour: A Systematic Literature Review of Algorithm Audits' À paraître, Proceedings of the ACM (PACM) Human-Computer Interaction, CSCW'21.

108. Voir Saleiro, P, Kuester, B, Hinkson, L, London, J, Stevens, A, Anisfield, A, Rodolfa, KT, Ghani, R (2018) 'Aequitas: A Bias and Fairness Audit Toolkit.' Arxiv and Oxford Internet Institute (2021) 'AI modelling tool developed by Oxford Academics incorporated into Amazon anti-bias software': <https://www.oii.ox.ac.uk/news-events/news/ai-modelling-tool-developed-by-oxford-academics-incorporated-into-amazon-anti-bias-software-2/>.

STIM (science, technologie, ingénierie et mathématiques), qui sont considérées comme un facteur majeur contribuant à une IA discriminatoire.

Certains exemples notables de partialité de l'IA due au manque de diversité ont été exposés dans un rapport de l'AI Now Institute, fondé par Meredith Whittaker, ancienne cadre de Google, et Kate Crawford, chercheuse principale chez Microsoft Research¹⁰⁹. Il s'agit notamment de services de reconnaissance d'images qui ont classé les personnes noires comme des gorilles et de la technologie Amazon qui ne reconnaît pas les utilisateur·ices à la peau plus foncée. La thèse du rapport (qui reflète une opinion largement répandue dans la communauté universitaire, politique et de l'IA au sens large) est que de tels exemples sont dus à des « angles morts », car les développeurs/développeuses conçoivent et testent les modèles en fonction de leur propre point de vue. L'absence d'une main-d'œuvre diversifiée conduit à une perspective limitée et peut entraîner des biais qui peuvent être difficiles à détecter et à corriger avant qu'ils ne conduisent à la discrimination.

Outre le problème répandu des **biais implicites**, un groupe homogène est susceptible d'avoir une **vision tronquée influencée par des identités et des expériences similaires**. À titre d'exemple, le programme Google AI Experiments a mis au point un jeu intitulé «Quick, Draw!». Les participant·es sont invité·es à dessiner des objets du quotidien, comme des chaussures, pour former un modèle¹¹⁰. Les cinq développeurs du jeu chez Google étaient des hommes. Les premiers utilisateurs du jeu et eux-mêmes dessinaient des baskets d'homme pour représenter une chaussure. Ce jeu ne savait donc pas que les escarpins à talons hauts étaient aussi des chaussures. Il ne s'agit pas d'une erreur intentionnelle puisqu'elle émane d'un groupe représentatif dominant qui conçoit les algorithmes dans le secteur technologique. **En tant que tel, tout algorithme construit par un groupe majoritaire risque de ne pas intégrer les points de vue des groupes minoritaires marginalisés et de ne fonctionner que pour les membres du groupe dominant.**

La diversité est importante car elle permet d'adopter des approches globales pour rendre les technologies d'IA plus responsables. Elle contribue à relever les défis plus rapidement et plus clairement, car les connaissances locales et l'expérience de première ligne seront intégrées au cœur de chaque processus de décision ou de travail. Il est essentiel d'obtenir le bon dosage de

109. Sarah Myers West, Meredith Whittaker et Kate Crawford, *Discriminating Systems: Gender, Race, and Power in AI*, AI Now Institute NYU, April 2019 : <https://ainowinstitute.org/discriminating-systems.pdf> (dernière consultation : 27 juillet 2022).

110. Josh Lovejoy, *Fair Is Not the Default – Why building inclusive tech takes more than good intentions*, 15 February 2018 <https://design.google/library/fair-not-default/> (dernière consultation le 28 juillet 2022).

talents afin d'avoir suffisamment de recul pour éliminer les biais et acquérir un avantage concurrentiel. **La diversité doit donc être considérée comme une « mission essentielle » en matière d'innovation.** Elle devrait se traduire par des politiques de recrutement plus diversifiées dans les communautés éducatives et professionnelles qui contribuent au développement des systèmes d'IA et à leur utilisation. Nous avons expliqué dans la section 3 que les obligations juridiques liées à la notion d'action positive pourraient jouer un rôle majeur à cet égard. En outre, les politiques favorables à la diversité dans le recrutement éducatif et professionnel devraient être complétées par une formation adéquate.

Le rapport *AI Now* (Université de New York)¹¹¹ a identifié une « crise de la diversité » dans le secteur de l'IA, en particulier dans le secteur technologique mondial, qui est très majoritairement composé d'hommes blancs, et affirme que cette prédominance a contribué aux biais sexistes et raciaux des algorithmes. Un rapport du Forum économique mondial de 2020¹¹² a dressé un tableau tout aussi sombre : malgré les discours sur une plus grande inclusion, la représentation des femmes dans les emplois liés à la technologie a diminué de 32 % depuis 1990. Selon une étude lancée par la Commission européenne en 2016, « seules 24 femmes diplômées sur 1 000 avaient une discipline liée aux TIC dans leur bagage de compétences ». En matière d'emploi, seules six de ces filles et femmes ont finalement trouvé un emploi dans le secteur numérique¹¹³.

Une start-up canadienne a constaté que les femmes ne représentent que 12 % des chercheurs et chercheuses de premier plan dans le domaine de l'apprentissage automatisé¹¹⁴. Un autre rapport¹¹⁵ de l'université de New York - *Discriminating Systems - Gender, Race, and Power in AI* - affirme que

111. Kari Paul, 'Disastrous' lack of diversity in AI industry perpetuates bias, study finds, The Guardian, 17 April 2019: <https://www.theguardian.com/technology/2019/apr/16/artificial-intelligence-lack-diversity-new-york-university-study> (dernière consultation le 27 juillet 2022).

112. Ronit Avi et Rana El Kaliouby, *Here's why AI needs a more diverse workforce*, Forum économique mondial, 21 septembre 2020: <https://www.weforum.org/agenda/2020/09/ai-needs-diverse-workforce/> (dernière consultation le 27 juillet 2022).

113. Women in AI: Promoting inclusive participation across society, Aimee Van WYNSBERGH, Alliance européenne de l'IA <https://futurium.ec.europa.eu/en/european-ai-alliance/blog/women-ai-promoting-inclusive-participation-across-society?language=hu>: <https://futurium.ec.europa.eu/en/european-ai-alliance/blog/women-ai-promoting-inclusive-participation-across-society?language=hu> (dernière consultation le 31 août 2022).

114. Archie de Berker, Women in Machine Learning: Negar Rostamzadeh, 20 février 2018: <https://medium.com/element-ai-research-lab/women-in-machine-learning-negar-rostamzadeh-dbb58dc75e81> (dernière consultation le 31 août 2022).

115. Sarah Myers West, Meredith Whittaker et Kate Crawford, *Discriminating Systems: Gender, Race, and Power in AI*, AI Now Institute NYU, avril 2019: <https://ainowinstitute.org/discriminating-systems.pdf> (dernière consultation le 27 juillet 2022).

la discrimination dans les systèmes d'IA est associée au manque de diversité dans les équipes qui travaillent sur ces technologies. **Que l'accent soit mis sur l'atténuation des biais dans les processus d'entrée ou sur l'équité dans les résultats, la diversité et l'inclusion sont l'un des outils les plus puissants dont disposent les entreprises.** Les angles morts créés par le manque de diversité (diversité en matière d'éducation, de perspectives, de vécu et d'origines) rendent plus difficile l'anticipation des biais dans les systèmes algorithmiques et leur impact potentiel sur les différents individus et groupes.

Les groupes déjà marginalisés sont systématiquement et de manière disproportionnée plus vulnérables au risque d'être lésés par des outils de décision algorithmiques qui ne représentent pas leurs perspectives et leurs intérêts. Au-delà de l'impératif moral de prévenir la discrimination raciale et sexuelle systémique dans la conception de nouveaux outils d'IA, il existe également un impératif économique. Des recherches ont montré que « les entreprises qui se situent dans le quartile supérieur en matière de diversité de genre sont 21 % plus susceptibles de connaître une rentabilité supérieure à la moyenne, tandis que la diversité ethnique et culturelle est corrélée à une augmentation de 33 % des performances¹¹⁶ ».

116. Les cinq avantages commerciaux d'une équipe diversifiée, CMI, 3 juillet 2019 : <https://www.managers.org.uk/knowledge-and-insights/listicle/the-five-business-benefits-of-a-diverse-team/> (dernière consultation le 31 août 2022).

Section 2

Le paysage juridique et politique en Europe : forces et faiblesses

Les responsables politiques sont généralement conscient-es que l'IA offre de nombreuses opportunités mais que cette technologie comporte le risque de renforcer et de perpétuer les inégalités existantes. Dans une enquête publiée par le Conseil de l'Europe pour connaître l'opinion des membres de la GEC ou du CDADI ou des observateurs auprès de ces organes, plus de **80 % des personnes ont estimé que l'IA présentait des risques pour les droits humains ; 40 % ont déclaré qu'il existe un risque direct de discrimination fondée sur le genre.**

Plusieurs initiatives qui ont été lancées dans des gouvernements englobent plusieurs aspects, notamment la participation des femmes dans les domaines des STIM, les vidéos manipulées, la cyberintimidation et la discrimination algorithmique. Certains pays comme la **Finlande**, par exemple, ont abordé de front la question du manque de transparence des systèmes algorithmiques conduisant à la discrimination, en publiant des recommandations et des orientations pour sensibiliser au problème¹¹⁷. Les **Pays-Bas** ont adopté une « analyse d'impact des algorithmes sur les droits fondamentaux » qui comprend une « ligne directrice sur la non-discrimination dès la conception¹¹⁸ ». Le Parlement néerlandais a récemment adopté une motion

117. Automaattisessa päätöksenteossa on turvattu virkavastuu ja hyvän hallinnon toteutuminen : <https://valtioneuvosto.fi/-/10623/automaattisessa-paatoksenteossa-on-turvattavavirkavastuu-ja-hyvan-hallinnon-toteutuminen> (dernière consultation le 28 juillet 2022).

118. Ministère de l'Intérieur et des Relations au sein du Royaume, «Fundamental rights and algorithms Impact Assessment» (mars 2022): <https://www.government.nl/binaries/government/documenten/reports/2022/03/31/impact-assessment-fundamental-rights-and-algorithms/Fundamental+Droits+et+Algorithmes+Impact+Assessment.pdf>.

rendant les évaluations d'impact sur les droits humains obligatoires pour les institutions publiques utilisant des algorithmes¹¹⁹.

Le gouvernement **autrichien** a publié un plan d'action sur les « vidéos manipulées » comprenant diverses mesures pour s'attaquer au problème. En **Finlande**, Aurora AI vise à orienter les citoyens et les citoyennes, notamment les jeunes, vers les services dont ils/elles ont besoin grâce à l'intelligence artificielle. Si la capacité des jeunes à trouver les services dont ils/elles ont besoin est améliorée, cela est susceptible de promouvoir l'égalité, par exemple dans l'accès aux services ou dans la fourniture d'une assistance et d'un soutien. L'Agence **portugaise** pour la modernisation administrative (AMA) a élaboré, avec l'aide de la Commission pour la citoyenneté et l'égalité de genre et d'autres parties prenantes concernées, le « Guide pour l'utilisation de l'intelligence artificielle dans l'administration publique ». Le guide est conçu pour répondre aux préoccupations suscitées par la non-discrimination en général ainsi que la protection des droits individuels et collectifs dans le développement des systèmes algorithmiques. Le document appelle l'attention sur la fiabilité et la représentativité des données à collecter et à traiter et met l'accent sur les questions liées à l'éthique, à la justice, à la transparence, à la responsabilisation et à la compréhension des systèmes.

Pourtant, les réponses nationales **manquent, dans une large mesure, de coordination**. Alors que des législateurs comme l'Union européenne sont en train d'adopter un cadre réglementaire uniforme sur l'IA, **le Conseil de l'Europe pourrait avoir une grande influence réglementaire dans le domaine des droits humains**. Sachant que l'UE préconise une « IA centrée sur la personne », l'action réglementaire du Conseil de l'Europe pourrait favoriser une **approche distincte de l'IA fondée sur les droits humains**.

Cette section de l'étude met en évidence les **instruments juridiques qui existent au niveau du Conseil de l'Europe et qui peuvent être utilisés pour traiter les divers aspects du problème de la discrimination algorithmique, allant de la non-discrimination à la protection des données et de la vie privée en passant par les réglementations sectorielles**. Elle dresse également une brève cartographie des instruments juridiques européens existants et à venir et montre que ces deux cadres présentent des **lacunes et des incertitudes lorsqu'il s'agit de traiter la discrimination algorithmique**. Ces lacunes appellent une action réglementaire du Conseil de l'Europe, dont certains contours possibles sont mis en évidence dans la section 3.

119. Voir European Centre for Not-for-Profit Law, "Netherlands sets precedent for human rights safeguards in use of AI" (2022): <https://ecnlp.org/news/netherlands-sets-precedent-human-rights-safeguards-use-ai>.

I. La discrimination et l'égalité : les instruments juridiques et politiques et leurs limites

Cette section met en lumière les instruments juridiques existants qui fournissent un socle juridique pour lutter contre la discrimination algorithmique et les formes connexes de violence algorithmique.

1) Les instruments juridiques contraignants du Conseil de l'Europe

La Convention européenne des droits de l'homme

L'article 14 de la CEDH et l'article 1 du Protocole n° 12 énoncent une interdiction de la discrimination qui fournit une **base juridique pour interdire la discrimination algorithmique**.

L'article 14 de la CEDH interdit toute discrimination fondée sur une liste non exhaustive de caractéristiques protégées :

« La jouissance des droits et libertés reconnus dans la présente Convention doit être assurée, sans distinction aucune, fondée notamment sur le sexe, la race, la couleur, la langue, la religion, les opinions politiques ou toutes autres opinions, l'origine nationale ou sociale, l'appartenance à une minorité nationale, la fortune, la naissance ou toute autre situation. »

L'article 14 de la CEDH n'est pas une disposition autonome, ce qui signifie qu'il ne peut être invoqué que dans le cadre d'une requête fondée sur une violation présumée d'un autre droit substantiel protégé par la Convention.

Entré en vigueur en 2005, le **Protocole n° 12 à la Convention** a été ratifié à ce jour par 20 des 46 États parties à la CEDH. L'article 1^{er} énonce une interdiction générale et autonome de la discrimination :

« 1. La jouissance de tout droit prévu par la loi doit être assurée, sans discrimination aucune, fondée notamment sur le sexe, la race, la couleur, la langue, la religion, les opinions politiques ou toutes autres opinions, l'origine nationale ou sociale, l'appartenance à une minorité nationale, la fortune, la naissance ou toute autre situation.

2. Nul ne peut faire l'objet d'une discrimination de la part d'une autorité publique quelle qu'elle soit fondée notamment sur les motifs mentionnés au paragraphe 1. »

La Convention d'Istanbul

La Convention du Conseil de l'Europe sur la prévention et la lutte contre la violence à l'égard des femmes et la violence domestique (la Convention d'Istanbul) fournit une **base juridique pour interdire la violence numérique à l'égard des femmes, y compris les stéréotypes algorithmiques, et la violence en ligne telle que le cyber-harcèlement, l'intimidation et le discours de haine sexiste en ligne.**

La Convention d'Istanbul, adoptée en 2011, est entrée en vigueur en 2014 et a été ratifiée par 37 États parties. Elle reconnaît que la violence à l'égard des femmes fondée sur le genre est une forme de discrimination. Ses dispositions sont axées sur la prévention, la protection, les poursuites et le développement de politiques intégrées en matière de lutte contre la violence à l'égard des femmes et la violence domestique.

La Convention d'Istanbul offre une base juridique pour **répondre au problème de la violence à l'égard des femmes** exercée en ligne et hors ligne.¹²⁰ Elle exige que le harcèlement, le harcèlement sexuel et la violence psychologique, y compris lorsqu'ils sont commis par le biais des technologies de l'information et de la communication, soient sanctionnés. En outre, elle donne un mandat clair aux pouvoirs publics des États parties pour s'attaquer aux racines sociétales et aux manifestations en ligne de la violence fondée sur le genre. L'**article 17** sur la « participation du secteur privé et des médias » est particulièrement pertinent pour les questions de violence fondée sur le genre en ligne. Il énonce à cet égard que :

« Les Parties encouragent le secteur privé, le secteur des technologies de l'information et de la communication et les médias, dans le respect de la liberté d'expression et de leur indépendance, à participer à l'élaboration et à la mise en œuvre des politiques, ainsi qu'à mettre en place des lignes directrices et des normes d'autorégulation pour prévenir la violence à l'égard des femmes et renforcer le respect de leur dignité ».

La Convention-cadre pour la protection des minorités nationales

La Convention-cadre pour la protection des minorités nationales fournit une **base juridique pour combattre la discrimination algorithmique fondée**

120. Voir la recommandation générale n° 1 sur la dimension numérique de la violence à l'égard des femmes adoptée par le Groupe d'experts sur la lutte contre la violence à l'égard des femmes et la violence domestique (GREVIO), disponible à l'adresse suivante : <https://www.coe.int/fr/web/istanbul-convention/general-recommendation>

sur le statut de minorité nationale ainsi que la violence en ligne, notamment le discours de haine.

Entrée en vigueur en 1998, la Convention compte 39 Etats parties. Dans son **article 4**, la Convention énonce que :

« 1. Les Parties s'engagent à garantir à toute personne appartenant à une minorité nationale le droit à l'égalité devant la loi et à une égale protection de la loi. À cet égard, toute discrimination fondée sur l'appartenance à une minorité nationale est interdite.

2. Les Parties s'engagent à adopter, s'il y a lieu, des mesures adéquates en vue de promouvoir, dans tous les domaines de la vie économique, sociale, politique et culturelle, une égalité pleine et effective entre les personnes appartenant à une minorité nationale et celles appartenant à la majorité. Elles tiennent dûment compte, à cet égard, des conditions spécifiques des personnes appartenant à des minorités nationales. »

L'**article 6(2)** prévoit que *« [l]es Parties s'engagent à prendre toutes mesures appropriées pour protéger les personnes qui pourraient être victimes de menaces ou d'actes de discrimination, d'hostilité ou de violence en raison de leur identité ethnique, culturelle, linguistique ou religieuse »*.

L'**article 9** relatif à la liberté d'expression, qui énonce que *« les Parties veillent, dans le cadre de leurs systèmes juridiques, à ce que les personnes appartenant à une minorité nationale ne fassent pas l'objet de discrimination dans leur accès aux médias »*, pourrait devenir particulièrement pertinent en ce qui concerne les questions de discrimination sur les plateformes de réseaux sociaux, de harcèlement informatique et de discours de haine.

La Charte européenne des langues régionales ou minoritaires

Entrée en vigueur en 1998, la Charte a été ratifiée à ce jour par 25 pays. L'**article 7(2)** de la Charte stipule que *« [l]es Parties s'engagent à éliminer, si elles ne l'ont pas encore fait, toute distinction, exclusion, restriction ou préférence injustifiées portant sur la pratique d'une langue régionale ou minoritaire et ayant pour but de décourager ou de mettre en danger le maintien ou le développement de celle-ci. »* Là encore, **cette disposition s'étend** en principe **aux domaines algorithmiques et en ligne, où elle peut être invoquée pour lutter contre la discrimination numérique sous ses nombreuses formes**. En outre, le Comité d'expert-es de la Charte européenne des langues régionales ou minoritaires a récemment rendu publique une déclaration soulignant que les applications d'IA peuvent *« faciliter l'usage quotidien des langues régionales ou minoritaires »* et aider les autorités dans leurs actions de promotion

de ces langues conformément à la Charte, et encourageant les États à « promouvoir la prise en compte des langues régionales ou minoritaires dans la recherche et l'étude de l'IA »¹²¹.

La Charte sociale européenne

À ce jour, 43 membres du Conseil de l'Europe ont ratifié soit la Charte sociale européenne (STE n° 35), adoptée en 1961, soit la Charte sociale européenne révisée (STE n° 163), adoptée en 1996. L'article E de la Charte révisée (**Non-Discrimination**) dispose que « [I]a jouissance des droits reconnus dans la présente Charte doit être assurée sans distinction aucune fondée notamment sur la race, la couleur, le sexe, la langue, la religion, les opinions politiques ou toutes autres opinions, l'ascendance nationale ou l'origine sociale, la santé, l'appartenance à une minorité nationale, la naissance ou toute autre situation ». En outre, l'article 20 garantit le « [d]roit à l'égalité de chances et de traitement en matière d'emploi et de profession, sans discrimination fondée sur le sexe ». Les deux dispositions sont applicables dans les cas de discrimination induite par les systèmes algorithmiques, d'autant plus que la Charte sociale européenne et sa version révisée portent sur les droits sociaux fondamentaux qui concernent l'emploi et les conditions de travail, le logement, l'éducation, la santé, l'aide médicale et la protection sociale, à savoir des domaines qui sont très touchés par les nouvelles formes de gestion algorithmique.

2) Les instruments politiques pertinents du Conseil de l'Europe

Un certain nombre de **normes et d'instruments de politique générale non contraignants** complètent les dispositions juridiques contraignantes et **prennent effet lorsqu'il s'agit de traiter les effets discriminatoires de l'IA et de la prise de décision algorithmique**.

En mars 2019, la « **Recommandation sur la prévention et la lutte contre le sexisme** » élaborée par la Commission pour l'égalité de genre a été adoptée par le Conseil des ministres¹²². Elle reconnaît que « [I]nternet a donné une nouvelle dimension à l'expression et à la transmission du sexisme et en

121. Déclaration du Comité d'experts de la Charte européenne des langues régionales et minoritaires sur la promotion des langues minoritaires par l'intelligence artificielle (2022), <https://rm.coe.int/declaration-ai-fr/1680a65800>.

122. Conseil de l'Europe, Recommandation CM/Rec(2019)1 sur la prévention et la lutte contre le sexisme adoptée par le Comité des ministres du Conseil de l'Europe (27 mars 2019): <https://rm.coe.int/cm-rec-2019-1-prevention-et-lutte-contre-le-sexisme/168094d895>. [en français]

particulier du discours de haine sexiste à un large public, même si les origines du sexisme ne sont pas à chercher du côté des technologies mais dans la persistance des inégalités entre les femmes et les hommes¹²³». Elle enjoint les États membres à « [i]ntégrer une perspective d'égalité entre les femmes et les hommes dans toutes les politiques, programmes et recherches en matière d'intelligence artificielle afin d'éviter les risques de perpétuation du sexisme et des stéréotypes de genre¹²⁴ ». La recommandation prévoit également un rôle positif pour l'IA puisqu'elle demande aux États parties « d'examiner comment l'intelligence artificielle pourrait aider à combler les écarts entre les femmes et les hommes et éliminer le sexisme¹²⁵. » Elle énumère des aspects essentiels tels que la participation des femmes et des filles à l'enseignement et à la pratique des technologies informatiques, l'intégration de l'égalité des genres dans la conception d'instruments axés sur les données, la sensibilisation aux stéréotypes de genre dans le big data, la transparence et la responsabilité. En outre, la récente recommandation « **sur la lutte contre les discours de haine** », rédigée conjointement par le Comité directeur sur l'anti-discrimination, la diversité et l'inclusion (CDADI) et le Comité directeur sur les médias et la société de l'information (CDMSI), indique que « les intermédiaires d'internet devraient identifier les formes de discours de haine qui sont diffusées par leurs systèmes et y réagir dans le cadre de leur responsabilité d'entreprise¹²⁶ ».

La **Stratégie 2018-2023 du Conseil de l'Europe pour l'égalité entre les femmes et les hommes** reconnaît également que le « sexisme et la discrimination à l'égard des femmes englobent le **discours de haine sexiste en ligne** » ainsi que la violence fondée sur le genre¹²⁷. En outre, en 2021, le Groupe d'experts sur la lutte contre la violence à l'égard des femmes et la violence domestique (GREVIO), qui surveille la mise en œuvre de la [Convention d'Istanbul](#), a adopté sa recommandation générale n° 1 sur la **dimension numérique de la violence à l'égard des femmes** en 2021, qui met en lumière

123. Ibid.

124. Recommandation II.B.7, *ibid.*, p. 19.

125. Ibid.

126. [Conseil de l'Europe, Recommandation CM/Rec\(2022\)16\[1\] du Comité des Ministres aux États membres sur la lutte contre le discours de haine \(20 mai 2022\)](#), [30].

127. [Stratégie du Conseil de l'Europe pour l'égalité entre les femmes et les hommes 2018-2023 adoptée par le Comité des ministres \(mars 2018\)](#), pages 10, 16, 18 : <https://www.coe.int/fr/web/genderequality/gender-equality-strategy>.

les problèmes juridiques liés au harcèlement sexuel en ligne, à la traque et à la dimension numérique de la violence psychologique¹²⁸.

En mai 2022, le Comité des Ministres a adopté une nouvelle **Recommandation sur la lutte contre les discours de haine** rédigée conjointement par le Comité directeur pour la lutte contre la discrimination, la diversité et l'inclusion (CDADI) et le Comité directeur sur les médias et la société de l'information (CDMSI)¹²⁹. Le texte reconnaît l'existence d'«**asymétries de pouvoir entre certaines plateformes numériques et leurs utilisateur·ices**», et formule des recommandations pour lutter contre le **discours de haine en ligne en ce qui concerne les politiques relatives à la modération du contenu, au microciblage et à la publicité en ligne, à l'amplification du contenu, aux systèmes de recommandation et aux stratégies de collecte de données sous-jacentes**.

En mai 2022, le Comité des Ministres a adopté une «**Recommandation sur la protection des droits des femmes et des filles migrantes, réfugiées et demandeuses d'asile**» qui exige que des évaluations d'impact sur les droits humains soient réalisées avant l'introduction de l'IA et des systèmes de prise de décision automatisés dans le domaine de la migration et que la conception, le développement et l'application de ces systèmes soient **non discriminatoires**¹³⁰. Elle appelle également à associer les femmes réfugiées, demandeuses d'asile et migrantes, ainsi que les OSC représentatives, aux «discussions sur le développement et le déploiement des nouvelles technologies qui les concernent».

D'autres instruments tels que les **Lignes directrices du Comité des Ministres du Conseil de l'Europe sur la défense de l'égalité et la protection contre la discrimination et la haine pendant la pandémie de covid-19 et les crises similaires à l'avenir** mentionnent la nécessité de veiller à ce que «les outils numériques permettant de faire face à la crise et aux risques qui en

128. Groupe d'experts sur la lutte contre la violence à l'égard des femmes et la violence domestique, Recommandation générale n° 1 sur la dimension numérique de la violence à l'égard des femmes (20 octobre 2021) : <https://rm.coe.int/recommandation-no-du-grevio-sur-la-dimension-numerique-de-la-violence-/1680a49148#:~:text=Le%20GREVIO%20consid%C3%A8re%20que%20la,11>. (dernière consultation le 22 juillet 2022). [en français]

129. Conseil de l'Europe, Recommandation CM/Rec(2022)16 sur la lutte contre le discours de haine : https://search.coe.int/cm/Pages/result_details.aspx?ObjectId=0900001680a67951. [en français]

130. Conseil de l'Europe, Recommandation CM/Rec(2022)17 du Comité des Ministres aux États membres sur la protection des droits des femmes et des filles migrantes, réfugiées et demandeuses d'asile, [22]-[25] : <https://rm.coe.int/prems-092122-fra-2573-recommandation-cm-rec-2022-17-a5-bat-web/1680a6ef9b>.

résultent « ne soient pas discriminatoires à l'égard des personnes issues de groupes vulnérables ou ne portent pas atteinte de toute autre manière à leurs droits¹³¹ ».

Ensemble, **ces recommandations abordent** un certain nombre de **problèmes contribuant à la discrimination algorithmique** (voir plus haut dans cette étude), à savoir : **le manque de diversité, de représentation et de participation équitables dans les domaines éducatifs et professionnels liés au secteur de l'IA, l'absence d'obligation contraignant les concepteurs à intégrer les préoccupations liées à l'égalité dans le développement des systèmes algorithmiques et l'absence de mécanismes de responsabilité clairement définis.**

Le Conseil de l'Europe a de plus adopté des instruments politiques concernant spécifiquement les droits humains dans l'espace numérique. Le Conseil des Ministres a adopté en 2020 la **Recommandation CM/Rec(2020)1 sur les impacts des systèmes algorithmiques sur les droits de l'homme**. Le texte accorde une attention spéciale à la discrimination, préconisant que les « acteurs du secteur privé qui conçoivent, développent ou mettent en œuvre des systèmes algorithmiques [se conforment] à un cadre normalisé de diligence raisonnable en matière de droits de l'homme pour éviter de favoriser ou de perpétuer la discrimination tout au long du cycle de vie de leur système », et veillent « à ce que la conception, le développement ou le déploiement en cours de leurs systèmes algorithmiques n'aient pas de conséquences discriminatoires directes ou indirectes sur les individus ou les groupes concernés, notamment sur ceux qui ont des besoins particuliers ou un handicap, ou ceux qui pourraient rencontrer des inégalités structurelles dans leur accès aux droits de l'homme »¹³². On peut aussi citer la **Recommandation CM/Rec(2022)13 du Comité des Ministres sur les effets des technologies numériques sur la liberté d'expression**, ainsi que plusieurs ensembles

131. Comité directeur pour la lutte contre la discrimination, la diversité et l'inclusion (CDADI), Lignes directrices du Comité des Ministres du Conseil de l'Europe sur la défense de l'égalité et la protection contre la discrimination et la haine pendant la pandémie de covid-19 et les crises similaires à l'avenir (2020), [27] : <https://rm.coe.int/prems-066621-fra-cdadi-lignes-directrices-a5-web-ok2-2764-3779-6356-1/1680a339c9>. [en français]

132. Comité des Ministres, Recommandation CM/Rec(2020)1 du Comité des Ministres aux États membres sur les impacts des systèmes algorithmiques sur les droits de l'homme (adoptée par le Comité des Ministres le 8 avril 2020, lors de la 1373^e réunion des Délégués des Ministres) : https://search.coe.int/cm/pages/result_details.aspx?ObjectId=09000016809e1124.

de lignes directrices, sur la reconnaissance faciale¹³³, la modération de contenu¹³⁴ et l'intelligence artificielle et la protection des données¹³⁵.

3) Un éclairage comparatif : autres dispositions européennes et internationales pertinentes

L'Union européenne dispose également d'un cadre juridique très étoffé en matière de discrimination et d'égalité. L'**article 21(1)** de la **Charte des droits fondamentaux de l'UE** interdit toute discrimination « *fondée notamment sur le sexe, la race, la couleur, les origines ethniques ou sociales, les caractéristiques génétiques, la langue, la religion ou les convictions, les opinions politiques ou toute autre opinion, l'appartenance à une minorité nationale, la fortune, la naissance, un handicap, l'âge ou l'orientation sexuelle* » et l'**article 21(2)** énonce que « [d]ans le domaine d'application des traités et sans préjudice des dispositions particulières qu'ils prévoient, est interdite toute discrimination exercée en raison de la nationalité ». L'**article 23** indique que « [l']égalité entre les femmes et les hommes doit être assurée dans tous les domaines, y compris en matière d'emploi, de travail et de rémunération » et autorise les actions positives. Dans le droit dérivé, la **directive 2000/43/CE** garantit l'égalité de traitement fondée sur la race ou l'origine ethnique au travail, dans l'accès aux biens et aux services et dans l'éducation. La **directive 2000/78/CE** interdit toute discrimination fondée sur le handicap, l'orientation sexuelle, la religion ou les convictions, l'âge sur le lieu de travail et la formation professionnelle. La **directive 2004/113/CE** garantit l'égalité entre les femmes et les hommes dans l'accès aux biens et services, tout comme la **directive 2006/54/CE** en matière d'emploi.

En 2022, la Commission européenne a publié une « **Déclaration européenne sur les droits et principes numériques pour la décennie numérique** » qui reflète la volonté de la Commission de développer une « **IA centrée**

133. Comité consultatif de la Convention pour la protection des personnes à l'égard du traitement automatisé des données à caractère personnel (Convention 108), Lignes directrices sur la reconnaissance faciale (2021) : <https://edoc.coe.int/en/artificial-intelligence/9749-lignes-directrices-sur-la-reconnaissance-faciale.html>.

134. Conseil de l'Europe, Note d'orientation : Modération de contenu. Meilleures pratiques en vue de la mise en place de cadres juridiques et procéduraux efficaces pour les mécanismes d'autorégulation et de corégulation de la modération de contenu (adoptée par le Comité directeur sur les médias et la société de l'information (CDMSI)) (2021) : <https://rm.coe.int/content-moderation-fr/1680a40294>.

135. Comité consultatif de la Convention pour la protection des personnes à l'égard du traitement automatisé des données à caractère personnel (Convention 108), Lignes directrices sur l'intelligence artificielle et la protection des données (2019) : <https://edoc.coe.int/fr/intelligence-artificielle/8252-intelligence-artificielle-et-protection-des-donnees.html>.

sur l'humain» et expose l'approche de l'UE en matière de transformation numérique. Le raisonnement de la Commission est que les droits numériques doivent garantir que les citoyens et citoyennes de l'UE ont accès aux technologies numériques et sont protégés de leurs conséquences nuisibles. Le chapitre III de la Déclaration comprend un engagement à «veiller à ce que les systèmes algorithmiques reposent sur des ensembles de données appropriés, afin d'éviter toute discrimination illicite et de permettre une surveillance humaine des résultats qui touchent des personnes¹³⁶.»

Au niveau de l'ONU, un certain nombre d'instruments protègent contre la discrimination au-delà des instruments généraux des droits humains existants: en particulier **la Convention internationale sur l'élimination de toutes les formes de discrimination raciale (CERD), la Convention sur l'élimination de toutes les formes de discrimination à l'égard des femmes (CEDAW) et la Convention relative aux droits des personnes handicapées (CRPD)**. Plus précisément, le Comité CERD a émis une **recommandation générale n° 36 relative à la prévention et la lutte contre le profilage racial par les services répressifs** en 2020. Ce document reconnaît que l'utilisation de l'intelligence artificielle contribue au renforcement des inégalités raciales et fait des recommandations pour prévenir et corriger les préjugés et la discrimination raciale.

Bien que ces instruments juridiques et politiques ne s'arrêtent pas aux frontières du monde numérique, leur applicabilité aux différentes formes de discrimination algorithmique souffre d'un certain nombre de lacunes.

4) Les limites et les incertitudes : où se situe la discrimination algorithmique ?

Ce patchwork juridique et politique aborde certains des risques discriminatoires de l'IA et de la prise de décision automatisée. Pourtant, **de nombreuses incertitudes subsistent quant à la mesure dans laquelle les dispositions légales existantes peuvent être utilisées pour promouvoir l'égalité et lutter contre la discrimination découlant de l'utilisation de ces technologies**. L'objectif de cette sous-section est donc d'examiner les lacunes existantes dans le cadre de l'égalité et de la non-discrimination décrit ci-dessus lorsqu'il s'agit de discrimination algorithmique. **Trois questions principales** se posent: 1) **l'absence de recoupement net entre les notions existantes de discrimination directe et indirecte et les formes de discrimination algorithmique**, 2) **les questions de procédure liées à la preuve** et

136. Commission européenne, « Déclaration européenne sur les droits et principes numériques pour la décennie numérique », COM(2022) 28 final (Bruxelles 2022).

à la **responsabilité**, 3) et les **défis liés à la protection par la législation de caractéristiques spécifiques**. Il a été expliqué à la section 3 que pour combler ces lacunes, il convient d'appliquer les obligations positives existantes en matière de promotion de l'égalité et d'intégrer des approches préventives de la discrimination algorithmique au titre de l'article 14 de la CEDH.

Les questions de qualification : discrimination algorithmique directe ou indirecte

Bien que l'article 14 de la CEDH ne fasse pas de distinction entre la discrimination directe et indirecte, la Cour européenne des droits de l'homme (la Cour) a établi cette distinction dans sa jurisprudence¹³⁷. **La discrimination directe** résulte « d'une **différence de traitement de personnes se trouvant dans des situations analogues ou similaires** », et lorsque cette différence est « **fondée sur une caractéristique identifiable** » ou un « **statut**¹³⁸ ». Par exemple, lorsque deux travailleurs qui ont des qualifications similaires sollicitent une promotion mais que l'un-e d'eux est préfér-e à l'autre « en raison de » son genre, il s'agit d'une discrimination directe fondée sur le genre.

Au début des années 2000, la **Cour a reconnu l'existence d'une discrimination indirecte** lorsque les États « n'appliquent pas un traitement différent à des personnes se trouvant dans des situations différentes¹³⁹ ». Elle a statué dans l'affaire *D.H. et autres c. République tchèque* qu'« une différence de traitement pouvait aussi consister en l'effet préjudiciable disproportionné d'une politique ou d'une mesure qui, bien que formulée de manière neutre, a un

137. La distinction a été établie en se référant au droit européen en matière d'égalité et à la jurisprudence de la Cour européenne de justice, voir *D.H. et autres c. République tchèque* Requête n° 57325/00 (Cour européenne des droits de l'homme, Grande Chambre, 13 novembre 2007), [184].

138. Voir par exemple *Kjeldsen, Busk Madsen et Pedersen c. Danemark*, requêtes n° 5095/71, 5920/72, 5926/72 (Cour européenne des droits de l'homme, 7 décembre 1976), [56] *Burden c. Royaume-Uni*, requête 13378/05 (Cour européenne des droits de l'homme, Grande Chambre, 29 avril 2008), [60] *Carson et autres c. Royaume-Uni*, requête n° 42184/05 (Cour européenne des droits de l'homme, Grande Chambre, 16 mars 2010), [61], et plus récemment *Biao c. Danemark*, requête n° 38590/10 (Cour européenne des droits de l'homme, Grande Chambre, 24 mai 2016), [89]. Voir également l'Agence des droits fondamentaux de l'Union européenne et le Conseil de l'Europe, *Manuel sur le droit européen en matière de non-discrimination* (Office des publications de l'Union européenne 2018), 43 et Cour européenne des droits de l'homme, *Guide sur l'article 14 de la Convention européenne des droits de l'homme et sur l'article 1 du Protocole n° 12 à la Convention* (Conseil de l'Europe 2020), 11.

139. *Thlimmenos c. Grèce*, requête n° 34369/97 (Cour européenne des droits de l'homme, 2 avril 2000), [44].

effet discriminatoire sur un groupe¹⁴⁰». Par exemple, une politique formulée de manière neutre qui conditionnerait le recrutement de candidat-es à une taille minimale pourrait avoir des effets indirectement discriminatoires sur les femmes, qui sont en moyenne plus petites que les hommes.

Dès lors qu'un constat *prima facie* de discrimination directe ou indirecte a été établi, un **système de justification ouvert** s'applique, selon lequel une **discrimination** ne peut être constatée que s'il n'y a «**aucune justification objective et raisonnable**¹⁴¹». En d'autres termes, tant la discrimination directe qu'indirecte peut être justifiée si elle poursuit un **but légitime** et s'il existe un «**rapport de proportionnalité entre les moyens employés et le but recherché**¹⁴²». Étant donné que le même système de justification s'applique en principe dans les deux cadres, qualifier la discrimination algorithmique de directe ou d'indirecte a des répercussions moins importantes sur les moyens de recours disponibles que dans le droit communautaire, où cette qualification conditionne l'applicabilité d'un système fermé ou ouvert de justifications¹⁴³. Néanmoins, il est important de comprendre comment les tribunaux, notamment la Cour européenne des droits de l'homme, qualifieront la discrimination algorithmique.

On a considéré jusqu'à présent que la discrimination algorithmique s'inscrivait principalement dans le cadre de la discrimination indirecte, notamment parce que les développeurs/développeuses sont peu susceptibles d'entrer des caractéristiques protégées dans les ensembles de données utilisés pour former les systèmes de prise de décision algorithmique (ADM)¹⁴⁴. Selon Hacker, par exemple, «dans les contextes d'apprentissage automatisé, la discrimination indirecte est le type de discrimination le plus pertinent» tandis que «la discrimination directe sera rare dans la prise de décision algorithmique, et se limitera en grande partie aux cas de partialité

140. *D.H. et autres c. République tchèque*, requête n° 57325/00 (Cour européenne des droits de l'homme, Grande Chambre, 13 novembre 2007), [184].

141. *Affaire «relative à certains aspects du régime linguistique de l'enseignement en Belgique» c. Belgique*, requêtes n° 1474/62; 1677/62; 1691/62; 1769/63; 1994/63; 2126/64 (Cour européenne des droits de l'homme, 23 juillet 1968), [10], p. 34.

142. *Ibid*, voir aussi *Marckx c. Belgique*, requête n° 6833/74 (Cour européenne des droits de l'homme, 13 juin 1979), [33].

143. Dans le droit de l'UE, la discrimination directe ne peut, en principe, être justifiée (sauf exceptions fermées), tandis que la discrimination indirecte donne lieu à un test de proportionnalité assorti d'un régime de justifications illimité.

144. Cet argument s'appuie sur une analogie avec le cadre américain de lutte contre la discrimination, voir par exemple Solon Barocas et Andrew D. Selbst, «Big Data's Disparate Impact» (2016) 104 *California law review* 671. Pourtant, la distinction entre la discrimination directe et indirecte dans le droit de la Cour diffère de la distinction américaine entre les notions de «traitement disparate» et d'«impact disparate».

implicite dans l'étiquetage¹⁴⁵ ». Borgesius et Kelly-Lyth font également valoir respectivement que « la loi sur la non-discrimination interdit de nombreux effets discriminatoires du processus décisionnel algorithmique, en particulier par le biais du concept de discrimination indirecte¹⁴⁶ » et que « la plupart des algorithmes biaisés relèveront du cadre de la discrimination indirecte¹⁴⁷ ».

Au moins trois arguments viennent à l'appui de ce point de vue : 1) la discrimination indirecte englobe les situations dans lesquelles des mesures formellement neutres entraînent des désavantages parce qu'elles interviennent dans un contexte social inégal et l'intègrent¹⁴⁸. Cela reflète la manière dont les technologies fondées sur les données intègrent et perpétuent le *statu quo* inégal de la société¹⁴⁹, 2) la discrimination indirecte met l'accent sur la dimension structurelle de la discrimination.¹⁵⁰ Cette orientation correspond au fait que les algorithmes d'apprentissage automatisé déduisent des règles à partir de modèles de groupes. Troisièmement, le concept de discrimination indirecte permet d'aborder les distinctions qui ne sont pas fondées sur des motifs légalement protégés mais qui, dans les faits, ont un impact sur des groupes protégés¹⁵¹. Étant donné que cette discrimination est l'une des formes dominantes de la discrimination algorithmique (voir plus loin), le cadre qui la définit présente un avantage supplémentaire.

Malgré le consensus sur la classification de la discrimination algorithmique comme étant indirecte, **une telle qualification « par défaut » soulève un**

145. Hacker, 'Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under EU law', 1152-1153.

146. Zuiderveen Borgesius, "Strengthening legal protection against discrimination by algorithms and artificial intelligence", 1578. Il reconnaît néanmoins l'existence d'une série de problèmes d'application.

147. Aislinn Kelly-Lyth, "Challenging Biased Hiring Algorithms" (2021) 41 Oxford Journal of Legal Studies 899, 906.

148. Voir Tobler, *Limits and potential of the concept of indirect discrimination*, 85. Sur le point de vue de l'auteur et de la victime, voir Alan David Freeman, « Legitimizing Racial Discrimination Through Anti-discrimination Law: A Critical Review of Supreme Court Doctrine' (1978) 62 Minnesota Law Review 1049.

149. Voir Anna Lauren Hoffmann, 'Where fairness fails: data, algorithms, and the limits of anti-discrimination discourse' (2019) 22 Information, Communication & Society 900.

150. Voir Hugh Collins et Tarunabh Khaitan, « Indirect Discrimination Law: Controversies and Critical Questions » in Hugh Collins et Tarunabh Khaitan (eds), *Foundations of Indirect Discrimination Law* (1 edn, Hart Publishing 2018), 19.

151. Par exemple, le travail à temps partiel est une question d'égalité de genre puisque la plupart des travailleurs à temps partiel sont des femmes. Voir Tobler, *Limits and potential of the concept of indirect discrimination*, 24 et Janneke Gerards, « Discrimination grounds », in: Dagmar Schiek, Lisa Waddington et Mark Bell (eds), *Cases, Materials and Text on National, Supranational and International Non-Discrimination Law*, Oxford and Portland, Oregon: Hart Publishing 2007, 33-184.

certain nombre de questions doctrinales et procédurales¹⁵². Des études récentes montrent que la notion de discrimination directe pourrait englober certains cas de discrimination algorithmique dans lesquels un groupe entier est systématiquement touché, quel que soit le critère utilisé pour la prise de décision¹⁵³. En outre, l'intégration des effets discriminatoires du biais algorithmique dans l'une ou l'autre notion soulève des questions normatives cruciales sur les concepts clés du droit de la non-discrimination¹⁵⁴. En ce sens, le CAHAL a reconnu dans son étude de faisabilité 2020 que « [l']importance accrue de la discrimination par procuration dans le contexte de l'apprentissage machine peut soulever des questions d'interprétation sur la distinction entre discrimination directe et indirecte ou, en fait, sur l'adéquation de cette distinction telle qu'elle est traditionnellement comprise¹⁵⁵ ». Par exemple, qu'est-ce qui peut être considéré comme un critère « neutre » pour la prise de décision à la lumière des boucles de rétroaction existantes et des problèmes de codage redondants ? La discrimination algorithmique, qui alimente les inégalités structurelles dans les décisions individuelles, est-elle une forme collective ou individuelle de traitement inéquitable ? L'utilisateur d'un algorithme doit-il être considéré comme un agresseur lorsqu'une machine apprend de manière autonome à discriminer ? Les réponses à ces questions détermineront, en théorie, si la notion de discrimination directe ou indirecte peut être utilisée pour appréhender la discrimination algorithmique¹⁵⁶.

152. Gerards J and Xenidis R, Algorithmic discrimination in Europe: Challenges and Opportunities for EU Gender Equality and Non-Discrimination Law (Réseau européen d'experts juridiques dans le domaine de l'égalité de genre et de la non-discrimination / European Commission, 2021).

153. Voir Adams-Prassl, Binns et Kelly-Lyth, «Directly discriminatory algorithms», *Modern Law Review* (à paraître).

154. Gerards J et Xenidis R, Algorithmic discrimination in Europe: Challenges and Opportunities for EU Gender Equality and Non-Discrimination Law (Réseau européen d'experts juridiques dans le domaine de l'égalité de genre et de la non-discrimination / European Commission, 2021).

155. CAHAL, "Feasibility Study on legal framework on AI design, development and application based on CoE standards" (2020), [13], p. 5.

156. Voir Gerards J et Xenidis R, Algorithmic discrimination in Europe: Challenges and Opportunities for EU Gender Equality and Non-Discrimination Law (Réseau européen d'experts juridiques dans le domaine de l'égalité de genre et de la non-discrimination / European Commission, 2021) et Xenidis R, «Tuning EU Equality Law to Algorithmic Discrimination: Three Pathways to Resilience» (2021) 27 *Maastricht Journal of European and Comparative Law* 736.

Les questions de procédure : preuve, proportionnalité, responsabilité et obligation

Dans la pratique, cependant, l'opacité des systèmes de décision algorithmiques laisse penser que les preuves nécessaires pour caractériser une discrimination directe feront souvent défaut. L'information pourrait n'être disponible qu'*ex post* et rester partielle, de sorte que l'on ne peut observer les effets d'un système algorithmique qu'après son utilisation. Par exemple, si un algorithme de notation de crédit refuse systématiquement un crédit aux personnes en situation de handicap, il est possible que l'on ne puisse pas accéder aux critères utilisés pour une telle décision mais que l'on soit en mesure d'observer un modèle de rejet par rapport aux personnes en situation de handicap. De même, il est possible que l'on ne puisse pas accéder aux informations concernant l'ensemble des demandeurs, de sorte qu'il n'y a pas de certitude concernant les demandeurs et demandeuses potentiel·les en situation de handicap qui ont obtenu un crédit ou les autres demandeurs et demandeuses qui se sont vu opposer un refus.

Le problème des preuves : pour les requérant·es potentiel·les, l'opacité des décisions algorithmiques constitue des obstacles substantiels à la réparation de la discrimination. Les asymétries d'information entre les utilisateurs et les sujets de la prise de décision algorithmique ou des systèmes d'aide à la décision signifient que les utilisateur·ices finaux/finales isolé·es n'auront pas la capacité de surveiller l'impact des décisions algorithmiques sur des groupes d'autres utilisateurs finaux. Ils/elles ne pourront pas non plus accéder aux informations sur les critères de décision. Même dans les cas potentiels de discrimination algorithmique indirecte, l'absence d'informations transparentes et significatives sur les critères de décision pertinents et le fait que les victimes n'aient pas une vue d'ensemble des décisions prises pourraient empêcher de prendre conscience de l'existence d'une discrimination. Cela peut éventuellement empêcher toute action en justice d'être engagée.

Les règles existantes en matière de charge de la preuve sont conçues pour aider les personnes qui saisissent la justice : quand il a été établi la preuve *prima facie* de discrimination, la charge de la preuve incombe en principe au défendeur, qui est chargé de démontrer que la différence de traitement est justifiée. **Pourtant, des problèmes juridiques continuent de se poser : comment fournir suffisamment d'éléments, et quel type d'informations présenter, pour établir une preuve *prima facie* de discrimination afin de déclencher le transfert de la charge de la preuve au défendeur ?** Dans le contexte

algorithmique, les asymétries d'information pourraient même empêcher de démontrer une discrimination *prima facie*¹⁵⁷.

Le test de proportionnalité : dès qu'une différence de traitement entre des personnes se trouvant dans une situation similaire ou l'absence de différence de traitement entre des personnes se trouvant dans une situation différente a été établie, les juges doivent effectuer un test de proportionnalité pour évaluer si elle peut être objectivement justifiée. Ce test en deux étapes vise à déterminer si la pratique répond à un objectif légitime et si les moyens employés sont raisonnablement proportionnés à l'objectif poursuivi¹⁵⁸. La réponse à ces questions conduit à une incertitude juridique considérable en raison de la nécessité pour les juges d'évaluer des compromis techniques qu'ils ne maîtrisent peut-être pas (par exemple, quelles mesures d'équité devaient être utilisées ? Comment équilibrer les compromis entre les différentes définitions de l'équité¹⁵⁹ ? Comment équilibrer l'exactitude et l'équité¹⁶⁰ ? Les barrières techniques qui en découlent pourraient contribuer à protéger les systèmes automatisés de prise de décision du contrôle judiciaire. Dans ces conditions, les études récentes vont dans le sens d'une application permissive du test de proportionnalité dans le contexte de l'opacité algorithmique¹⁶¹.

La responsabilité et l'obligation : la question de la responsabilité de la discrimination algorithmique est épineuse. Certain-es commentateurs/commentatrices font valoir que la loi devrait permettre « une extension des

157. Dans ce contexte, la recommandation CM/Rec(2020)1 du Comité des Ministres aux États membres sur les impacts des systèmes algorithmiques sur les droits de l'homme pourrait faciliter la production de preuve par les requérant-e-s car elle propose un principe de contestabilité : « Il faudrait instaurer comme condition préalable la nécessité d'expliquer et de préciser, de manière opportune, impartiale, lisible et accessible aux personnes dont les droits ou les intérêts légitimes peuvent être affectés, ainsi qu'aux autorités publiques compétentes, l'existence, le processus, la raison d'être, le raisonnement et les produits éventuels des systèmes algorithmiques aux niveaux individuel et collectif. »

158. Greffe de la Cour européenne des droits de l'homme, Guide sur l'article 14 de la Convention européenne des droits de l'homme et sur l'article 1 du Protocole n° 12 à la Convention (30 avril 2022) : https://www.echr.coe.int/Documents/Guide_Art_14_Art_1_Protocol_12_FRA.pdf (dernière consultation le 22 juillet 2022).

159. L'équité est un terme philosophique et statistique utilisé pour décrire si un système algorithmique traite différents groupes de manière équitable. Il existe différentes définitions de l'équité (par exemple, tous les groupes obtiennent des taux similaires de faux positifs et de faux négatifs, ou les performances d'un algorithme sont calibrées pour être similaires pour tous les groupes) qui peuvent être incompatibles entre elles. Le terme statistique « équité » et le terme juridique « égalité de traitement » ne se recoupent pas nettement.

160. Voir Binns R, "Algorithmic Decision-making: A Guide For Lawyers" (2020) 25 Judicial Review 2.

161. Pablo Martínez-Ramil, "Discriminatory algorithms. A proportionate means of achieving a legitimate aim?" (2022) Journal of Ethics and Legal Technologies 4(1).

motifs de défense des défendeurs [qui] pourrait leur permettre d'établir que les biais ont été développés de manière autonome par un algorithme¹⁶²». Cependant, un tel argument soulève la question difficile de savoir qui doit être tenu responsable de la discrimination algorithmique en l'absence de personnalité juridique des systèmes d'IA? En outre, la répartition de la responsabilité entre les fournisseurs et les utilisateurs d'IA (ceux qui les déploient) est une autre difficulté, car les deux pourraient porter la responsabilité d'un système discriminatoire. Compte tenu des nombreuses sources de biais algorithmique, qu'il s'agisse des données, des caractéristiques des modèles ou de leur mise en œuvre, il est presque impossible d'identifier une cause unique et précise de discrimination algorithmique.

Les questions relatives au champ d'application personnel de la loi sur la non-discrimination : l'inadéquation entre les systèmes algorithmiques et les motifs de discrimination protégés

La dernière série de défis qui se pose concerne l'absence de chevauchement entre le champ d'application personnel des dispositions juridiques en matière de non-discrimination et les formes idiosyncratiques de la subjectivité algorithmique.

La discrimination secondaire et la discrimination indirecte: Les études montrent que la discrimination algorithmique a lieu même lorsque les caractéristiques protégées sont supprimées d'un ensemble de données particulier. En effet, le profilage algorithmique repose sur des points de données qui, combinés, peuvent conduire à un regroupement qui chevauche les groupes protégés. Par exemple, le temps de trajet entre le domicile et le lieu de travail ou le code postal pourrait conduire à des déductions sur le statut socio-économique et l'ethnicité, étant donné la spatialisation existante des inégalités socio-économiques et raciales¹⁶³. En particulier, les problèmes de **codage redondant** se posent lorsque les variables d'un ensemble de données sont en corrélation avec une catégorie protégée, par exemple le temps de trajet domicile-travail et l'origine ethnique, qui peuvent être déduits par des algorithmes d'apprentissage automatique. S'ajoutent à cela les questions de **boucles de rétroaction**, qui décrivent des situations où un système s'appuie sur des données issues de discriminations passées pour établir des

162. Grozdanovski L, 'In search of effectiveness and fairness in proving algorithmic discrimination in EU law' (2021) 58 Common Market Law Review, 99.

163. Voir Williams BA, Brooks CF et Shmargad Y, « How Algorithms Discriminate Based on Data They Lack: Challenges, Solutions, and Policy Implications » (2018) 8 Journal of Information Policy 78.

prédictions. La discrimination algorithmique a donc de fortes chances de prendre la forme d'une **discrimination secondaire**.

L'article 14 de la CEDH interdit toute discrimination «fondée notamment sur le sexe, la race, la couleur, la langue, la religion, les opinions politiques ou toutes autres opinions, l'origine nationale ou sociale, l'appartenance à une minorité nationale, la fortune, la naissance ou toute autre situation». La discrimination indirecte qui repose, par exemple, sur des **données comportementales telles que le temps d'écran, l'utilisation du wifi, les données de géolocalisation**, etc. pourrait donc relever de l'article 14 CEDH **par la voie de la discrimination indirecte**, en démontrant un effet très désavantageux fondé sur l'un des motifs explicitement énumérés¹⁶⁴. **Le problème est que cette discrimination indirecte pourrait échapper à la protection juridique contre la discrimination en raison des difficultés de procédure** exposées dans la section ci-dessus¹⁶⁵.

Les «nouveaux» groupes algorithmiques et la notion d'«autre statut»: en outre, les algorithmes peuvent créer de nouvelles catégorisations fondées sur des caractéristiques apparemment inoffensives, telles que les préférences du navigateur web ou le numéro d'appartement, ou des catégories plus compliquées combinant de nombreux points de données. Par exemple, un magasin en ligne peut constater que la plupart des consommateur-ices utilisant un certain navigateur web font moins attention aux prix et décider de faire payer plus cher ces consommateur-ices. Bien qu'ils ne correspondent pas à des critères protégés par la législation sur la non-discrimination, certains de ces groupes algorithmiques pourraient mériter une protection juridique, par exemple si des modèles de différenciation algorithmique les exposent à un désavantage socio-économique systématique.

Lorsque la discrimination résultant de certains groupes algorithmiques ne recoupe pas des catégories explicitement protégées par l'article 14 de la CEDH, **la liste non limitative des motifs protégés de l'article 14 et l'approche**

164. La discrimination indirecte pourrait, dans certains cas, être traitée comme une discrimination directe, en fonction de la manière dont sont délimitées la portée et les limites des groupes protégés. Pour une discussion de ce problème dans le cadre de la notion de discrimination directe dans le contexte européen, voir Xenidis R, «Tuning EU Equality Law to Algorithmic Discrimination: Three Pathways to Resilience» (2021) 27 Maastricht Journal of European and Comparative Law 736.

165. Voir, par exemple, Anton Vedder et Laurens Naudts (2017) Accountability for the use of algorithms in a big data environment, *International Review of Law, Computers & Technology*, 31:2, 206-224 et Naudts, L. (2019). How Machine Learning Generates Unfair Inequalities and How Data Protection Instruments May Help in Mitigating Them. In R. Leenes, R. van Brakel, S. Gutwirth & P. De Hert (Editors), *Data Protection and Privacy: The Internet of Bodies* (Computers, Privacy and Data Protection).

flexible de la Cour européenne des droits de l'homme (la Cour) à l'égard de la protection des « nouveaux motifs » offrent sans doute une possibilité de protection¹⁶⁶. Il a été avancé que les clauses anti-discrimination « semi-ouvertes » telles que l'article 14 de la CEDH offrent de meilleures solutions pour remédier à la discrimination algorithmique que les dispositions anti-discriminatoires totalement fermées telles que celles du droit dérivé de l'UE¹⁶⁷. Par exemple, la Cour a protégé des groupes sur la base de leur statut professionnel ou de leur lieu de résidence¹⁶⁸. Cette approche ouverte, basée sur la notion d'« **autre statut** », pourrait faciliter l'extension de la portée de nouveaux groupes algorithmiques au titre de l'article 14 de la CEDH. Elle pose pourtant la question des **limites normatives de la législation contre la discrimination** : quels sont ses contours ? À quels types d'injustices est-elle censée s'attaquer ?

En outre, **certains groupes algorithmiques n'ont pas d'importance sociale et il est donc difficile de les considérer comme des groupes méritant d'être protégés par la loi sur la discrimination**¹⁶⁹. Les « nouveaux » groupes algorithmiques issus d'un regroupement algorithmique intangible font l'objet de distinctions qui ont des effets socio-économiques très tangibles et pourraient former à la longue une **discrimination structurelle « émergente**¹⁷⁰ ». Contrairement aux groupes algorithmiques fondamentaux sur le plan social, ces distinctions échapperont systématiquement à la législation sur l'égalité découlant de la CEDH. Des études récentes ont proposé d'étendre le champ d'application de la législation sur la lutte contre la discrimination pour couvrir ces distinctions algorithmiques préjudiciables¹⁷¹.

Il se pose un dernier problème relatif au champ d'application de la législation sur l'égalité découlant de la CEDH lorsque la **prise de décision algorithmique brouille les frontières entre l'individu et le groupe**. On s'aperçoit, en

166. Voir Gérards, Janneke et Frederik Zuiderveen Borgesius. « Protected Grounds and the System of Non-Discrimination Law in the Context of Algorithmic Decision-Making and Artificial Intelligence. » *Colorado Technology Law Journal*, à paraître (2020).

167. Ibid.

168. Voir Van der Mussel c. Belgique, requête n° 8919/80 (Cour européenne des droits de l'homme, 23 novembre 1983) et Carson et autres c. Royaume-Uni (2010), [70]-[71].

169. Voir Matthias Leese, The new profiling: Algorithms, black boxes, and the fail of anti-discriminatory safeguards in the European Union, 45 SECURITY DIALOGUE 494–511, 501 (2014) ; Monique Mann & Tobias Matzner, Challenging algorithmic profiling: The limits of data protection and anti-discrimination in responding to Emerging discrimination, 6 BIG DATA & SOCIETY, 5–6 (2019).

170. Ibid.

171. Voir Wachter S, « The Theory of Artificial Immutability: Protecting Algorithmic Groups Under Anti-Discrimination Law » (2022), Tulane Law Review (à paraître).

particulier, que les modèles fondés sur le groupe sont utilisés pour prendre des décisions concernant les individus. Cela présuppose que l'appartenance à des groupes algorithmiques donnés peut être attribuée à des individus, même si cela n'est pas exact dans les faits. Par exemple, un utilisateur ou une utilisatrice qui correspond au modèle de trafic web typique d'une femme de 25 à 30 ans résidant dans un environnement urbain peut se voir attribuer cette identité de genre et d'âge qui servira de base à une prise de décision ultérieure. Si le groupe algorithmique attribué ne correspond pas à l'identité réelle de la personne, celle-ci n'aura aucune possibilité de corriger les résultats du profilage algorithmique et du traitement qui en découle. Cependant, l'utilisateur-ice qui a été victime d'une discrimination fondée sur le sexe, par exemple une hausse des prix de l'assurance maladie, peut invoquer une « **discrimination par association** », notion reconnue par la Cour en 2008¹⁷².

La **discrimination intersectionnelle**: enfin, la discrimination algorithmique est susceptible d'être de nature intersectionnelle, c'est-à-dire d'impliquer plusieurs motifs de discrimination ou vecteurs de désavantage¹⁷³. En raison de la granularité du profilage algorithmique, les systèmes d'IA sont capables de déduire plusieurs appartenances sociales protégées et de **regrouper** potentiellement **les utilisateurs en fonction de différentes classifications problématiques**. Par exemple, les profils algorithmiques peuvent contenir des informations concernant le sexe, l'âge, l'origine ethnique, les croyances religieuses, l'orientation sexuelle ou l'identité de genre, sur la base de l'analyse des comportements en ligne, des préférences des consommateur-ices, etc. L'identification et la correction des cas de discrimination algorithmique intersectionnelle s'avèrent encore plus difficiles que les cas à facteur unique en raison du manque de données désagrégées sur l'égalité, qui ne permet pas de comparer les disparités potentielles entre les résultats algorithmiques et la situation réelle des groupes marginalisés de façon intersectionnelle¹⁷⁴.

172. Molla Sali c. Grèce requête n° 20452/14 (Cour européenne des droits de l'homme, 19 décembre 2018), [141].

173. L'exposé des motifs de la Recommandation de politique générale n°14 de l'ECRI [1] indique que la discrimination intersectionnelle est « une situation dans laquelle plusieurs motifs de discrimination interagissent au point de devenir inséparables, leur combinaison créant alors un motif nouveau ». Voir également Gerards J et Xenidis R, *Algorithmic discrimination in Europe: Challenges and Opportunities for EU Gender Equality and Non-Discrimination Law* (European network of legal experts in gender equality and non-discrimination / European Commission, 2021).

174. La catégorisation des données peut également être problématique et manquer de représentativité, ce qui peut avoir des conséquences sur les tentatives de correction de la discrimination algorithmique. Voir Ruberg, B. et Ruelos, S., « Data for queer lives: How LGBTQ gender and sexuality identities challenge standards of demographics » (2020), *Big Data & Society*, vol.

Les approches de débiaisage montrent également des limites lorsqu'il s'agit de remédier aux conséquences discriminatoires des biais visant les minorités intersectionnelles¹⁷⁵. Dans ce contexte, la discrimination intersectionnelle est souvent passée à travers les mailles du filet des recours judiciaires. Bien que la Cour européenne des droits de l'homme se soit attaquée avec succès (même implicitement) à la discrimination intersectionnelle dans une affaire comme *BS c. Espagne*,¹⁷⁶ elle ne l'a pas reconnue explicitement et n'y a pas remédié dans d'autres affaires comme *SAS c. France* ou *Garib c. Pays-Bas*¹⁷⁷. Cette **absence de cadre juridique solide contre la discrimination intersectionnelle**, souvent due à des conceptions formalistes de l'égalité fondées sur la comparaison, se révélera particulièrement problématique dans le contexte de la discrimination algorithmique.

II. Le droit relatif à la vie privée et à la protection des données : équité et exactitude

Outre les instruments juridiques relatifs à l'égalité et à la discrimination, la **législation relative à la vie privée et à la protection des données peut également être mise à profit pour lutter contre la discrimination algorithmique**. Le concept d'équité dans la loi sur la protection de la vie privée concerne l'intention d'une organisation d'utiliser de bonnes fois des informations à caractère personnel dans le but d'équilibrer les intérêts des responsables du traitement des données et des personnes concernées (les individus). On s'accorde généralement à dire, par exemple, que le traitement d'informations à caractère personnel qui est effectué sans que l'intéressé n'en soit avisé ou n'y consente conduirait à une situation injuste aux yeux des régulateurs de la vie privée. **Cependant, l'idée d'équité peut avoir de nombreuses nuances possibles : non-discrimination, juste équilibre, équité procédurale, bonne foi, etc.**

La relation entre discrimination et (in)équité se retrouve dans de nombreux textes législatifs, propositions et documents politiques à travers le monde. La Convention 108+, aux côtés du RGPD et de nombreuses autres lois sur la protection de la vie privée, énonce que le responsable du traitement doit, afin d'assurer un traitement équitable et transparent à l'égard de la personne

175. Balayn A et Gürses S, *Beyond Debiasing: Regulating AI and its inequalities* (European Digital Rights 2021), 62-63.

176. *B.S. c. Espagne* Requête n° 47159/08 (Cour européenne des droits de l'homme, 24 juillet 2012).

177. Voir, par exemple, *S.A.S. c. France* requête n° 43835/11 (Cour européenne des droits de l'homme, 1^{er} juillet 2014) ou *Garib c. Pays-Bas* requête n° 43494/09 (Cour européenne des droits de l'homme, Grande Chambre, 6 novembre 2017).

concernée, utiliser des procédures mathématiques ou des statistiques appropriées pour le profilage et mettre en œuvre des mesures techniques et organisationnelles appropriées pour prévenir les risques pouvant nuire aux intérêts et aux droits de la personne concernée. Les risques peuvent inclure la discrimination fondée sur l'origine raciale ou ethnique, les opinions politiques, l'appartenance syndicale, le statut génétique ou l'orientation sexuelle.

L'équité est un principe fondamental selon lequel des données à caractère personnel ne doivent pas être traitées d'une manière préjudiciable, discriminatoire, inattendue ou trompeuse pour la personne concernée. On peut affirmer que l'équité dans la législation sur la protection de la vie privée est liée à la **nécessité de remédier au déséquilibre des pouvoirs entre les personnes concernées (individus) et l'écosystème numérique**. C'est pour cette raison que la législation sur la protection de la vie privée a récemment été assez largement exploitée pour faire face aux méfaits de l'IA et de la prise de décision algorithmique, comme le souligne un rapport publié par le Future Privacy Forum¹⁷⁸. Le rapport met en lumière les mesures prises par les autorités chargées de la protection des données, notamment des obligations de transparence détaillées sur les paramètres qui ont conduit à une décision individuelle automatisée, une lecture large du principe d'équité pour éviter les situations de discrimination, et des conditions strictes pour un consentement valable en cas de profilage et de prise de décision automatisée.

Dans le cadre de cette étude, nous nous intéressons à deux éléments d'équité du point de vue de la vie privée :

- ▶ **L'équité en tant que procédures** : la transparence et l'équité sont inextricablement liées parce qu'il est possible de soutenir que l'ouverture du code source à un examen externe ou la fourniture d'une explication significative sur le traitement des informations à caractère personnel par le système d'IA pourrait conduire à l'identification des biais et de leurs causes profondes, et donc à une augmentation positive de la responsabilité publique. Par exemple, la *Corte di Cassazione* italienne a rendu une sentence en 2021 indiquant que le consentement d'une personne concernée ne peut être considéré comme valide si l'algorithme n'est pas transparent, car la personne concernée n'est pas

178. AUTOMATED DECISION-MAKING UNDER THE GDPR - A COMPREHENSIVE CASE-LAW ANALYSIS, Future Privacy Forum : <https://fpf.org/blog/fpf-report-automated-decision-making-under-the-gdpr-a-comprehensive-case-law-analysis/>.

en mesure de comprendre ce à quoi elle consent¹⁷⁹. Cette affaire a été accueillie favorablement par le régulateur italien de la protection de la vie privée, Garante, qui a estimé qu'elle montrait la manière dont la loi sur la protection de la vie privée (et le RGPD en l'espèce) est apte à défendre les droits des individus à l'ère de l'IA.

- **L'équité en tant que protection des vulnérabilités individuelles**: dans le droit de la vie privée, l'équité est souvent conçue comme un outil de correction visant à rééquilibrer les relations asymétriques ou déséquilibrées entre les organisations et les individus. Prenons par exemple le cas des plateformes algorithmiques au sujet desquelles le Conseil d'État français (tel que reformulé par la Commission Nationale de l'Informatique et des Libertés) affirme que « [l]a loyauté consiste à assurer de bonne foi le service de classement ou de référencement, sans chercher à l'altérer ou à le détourner à des fins étrangères à l'intérêt des utilisateurs¹⁸⁰ ». À un niveau plus général, dans l'environnement algorithmique, « l'équité pourrait bien représenter une solution au problème des *relations déséquilibrées* entre les contrôleurs d'algorithmes et les utilisateurs¹⁸¹ ».

Pour de nombreux pays, européens ou non européens, la modernisation de la Convention 108, qui a consisté à introduire de **nouveaux droits pour les personnes concernées dans des contextes de prise de décision algorithmique**, notamment en lien avec l'intelligence artificielle, représente un terrain d'entente, car le traité sert de norme limite sur la manière dont les pays doivent s'y prendre pour protéger le droit à la vie privée de leurs citoyen·nes à l'ère de l'IA. Le RGPD, qui présente de nombreuses similitudes avec la **Convention 108 +** (bien que le Conseil de l'Europe ait une portée et une territorialité beaucoup plus larges que celles de l'UE), contient également des dispositions visant à soutenir les droits individuels dans le contexte de l'IA et des algorithmes, y compris le célèbre article 22, qui protège les individus de la prise de décision automatisée.

Plusieurs autres protections s'appliquent à ces activités de traitement des données, notamment celles qui découlent des principes généraux de

179. *Corte di Cassazione, Civile Ord. Sez. 1 Num. 14381*, ItalgIureWeb, 25 May 2021 : <http://www.italgiure.giustizia.it/xway/application/nif/clean/hc.dll?verbo=attach&db=snciv&id=.:20210525/snciv@s10@a2021@n14381@tO.clean.pdf> (dernière consultation le 26 mai 2021).

180. Conseil d'État, « Le Numérique et les droits fondamentaux », 2014, p. 273 et 278-281 »

181. Understanding algorithmic decision-making: Opportunities and challenges, disponible à l'adresse : [https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624261/EPRS_STU\(2019\)624261_FR.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624261/EPRS_STU(2019)624261_FR.pdf).

traitement des données énoncés à l'article 5, des motifs juridiques du traitement énoncés à l'article 6, des règles relatives au traitement de catégories particulières de données (telles que les données biométriques) énoncées à l'article 9, des exigences spécifiques en matière de transparence et d'accès concernant la prise de décision algorithmique (ADM) énoncées aux articles 13 à 15, et de l'obligation de procéder à des évaluations d'impact sur la protection des données dans certains cas prévue à l'article 35.

Toutefois, les instruments actuels de protection de la vie privée présentent des **limites** lorsqu'il s'agit de l'IA et de la prise de décision algorithmique, notamment :

- L'exercice des droits des personnes concernées dans le contexte de l'IA et de la prise de décision algorithmique est assez complexe. Par exemple, même avec les conseils du groupe de travail 29 sur la protection des données concernant la prise de décision et le profilage individuels automatisés, l'affirmation de l'article 22 du RGPD (« **unique-ment** » automatisé, et « **effets juridiques ou similaires significatifs** ») présente des défis pratiques.

La transparence de la gestion algorithmique est le premier pas vers une véritable responsabilisation. Toutefois, les obligations de **transparence et d'explicabilité** relatives à l'atténuation des biais soulèvent des questions concernant l'**intersection des lois sur la protection de la vie privée et sur les secrets commerciaux**. Il est important que les systèmes algorithmiques atteignent un certain seuil d'accessibilité et d'intelligibilité, que ce soit pour des auditeurs internes ou externes, un organisme de réglementation ou un tribunal. Cependant, l'algorithme propre à une entreprise peut également être couvert par la législation sur les secrets commerciaux. Il existe des développements intéressants dans ce sens grâce à l'émergence du calcul multipartite sécurisé qui peut permettre d'interroger une IA sans avoir accès au code réel. Mais on en est encore loin. Les initiatives actuelles en matière de réglementation vont dans ce sens; la recommandation CM/Rec(2020)1 du Conseil de l'Europe sur les impacts des systèmes algorithmiques sur les droits de l'homme prévoit ainsi que « [l]es cadres législatifs relatifs à la protection de la propriété intellectuelle ou du secret d'affaires ne devraient pas empêcher une telle transparence, pas plus que les États ou les parties privées ne devraient chercher à les exploiter à cette fin », et que « [l]es considérations de confidentialité ou les secrets commerciaux ne devraient pas entraver la mise en œuvre d'études d'impact effectives sur les droits humains ».

III. Les réglementations sectorielles de l'IA : forces et limites pour la promotion de l'égalité et la lutte contre la discrimination

Outre les lois sur la discrimination, la vie privée et la protection des données, les réglementations sectorielles seront également pertinentes pour lutter contre la discrimination algorithmique.

Le Conseil de l'Europe élabore actuellement une réglementation qui traiterait de la discrimination algorithmique dans le cadre d'une initiative visant à promouvoir les droits humains, la démocratie et l'État de droit. Il pourrait s'agir d'un instrument transversal juridiquement contraignant traitant des questions relatives au secteur public, ainsi que de réglementations sectorielles contraignantes et non contraignantes¹⁸². En 2020, le CAHAI a préparé une « **Étude de faisabilité sur un cadre juridique relatif à la conception, au développement et à l'application de l'IA, fondé sur les normes du Conseil de l'Europe** », qui reconnaît que « les systèmes d'IA [peuvent] être utilisés d'une manière qui perpétue ou amplifie les biais injustes, qui sont aussi fondés sur de nouveaux motifs de discrimination en cas de discrimination dite « par procuration¹⁸³ ». En même temps, le CAHAI considère que « les systèmes d'IA peuvent promouvoir et renforcer les droits humains de manière plus générale, et contribuer à faire en sorte qu'ils soient respectés et effectivement appliqués », par exemple « en détectant des décisions (humaines ou automatisées) biaisées, en surveillant les modes de représentation de différents groupes (comme les femmes dans les médias) ou en analysant les structures discriminatoires au sein des organisations¹⁸⁴ ».

Dans son document de 2021 intitulé « **Éléments potentiels d'un cadre juridique sur l'intelligence artificielle, fondés sur les normes du Conseil de l'Europe en matière de droits de l'homme, de démocratie et d'État de droit** », le CAHAI recommande d'inclure « une disposition sur le respect de l'égalité de traitement et de la non-discrimination des individus en lien avec le développement, la conception et l'application des systèmes d'IA, afin d'éviter que des biais injustifiés ne soient intégrés dans ces systèmes et l'utilisation de systèmes d'IA entraînant des effets discriminatoires » dans

182. Voir CAHAI, « Étude de faisabilité sur un cadre juridique relatif à la conception, au développement et à l'application de l'IA, fondé sur les normes du Conseil de l'Europe » (2020), [54].

183. Comité sur l'intelligence artificielle, « Éléments potentiels d'un cadre juridique sur l'intelligence artificielle, fondés sur les normes du Conseil de l'Europe en matière de droits de l'homme, de démocratie et d'État de droit », *Conseil de l'Europe* (2022), [13]

184. *Ibid.*, [20].

la convention-cadre transversale juridiquement contraignante sur la réglementation de l'IA, actuellement en cours d'élaboration¹⁸⁵.

Le CAHAI propose également une réglementation complémentaire pour le secteur public et recommande que « le processus de documentation et de journalisation » relatif au développement du système soit « méticuleusement conservé afin d'assurer la transparence et la traçabilité du système ». Il recommande également que « des processus de test et de validation adéquats, ainsi que des mécanismes de gouvernance des données, soient mis en place » pour évaluer « le risque potentiel d'accès ou de traitement inégal, les différentes formes de préjugés et de discrimination, ainsi que l'impact sur l'égalité de genre¹⁸⁶ ».

Étant donné que d'autres réglementations sectorielles sont envisagées en Europe, il est important de préciser la **valeur ajoutée d'une réglementation de l'IA au niveau du Conseil de l'Europe**. On peut soutenir que la réglementation du Conseil de l'Europe peut avoir **une forte influence au niveau mondial** en raison du grand nombre de ses membres, de son approche particulière fondée sur les droits humains et du fait que l'instrument serait également ouvert à la ratification de parties non étatiques. Le document « Éléments potentiels » du CAHAI fait référence à des normes minimales et une approche axée sur le secteur public, conformément au mécanisme de la Convention européenne des droits de l'homme, ce qui diffère de « l'approche axé sur le marché » adoptée par l'UE dans son projet de législation européenne relative à l'IA¹⁸⁷. L'approche fondée sur les risques qu'ils adoptent tous deux pour les systèmes d'IA est une caractéristique commune aux deux règlements¹⁸⁸. Pourtant, le Conseil de l'Europe a la capacité de favoriser une **approche distincte de l'IA et des technologies algorithmiques, fondée sur les droits humains**.

Une réglementation sectorielle de l'IA est également en cours dans l'UE. Le projet de **législation de l'UE** relative à l'IA suit une approche fondée sur le **risque et classe les systèmes d'IA comme « à haut risque »** s'ils sont déployés dans les domaines suivants : identification et catégorisation biométriques des personnes physiques, gestion et exploitation des infrastructures critiques (circulation routière, eau, gaz, chauffage et fourniture d'électricité),

185. Ibid, [27]

186. Ibid, [60]

187. Voir Marten Breuer, "The Council of Europe as an AI Standard Setter" *Verfassungsblog* (4 avril 2022) : <https://verfassungsblog.de/the-council-of-europe-as-an-ai-standard-setter/>.

188. Voir Comité sur l'intelligence artificielle, « Éléments potentiels d'un cadre juridique sur l'intelligence artificielle, fondé sur les normes du Conseil de l'Europe en matière de droits de l'homme, de démocratie et d'État de droit », *Conseil de l'Europe* (2022), [19].

éducation et formation professionnelle, emploi, gestion des travailleurs et accès à l'emploi indépendant, accès et jouissance des services privés essentiels et des services et avantages publics, application des lois, migration, asile et gestion du contrôle aux frontières, administration de la justice et processus démocratiques. Les systèmes d'IA qui présentent un « **risque inacceptable** » sont interdits, par exemple « les pratiques qui présentent un risque important de manipuler des personnes par des techniques subliminales agissant sur leur inconscient, ou d'exploiter les vulnérabilités de groupes vulnérables spécifiques tels que les enfants ou les personnes en situation de handicap afin d'altérer sensiblement leur comportement d'une manière susceptible de causer un préjudice psychologique ou physique à la personne concernée ou à une autre personne ». Les systèmes d'IA qui présentent un **risque limité** sont soumis à des obligations de transparence spécifiques et ceux qui présentent un **risque faible ou minime** à des codes de conduite.

Bien que la réglementation de l'UE relative à l'IA prévoit des obligations de transparence prometteuses en vue d'atténuer les biais, en particulier en ce qui concerne les données de formation et les critères de décision¹⁸⁹, plusieurs **critiques** ont été formulées concernant la manière dont ladite réglementation propose de garantir le respect des droits fondamentaux. Par exemple, elle aborde les systèmes d'IA du point de vue de la responsabilité du fait des produits et **ne prévoit donc pas de mécanismes de plainte** qui permettraient **aux victimes de discrimination algorithmique ou aux ONG ayant un intérêt légitime de demander que des modifications soient apportées à ces systèmes après leur déploiement** conformément à la loi contre la discrimination¹⁹⁰. En outre, les commentateurs et commentatrices ont critiqué le fait que les fournisseurs et les utilisateurs de systèmes d'IA n'ont pas l'obligation légale **de procéder à des évaluations ex ante de l'impact sur les droits humains**¹⁹¹. L'absence de toute clause d'intégration de l'égalité ou d'obligation positive exigeant que l'IA et les systèmes algorithmiques favorisent l'égalité est également regrettable. **Il s'agit là d'aspects sur lesquels l'instrument du Conseil de l'Europe devrait se concentrer afin de créer une complémentarité avec les réglementations sectorielles de**

189. Voir en particulier l'article 10 sur la gouvernance des données et des données de la réglementation de l'UE sur l'IA.

190. Voir Joan Lopez Solano, Aaron Martin, Siddharth de Souza et Linnet Taylor, « Governing data and artificial intelligence for all Models for sustainable and just data governance » (Panel pour l'avenir de la science et de la technologie, Service de recherche du Parlement européen 2022), 52 disponible sur [https://www.europarl.europa.eu/RegData/etudes/STUD/2022/729533/EPRS_STU\(2022\)729533_FR.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2022/729533/EPRS_STU(2022)729533_FR.pdf).

191. Voir *ibid.*

l'IA de l'UE et de garantir que son mandat en matière de droits humains soit au cœur des nouvelles dispositions légales.

En 2022, la Commission européenne a **proposé une nouvelle directive relative à l'adaptation des règles en matière de responsabilité civile extracontractuelle au domaine de l'intelligence artificielle**¹⁹². La proposition a pour objectif de permettre « une application effective des droits fondamentaux dans la sphère privée et [de préserver] le droit à un recours effectif en cas de réalisation de risques spécifiques à l'IA », y compris la non-discrimination. La Commission explique que la proposition vient « compléter d'autres instruments de la politique de la Commission en matière d'IA qui repose sur des obligations préventives en matière de réglementation et de surveillance visant directement à éviter la violation des droits fondamentaux (comme la discrimination) ». Comme elle « ne crée pas de devoirs de vigilance ni de responsabilité pour les différentes entités dont l'activité est régie par [la législation sur la non-discrimination] et ne prévoit pas d'harmonisation en la matière [...], elle ne prévoit pas de nouvelles actions en responsabilité et n'a pas d'incidence sur les exemptions de responsabilité prévues [par la législation sur la non-discrimination] ; en revanche, elle introduit « un **allègement de la charge de la preuve pour les victimes de dommages causés par des systèmes d'IA** pour les actions qui peuvent être fondées sur le droit national ou ces autres actes législatifs de l'UE » en matière de non-discrimination. La présente étude estime que **le Conseil de l'Europe pourrait s'inspirer de ces règles pour faciliter l'accès des requérant-es à la justice** en ce qui concerne les plaintes pour discrimination algorithmique, en particulier s'agissant des questions de preuve.

Il est expliqué dans la section 3 ci-dessous que les futures réglementations sectorielles de l'IA au niveau du Conseil de l'Europe devraient également inclure une **obligation légale pour l'IA et les systèmes algorithmiques de promouvoir l'égalité**. La législation norvégienne sur l'égalité pourrait constituer une référence utile dans ce contexte, car elle indique que la promotion de l'égalité doit être considérée comme une obligation légale¹⁹³.

192. Commission européenne, Proposition de Directive du Parlement européen et du Conseil relative à l'adaptation des règles en matière de responsabilité civile extracontractuelle au domaine de l'intelligence artificielle (Directive sur la responsabilité en matière d'IA) COM(2022) 496 final.

193. Voir le chapitre 4 de la loi norvégienne relative à l'égalité et à l'interdiction de la discrimination (loi sur l'égalité et l'interdiction de la discrimination): https://lovdata.no/dokument/NLE/lov/2017-06-16-51#KAPITTEL_4.

Section 3

La promotion de l'égalité dans et par l'utilisation de l'IA : le rôle des actions et obligations positives

Si la section précédente a mis en évidence les instruments juridiques et politiques pertinents du Conseil de l'Europe, de l'UE et au niveau international, elle a également signalé les lacunes, les insuffisances et les incertitudes liées à l'applicabilité de ces instruments au problème de la discrimination algorithmique. Cette section montre que la résolution de ces problèmes passe par un **changement de paradigme**. Nous suggérons d'abord de **revoir les règles existantes à la lumière des nouvelles asymétries de pouvoir et d'information propres aux technologies algorithmiques**. Nous recommandons ensuite que les **actions et obligations positives soient utilisées comme un moyen d'élaborer une obligation juridique de prévenir la discrimination et de promouvoir l'égalité dans et par l'utilisation des systèmes algorithmiques**. L'adoption de ces deux mesures ferait du principe d'«**égalité dès la conception**» une **caractéristique importante de l'approche du Conseil de l'Europe fondée sur les droits humains** visant à lutter contre la discrimination algorithmique.

I. La révision des règles existantes à la lumière des nouvelles asymétries de pouvoir

La présente section a pour objet de définir les moyens de répondre aux questions soulevées à la section 2 en ce qui concerne l'applicabilité des dispositions juridiques existantes.

Premièrement, à la lumière des études actuelles qui montrent qu'en l'absence de protections, le biais algorithmique imprègne systématiquement les décisions algorithmiques, une **présomption de biais algorithmique** pourrait être posée lorsqu'aucune mesure préventive n'a été prise par les utilisateurs des systèmes algorithmiques. La présomption se justifie par l'omniprésence des biais dans le processus de conception des systèmes d'IA, notamment

les biais dans la collecte des données et les ensembles de données, les biais dans la conception des problèmes, les modèles algorithmiques et la mise en œuvre des recommandations en matière d'IA¹⁹⁴. Selon Eubanks, « lorsque les outils de décision automatisés ne visent pas explicitement à éliminer les inégalités structurelles, leur vitesse accrue et leur ampleur les aggravent considérablement¹⁹⁵ ». En d'autres termes, la probabilité que la discrimination algorithmique se produise est très grande lorsqu'aucune mesure de protection n'a été mise en place. Lorsqu'elle perpétue l'inégalité, l'utilisation de systèmes d'IA biaisés devrait être comparée à une mise en œuvre active d'un désavantage structurel et à l'amplification de la distribution injuste de biens sociaux essentiels. La **prévisibilité des préjudices discriminatoires découlant d'un biais algorithmique** justifie donc que la discrimination algorithmique soit considérée conceptuellement comme une forme de **négligence**. À cet égard, les travaux de Moreau sur la discrimination et les théories de la discrimination fondées sur la responsabilité civile¹⁹⁶ nous permettent de déduire qu'il existe une **responsabilité sociale pour les utilisateurs de systèmes algorithmiques de prendre des mesures raisonnables pour prévenir l'aggravation de la discrimination** dans la société. Cette approche fait écho aux discussions qui ont lieu actuellement dans le contexte de l'UE et, en particulier, à la proposition de la Commission relative à une « présomption réfragable pour les dommages liés à l'IA¹⁹⁷ ».

Deuxièmement, l'utilisation généralisée des systèmes d'IA crée **de nouvelles asymétries de pouvoir et d'information**. Il devient très **difficile pour ceux et celles qui sont l'objet de décisions algorithmiques d'identifier la discrimination** car elle résulte de la combinaison de la personnalisation, de l'automatisation et de l'opacité des processus décisionnels. La comparaison avec des personnes de même rang et les interactions sociales sont des dispositifs heuristiques importants lorsqu'il s'agit de poser des présomptions de discrimination. Or il devient impossible de prendre connaissance d'indices sociaux ou de se comparer à d'autres personnes demandant des prêts dans

194. Grozdanovski suggère qu'il est possible de remarquer l'existence d'une telle présomption dans le livre blanc de l'UE sur l'intelligence artificielle. Voir Grozdanovski L, "In search of effectiveness and fairness in proving algorithmic discrimination in EU law" (2021) 58 Common Market Law Review.

195. Eubanks V, Automating inequality: how high-tech tools profile, police, and punish the poor (First edition. edn, St. Martin's Press 2018).

196. Voir Sophia Moreau, "Discrimination as negligence" (2010) 40 Canadian Journal of Philosophy 123; Oppenheimer DB, "Negligent Discrimination" (1993) 141 University of Pennsylvania law review 899.

197. Voir en ce sens Luca Bertuzzi, « LEAK : La Commission va proposer une présomption réfragable pour les dommages liés à l'IA » (Euractiv, 2022) : <https://www.euractiv.com/section/digital/news/leak-commission-to-propose-rebuttable-presumption-for-ai-related-damages/>.

le contexte d'un service de crédit en ligne¹⁹⁸. Cette asymétrie d'information fait qu'il est difficile de soupçonner une discrimination en premier lieu. Même en cas de suspicion, **la collecte de preuves représente un problème supplémentaire** car les décisions ou les recommandations algorithmiques qui les sous-tendent ne sont pas facilement consultables et souvent non divulguées par les utilisateurs des systèmes de prise de décision algorithmique. Par conséquent, **la présentation de preuves établissant une présomption de discrimination devant les tribunaux est un défi juridique majeur**. Même si le déplacement de la charge de la preuve peut contribuer à atténuer les asymétries de pouvoir créées par des systèmes algorithmiques opaques¹⁹⁹, le seuil de déclenchement de ce déplacement devrait tenir compte de la position des utilisatrices finales/finales et de leur accès limité aux preuves *prima facie*.

Le fait de rapprocher la prévisibilité des biais algorithmiques et les asymétries d'information existantes montre que le déploiement généralisé des systèmes d'IA dans **les processus décisionnels perturbe l'équilibre entre la situation des victimes potentielles de discrimination et celle des fournisseurs et des utilisateurs de ces systèmes**. D'un côté, les victimes font l'objet d'une discrimination généralisée qu'elles ne sont pas actuellement en mesure d'identifier et de prouver, de l'autre, les entreprises à but lucratif jouissent d'un pouvoir accru grâce à des systèmes d'IA qui augmentent leurs profits tout en les dégageant éventuellement de toute responsabilité à l'égard de leurs conséquences discriminatoires en raison des obstacles juridiques énumérés ci-dessus. Il importe donc que **le cadre juridique soit ajusté pour mieux refléter et intégrer les changements de pouvoir et les déséquilibres** qui découlent de l'utilisation des systèmes d'IA dans un large éventail de décisions qui offrent ou non de meilleures chances à chacun et peuvent donc aggraver les inégalités dans la société.

La révision des règles existantes sur la charge de la preuve peut contribuer à restaurer l'efficacité de la législation sur la lutte contre la discrimination à la lumière des nouvelles asymétries de pouvoir et d'information entre les utilisateurs et les sujets des systèmes décisionnels algorithmiques. Poser une présomption de biais algorithmique (voir ci-avant) permettrait de **transférer la charge de la preuve au défendeur dès lors qu'aucune mesure**

198. Lorsque des systèmes d'IA sont utilisés dans le secteur privé, en particulier, les personnes ne sont pas nécessairement informées qu'un système algorithmique intervient dans la prise d'une décision les concernant. Il est donc encore plus difficile de savoir quand solliciter un réexamen par un être humain ou de prêter attention aux spécificités d'une éventuelle discrimination algorithmique.

199. Voir C-109/88 Handels- og Kontorfunktionærernes Forbund I Danmark v Dansk Arbejdsgiverforening, agissant au nom de Danfoss EU:C:1989:383.

préventive n'a été prise. Ces mesures préventives pourraient prendre la forme, par exemple, d'une analyse d'impact, d'un audit ou d'une certification du système algorithmique utilisé (voir la section des recommandations). Le fait de ne pas prendre de mesures préventives adéquates pourrait alors constituer une négligence. Ce mécanisme aiderait les victimes potentielles à présenter des preuves *prima facie* accessibles en vue de renverser la charge de la preuve sur les utilisateurs. Une telle adaptation du cadre juridique **intégrerait également l'action positive** et les **obligations préventives** contre les biais algorithmiques, (voir ci-dessous).

Troisièmement, l'adaptation des règles existantes suggérées ci-dessus devrait être combinée avec une approche de contrôle public²⁰⁰. À cet égard, il serait bon de **donner aux organismes de promotion de l'égalité, aux médiateurs pour les questions de discrimination et aux institutions nationales de défense des droits humains les moyens de contrôler l'impact discriminatoire des systèmes algorithmiques de prise de décision et des systèmes qui viennent en soutien.** Il s'agit donc de fournir à ces institutions les droits légaux et les pouvoirs d'investigation nécessaires (par exemple, pour accéder aux ensembles de données et aux critères de décision), les ressources adéquates, mais aussi la capacité de prévenir la discrimination en coopérant avec les utilisateurs des systèmes de prise de décision automatisée (par exemple des entreprises qui utilisent ces systèmes à l'appui des procédures de recrutement) afin de collecter des données pertinentes sur l'impact de leurs décisions, et d'aider les victimes potentielles à obtenir réparation. Le contrôle pourrait prendre la forme d'un **test de situation** qui permettrait à ces autorités de tester les résultats d'un système donné en comparant les résultats obtenus pour différents groupes. Elles pourraient par exemple tester des CV ou des demandes de crédit provenant de groupes majoritaires et minoritaires pour détecter une éventuelle discrimination algorithmique dans les contextes où les entreprises utilisent des systèmes de prise de décision automatisée. Les autorités pourraient également effectuer des **audits** pour détecter les biais potentiels si elles ont accès aux systèmes pertinents. Ces méthodes de **contrôle public** pourraient aider les victimes en atténuant les obstacles existants à l'établissement d'une discrimination *prima facie*.

La fonction de contrôle des organismes de promotion de l'égalité doit être renforcée par **des obligations légales en matière de transparence.** Les utilisateurs de systèmes algorithmiques devraient être tenus de **fournir des informations significatives et intelligibles sur les critères utilisés pour la prise**

200. Voir Xenidis R et Senden L, « EU Non-discrimination Law in the Era of Artificial Intelligence: Mapping the Challenges of Algorithmic Discrimination » in Bernitz U and others (eds), *General Principles of EU Law and the EU Digital Order* (Wolters Kluwer 2019).

de décision. Pour l'instant, le RGPD n'offre pas de droit à une explication²⁰¹. Dans le domaine des biens et services, la protection des consommateur-ices devrait également être étudiée en tant qu'outil permettant de demander des informations sur les décisions algorithmiques concernant des personnes qui ont été potentiellement victimes de discrimination. Ce domaine pourrait contribuer à corriger les asymétries de pouvoir créées par l'opacité des systèmes de prise de décision automatisée entre les sujets des décisions algorithmiques et leurs auteurs. La récente recommandation CM/Rec(2020)1 du Conseil de l'Europe sur les impacts des systèmes algorithmiques sur les droits de l'homme offre à cet égard un certain nombre de pistes intéressantes.

Quatrièmement, il est nécessaire de garantir la possibilité de révision des systèmes algorithmiques à la lumière des obligations de non-discrimination. Lorsque les parties demanderesses, les avocat-es ou les juges se voient présenter des informations techniques concernant un système spécifique, il est peu probable que ces informations soient suffisamment intelligibles pour qu'ils et elles puissent déterminer la nature discriminatoire ou non discriminatoire du système. **Les discussions techniques sur l'adéquation des paramètres d'équité donnés et les seuils appropriés des compromis entre l'exactitude et l'équité sont difficiles à évaluer du point de vue des obligations juridiques découlant de la législation contre la discrimination.** Dans ce contexte, comment faire en sorte que les systèmes de prise de décision automatisée soient soumis à un **test de proportionnalité** qui garantisse l'efficacité du droit de la non-discrimination ? Là encore, plusieurs solutions peuvent être envisagées (voir la section des recommandations de cette étude). D'une part, les **obligations de transparence** qui incombent aux utilisateurs des systèmes de prise de décision automatisée pourraient garantir l'accès à un compte rendu intelligible des choix techniques et équitables effectués par les développeurs/développeuses et les utilisateurs. D'autre part, **l'intégration de l'action positive** pourrait conduire à une **obligation positive de prévenir les biais algorithmiques qui déplacerait l'évaluation de la proportionnalité du terrain technique au terrain juridique.** Dans ce cas, les juges examineraient plutôt la pertinence des mesures préventives prises pour éviter les biais que les choix techniques en faveur de l'équité.

Enfin, nous suggérons que **l'obligation juridique, en tant que notion judiciaire se rapprochant de la responsabilité, devrait être répartie stratégiquement afin de faciliter l'accès à la justice et aux recours** dans les cas de discrimination algorithmique. Dans le contexte de la CEDH et d'autres

201. Voir Wachter, Sandra, Brent Mittelstadt, and Luciano Floridi. "Why a right to explanation of automated decision-making does not exist in the general data protection regulation." *International Data Privacy Law* 7.2 (2017): 76-99.

instruments juridiques du Conseil de l'Europe, qui imposent des obligations aux autorités publiques, nous suggérons que **les États parties tiennent les utilisateurs des systèmes d'IA responsables de la discrimination algorithmique découlant du déploiement de leur système**. La section consacrée aux recommandations explique que **cette responsabilité peut être complétée par l'obligation juridique faite aux prestataires de procéder à des évaluations *ex ante* de l'impact sur les droits humains afin de prévenir les préjudices discriminatoires**. Un tel dispositif permettra également d'encourager la **consignation de toutes les mesures préventives** prises par le fournisseur afin **que des informations essentielles puissent être fournies à l'utilisateur et aux utilisateur·ices finaux/finales** du système en cas de procédure judiciaire.

L'approche proposée ici, qui s'articule autour d'une présomption de biais algorithmique, de négligence et de prévention, pourrait contribuer à la sécurité juridique et à l'efficacité des dispositions antidiscriminatoires de la CEDH en allégeant la charge de la preuve des victimes, en favorisant les garanties préventives, en clarifiant la répartition des responsabilités et en aidant à mieux définir les justifications auxquelles les défenseurs peuvent avoir accès. Dans l'ensemble, nous suggérons qu'**une approche plus réelle de l'égalité devrait guider l'interprétation des dispositions antidiscriminatoires** afin de préserver leur efficacité dans le contexte de la discrimination algorithmique.

II. Une obligation de promouvoir l'égalité dans et par l'utilisation de systèmes algorithmiques : le rôle des actions et obligations positives

Ce rapport a montré comment les systèmes d'IA, sans les garde-fous et les contrôles appropriés, peuvent conduire à une plus grande exclusion des femmes et des groupes vulnérables. Malgré le potentiel discriminatoire de l'IA, les chercheurs/chercheuses et les développeurs/développeuses ont étudié les possibilités offertes par l'IA pour déterminer et corriger les inégalités. L'analyse nécessite un **changement de paradigme dans lequel les bases de la conception et du déploiement des logiciels sont systématiquement remises en question et vérifiées par rapport à leur impact en matière d'inclusion ou d'exclusion**. En d'autres termes, le déploiement d'un nouveau système d'IA devrait être « délibérément et intentionnellement inclusif » et « doit autonomiser les communautés et présenter un avantage

pour toute la société²⁰²». Pour ce faire, il faut exiger des entreprises qu'elles respectent une série d'obligations et mettre en place un ensemble de contrôles avant et après la mise sur le marché. Selon nous (voir ci-après), un tel changement de paradigme exige que le vaste éventail de mesures d'action positive disponibles, notamment la sensibilisation, les mesures basées sur la promotion, les mesures temporaires spéciales et les quotas, soit utilisé à des fins d'égalité, de diversité et d'inclusion dans tous les domaines.

L'éradication des biais et des inégalités impose de faire un choix conscient, sans doute politique et social. Dans un premier temps, il convient d'admettre que les systèmes d'IA ne sont pas neutres mais reproduisent et amplifient les inégalités structurelles et les systèmes d'exclusion et de désavantage qui sont institutionnalisés dans la société. Il faut donc s'éloigner du point de vue de l'auteur de la discrimination et reconnaître au contraire que les normes prédominantes et les hypothèses incontestées qui sous-tendent le développement et le déploiement des logiciels empêchent de prendre en compte les besoins des femmes et des groupes minoritaires²⁰³. Le fait de supposer qu'un système répondra de manière égale aux besoins de différents groupes empêchera *de facto* les groupes minoritaires de bénéficier des applications de l'IA et des opportunités qui y sont liées dans la même mesure que les autres groupes. Par conséquent, **l'égalité réelle et la lutte contre la discrimination « dès la conception » devraient être placées au centre de la réglementation juridique du développement et du déploiement de l'IA.**

1) Qu'est-ce qu'une action positive ?

L'action positive, également appelée mesures temporaires spéciales ou mesures positives dans le contexte européen, est une série de politiques qui peuvent être adoptées en vue de parvenir à l'égalité totale ou *de facto*. Elle s'appuie sur une critique de l'égalité formelle ou de l'égalité des chances qui dénonce l'aveuglement de ces cadres vis-à-vis des différentes situations de départ des différents groupes sociaux. Par exemple, donner la même possibilité d'emploi à un.e travailleur/travailleuse en situation de handicap et à

202. Renee Cummings, "This is how AI can support diversity, equity and inclusion", World Economic Forum: <https://www.weforum.org/agenda/2022/03/ai-support-diversity-equity-inclusion/>. Voir aussi Equality Now, A Call For An Intersectional Feminist Informed Universal Declaration On Digital Rights : https://www.equalitynow.org/news_and_insights/universal-declaration-on-digital-rights/.

203. Pour un compte rendu éloquent du point de vue de l'agresseur sur la discrimination par rapport à la compréhension de la discrimination comme un phénomène structurel, voir par exemple Freeman AD, « Legitimizing Racial Discrimination Through Anti-discrimination Law: A Critical Review of Supreme Court Doctrine » (1978) 62 Minnesota Law Review. La distinction a été reconnue par la loi à travers la notion de discrimination indirecte.

un-e travailleur/travailleuse valide pourrait conduire à un taux d'abandon plus élevé dans le premier cas parce qu'aucune mesure d'adaptation n'a été prise pour que la personne en situation de handicap soit effectivement capable d'accomplir ses tâches. En revanche, l'ancrage de politiques dans les théories de l'égalité réelle exige l'adoption de mesures d'accommodement spéciales qui créent des conditions permettant aux femmes et aux groupes historiquement défavorisés de participer à la société et de récolter les bénéfices de cette participation au même titre que les groupes privilégiés. Il s'agit concrètement de veiller à ce que les personnes en situation de handicap puissent accéder à un environnement de travail physique et psychologique sûr et adapté, par exemple grâce à des équipements spéciaux, des horaires de travail flexibles, etc. Les théories dites de l'égalité transformative vont dans le même sens, mais mettent davantage l'accent, d'un point de vue conceptuel, sur la transformation du statu quo inégalitaire à long terme, par exemple en accordant des avantages spécifiques et temporels aux femmes et aux groupes structurellement défavorisés. Un exemple de ces politiques d'égalité est constitué par les systèmes de quotas flexibles selon lesquels, par exemple, un employeur confronté à des candidat-es hommes et femmes également qualifiés dans un processus de recrutement donnera la préférence à la candidate féminine lorsque les femmes sont sous-représentées dans la communauté professionnelle concernée.

Dans le contexte du Conseil de l'Europe, l'action positive n'est pas une obligation juridique, mais elle a été encouragée par la Commission européenne contre le racisme et l'intolérance (ECRI) qui estime qu'il s'agit d'un « outil efficace pour parvenir à des conditions équitables dans la société pour les membres des groupes défavorisés²⁰⁴ ». La Recommandation CM/Rec(2003)3 du Comité des Ministres sur la participation équilibrée des femmes et des hommes à la prise de décision politique et publique adoptée en 2003 encourage également les Etats membres du Conseil de l'Europe à veiller à ce que la représentation des femmes ou des hommes dans tout organe de décision de la vie politique ou publique ne soit pas inférieure à 40 %.²⁰⁵ Tous les États membres du Conseil de l'Europe ont également ratifié la Convention sur l'élimination de toutes les formes de discrimination à l'égard des

204. Commission européenne contre le racisme et l'intolérance, Séminaire réunissant les organismes nationaux spécialisés dans la lutte contre le racisme et la discrimination raciale sur l'action positive : note explicative (2007) : <https://rm.coe.int/seminar-with-national-specialised-bodies-to-combat-racism-and-racial-d/16808b54b0>.

205. Conseil de l'Europe, CM/Rec(2003)3 sur la participation équilibrée des femmes et des hommes à la prise de décision politique et publique, adopté par le Comité des Ministres du Conseil de l'Europe, (12 mars 2003), disponible à l'adresse https://search.coe.int/cm/Pages/result_details.aspx?ObjectID=09000016805e0848.

femmes des Nations Unies, qui autorise clairement les actions positives en stipulant « L'adoption par les Etats parties de mesures temporaires spéciales visant à accélérer l'instauration d'une égalité de fait entre les hommes et les femmes n'est pas considérée comme un acte de discrimination tel qu'il est défini dans la présente Convention, mais ne doit en aucune façon avoir pour conséquence le maintien de normes inégales ou distinctes; ces mesures doivent être abrogées dès que les objectifs en matière d'égalité de chances et de traitement ont été atteints ». ²⁰⁶ Dans l'UE, la législation sur la non-discrimination autorise les mesures spéciales dans le cadre de l'action positive, dans certaines limites telles que l'interdiction d'un quota strict qui donnerait une préférence automatique aux groupes sous-représentés et la nécessité pour les mesures spéciales de viser à transformer le statu quo à long terme ²⁰⁷. La définition de l'action positive dans le contexte du Conseil de l'Europe et de la Convention européenne des droits de l'homme est similaire. La notion de « mesures temporaires spéciales » est souvent utilisée. La Recommandation de politique générale n°7 de l'ECRI indique par exemple que « [l]a loi doit prévoir que l'interdiction de la discrimination raciale n'empêche pas de maintenir ou d'adopter des mesures spéciales temporaires destinées à prévenir ou à compenser les désavantages subis par [les groupes protégés] ou à faciliter leur pleine participation dans tous les domaines de la vie ²⁰⁸ ». Elle indique également que « [c]es mesures ne doivent pas être maintenues une fois atteints les objectifs visés ²⁰⁹ ».

2) Les obligations positives découlant de la CEDH

Pour aborder la question de savoir comment promouvoir l'égalité dans et par l'utilisation de l'IA, la base juridique exposée ci-dessus, qui autorise les actions positives, peut être examinée conjointement avec une autre caractéristique spécifique importante de la CEDH qui est la notion d'**obligations positives**. Ces obligations impliquent que les États ont, dans certaines circonstances, le devoir de prendre activement des mesures pour parvenir à

206. Assemblée générale des Nations unies, Convention sur l'élimination de toutes les formes de discrimination à l'égard des femmes, article 4, paragraphe 1, (18 décembre 1979) disponible à l'adresse <https://www.ohchr.org/en/instruments-mechanisms/instruments/convention-elimination-all-forms-discrimination-against-women>.

207. Pour un compte rendu détaillé, voir Raphaële Xenidis et Héléne Masse-Dessen, « Positive action in practice: some dos and don'ts in the field of EU gender equality law » (2018) 2 *European equality law review* 36.

208. Recommandation de politique générale n° 7 de l'ECRI sur la législation nationale pour lutter contre le racisme et la discrimination raciale (2002), [5].

209. Ibid.

l'égalité et prévenir la discrimination²¹⁰. Elles vont plus loin que les obligations passives ou négatives limitées de ne pas discriminer, parce qu'elles imposent de prendre des mesures préventives contre la discrimination ou des mesures d'action positive pour promouvoir l'égalité comme moyen de se conformer à l'article 14 de la CEDH.

Dans sa Recommandation de politique générale n° 7, l'ECRI souscrit spécifiquement aux obligations positives de promouvoir l'égalité et de prévenir la discrimination sous la forme de dispositions constitutionnelles, d'obligations pour les autorités publiques, ainsi que d'obligations pour les organismes publics de conditionner l'attribution de marchés, de prêts, de subventions ou d'autres avantages au respect de l'obligation positive de promouvoir l'égalité et de prévenir la discrimination²¹¹. Ces obligations peuvent être utilisées comme base juridique pour créer une obligation d'intégration de l'égalité dans le contexte de l'utilisation de l'IA par les autorités publiques.

Les obligations et actions positives constituent un socle juridique intéressant pour utiliser l'IA afin de promouvoir l'égalité à deux égards. On peut en effet considérer que les obligations positives de prévention de la discrimination doivent imposer aux États de recourir à des actions positives afin de créer des garanties pour empêcher l'apparition de biais algorithmiques illégaux à n'importe quel niveau du cycle de vie de l'IA. On peut aussi estimer que les obligations positives de promouvoir l'égalité sont une obligation pour les États d'investir dans l'utilisation des nouvelles opportunités créées par l'IA pour mieux servir les communautés défavorisées afin qu'elles puissent jouir pleinement des droits garantis par la CEDH. Les paragraphes suivants présentent des stratégies pour y parvenir.

3) La nécessité de recentrer l'action positive

L'action positive est **une condition *sine qua non*** pour utiliser l'IA à bon escient. Elle peut prendre de nombreuses formes, notamment des mesures de soutien telles que la diffusion d'informations auprès des communautés ciblées, des programmes de formation et de financement ciblés, des

210. Voir Cour européenne des droits de l'homme, Guide sur l'article 14 de la Convention (interdiction de la discrimination) et sur l'article 1 du Protocole n° 12 (interdiction générale de la discrimination) (2022), [42-43] : https://www.echr.coe.int/Documents/Guide_Art_14_Art_1_Protocol_12_FRA.pdf. Voir aussi, par exemple, Cour européenne des droits de l'homme, requête n° 34369/97 *Thlimmenos c. Grèce* (2 avril 2000) et Cour européenne des droits de l'homme, requête n° 11146/11 *Horváth et Kiss c. Hongrie* (29 janvier 2013).

211. *Ibid.*, [2], [8] et [9].

mesures spéciales temporaires et des systèmes de quotas flexibles²¹². Il faudrait par exemple que les principales priorités incluent la **diversification** des communautés éducatives et professionnelles impliquées dans toutes les phases du développement et du déploiement des applications de l'IA, grâce à un soutien financier et à des efforts de sensibilisation. La diversification peut s'inscrire dans le cadre d'un effort plus large visant à attirer et à retenir davantage de femmes et de filles et de personnes issues de communautés marginalisées dans les domaines STIM.

Si nécessaire, des mesures spéciales temporaires et des systèmes de quotas flexibles seront utilisés pour assurer la parité et l'inclusion dans les communautés éducatives et professionnelles. Des mesures d'action positive sous la forme, par exemple, de mesures spéciales et de lutte contre les stéréotypes devraient viser à rendre ces environnements plus inclusifs afin de fidéliser les groupes minoritaires à long terme et de réduire les taux d'abandon scolaire. Dans le même ordre d'idées, les dispositions relatives à l'intégration des obligations de non-discrimination, y compris l'intégration de la dimension de genre, peuvent être considérées comme une base juridique utile pour réaliser « l'égalité dès la conception » dans le domaine de l'IA.

La **formation** de ces communautés doit passer par une transformation des programmes d'enseignement et l'adoption d'une approche des questions éthiques, des obligations juridiques et des sciences sociales en matière de discrimination et d'inégalité faisant partie intégrante de l'enseignement supérieur et professionnel. Il faudrait également que des formations complémentaires soient dispensées régulièrement aux expert-es, aux parties prenantes et aux communautés professionnelles du secteur de l'IA, sur une base *ad hoc* ou en tant que formation continue. Cette formation doit aborder des questions telles que les inégalités structurelles, l'intégration de la dimension de genre et les stéréotypes. Des formations devraient également être dispensées à d'autres groupes cibles concernés, notamment les organes de surveillance (y compris les organismes de promotion de l'égalité, les institutions nationales des droits humains, les organismes de médiation, etc.) et les organisations de la société civile, les professionnel·les du droit et les juges qui traitent des affaires de droits numériques et de la discrimination.

Une approche centrée sur l'égalité réelle et l'action positive pourrait également nécessiter l'adaptation des dispositions juridiques existantes. En effet, à mesure que l'émergence des nouvelles technologies déplace la dynamique du pouvoir entre les utilisateurs et les sujets des systèmes d'IA, les arrangements judiciaires et les dispositions normatives qui

212. Voir Christopher McCrudden, Resurrecting positive action (2020) 18(2) *International Journal of Constitutional Law*, 429.

sous-tendent les règles juridiques deviennent instables. Rééquilibrer ces asymétries de pouvoir implique donc d'adapter l'architecture juridique. Il est expliqué ci-après que les règles relatives au renversement de la charge de la preuve pourraient être allégées pour les victimes de discrimination algorithmique en posant une présomption de biais algorithmique²¹³. Une telle présomption pourrait être envisagée tant que les usagers d'un système d'IA n'ont pas mis en place de mesures de protection contre la discrimination, c'est-à-dire tant qu'ils supposent que les systèmes d'IA sont neutres vis-à-vis des groupes protégés. On verra ci-après que les protections appropriées peuvent prendre plusieurs formes telles que des audits, des certifications, des évaluations de l'impact sur l'égalité. De plus amples détails sur cette proposition d'adaptation juridique sont présentés à la section 4.

4) L'utilisation de l'analyse des données pour détecter les discriminations

Une deuxième possibilité pour l'IA d'être utilisée pour promouvoir l'égalité consiste à exploiter les capacités de l'analyse des données pour détecter les modèles discriminatoires dans l'allocation des ressources, la diffusion de l'information, la représentation des groupes ou la performance de systèmes donnés. Plusieurs exemples montrent que l'analyse des données peut également être utilisée pour débusquer les mauvais modèles et les pratiques finales qui reproduisent les préjugés. Par exemple, les technologies de reconnaissance d'image de l'IA pourraient être utilisées pour analyser de grandes quantités de données et évaluer les représentations des femmes et des minorités dans différents secteurs des médias, des programmes télévisés aux films, en passant par la publicité en ligne et physique, etc. Dans le domaine de la modération de contenu, l'IA a été utilisée pour détecter les discours de haine afin de signaler et de supprimer les contenus offensants²¹⁴. Dans le même temps, il faut absolument éviter que ce déploiement de l'IA ne réduise

213. À ne pas confondre avec une présomption de discrimination algorithmique, car un tel biais pourrait ou non être discriminatoire. Pour d'autres suggestions sur l'allègement de la charge de la preuve en matière de discrimination algorithmique, voir Janneke Gerards et Raphaële Xenidis, *Algorithmic discrimination in Europe: Challenges and Opportunities for EU Gender Equality and Non-Discrimination Law* (Réseau européen d'experts juridiques dans le domaine de l'égalité des genres et de la non-discrimination / Commission européenne, 2021) et AlgorithmAudit, White Paper: Reversing the burden of proof in the context of (semi-)automated decision-making (2022) : <https://drive.google.com/file/d/1RHdqoGVgwwv-FTv8qC9fAlsVl8eUTcR7s/preview>.

214. Commission européenne contre le racisme et l'intolérance, Recommandation de politique générale n° 15 sur la lutte contre les discours de haine CRI(2016)15, [140] : <https://rm.coe.int/recommandation-de-politique-generale-n-15-de-l-ecri-sur-la-lutte-contr/16808b5b03>. [en français]

au silence les groupes discriminés ou minoritaires²¹⁵. Détecter automatiquement les propos discriminatoires dans les offres d'emploi pourrait également être un moyen de mettre l'IA au service de la promotion de l'égalité. En allant encore plus loin, les systèmes de recommandation pourraient être utilisés pour recommander un langage inclusif alternatif pour remplacer le contenu discriminatoire des offres d'emploi.

5) L'IA comme moyen de rendre des services aux personnes discriminées et communautés défavorisées et d'améliorer l'accessibilité

Au-delà de la détection, les systèmes d'IA peuvent également être développés à dessein pour rendre des services aux personnes discriminées et communautés marginalisées, à risque ou défavorisées. Ils peuvent être utilisés, par exemple, pour améliorer l'accessibilité aux informations ou aux biens et services existants. Le recours à des systèmes automatisés de traduction dans les langues régionales ou minoritaires qui ne sont parlées que par un petit nombre de personnes améliorerait l'accès à des services essentiels. L'IA pourrait également servir à promouvoir l'égalité dans le secteur pénal et policier, par exemple lorsqu'elle est utilisée pour prévenir les risques de violence fondée sur le genre, comme en Espagne avec le logiciel VioGen. Dans le secteur de la santé, l'IA pourrait être utilisée pour améliorer l'accès aux soins dans les zones défavorisées et renforcer les capacités de diagnostic des groupes traditionnellement sous-représentés.

La condition de ces utilisations positives de l'IA est cependant de mobiliser des ressources au profit de la diversification et de la formation des communautés professionnelles impliquées dans son développement et son utilisation et de prendre des mesures d'action positive pour garantir que ces systèmes sont au service des groupes marginalisés. Dans le même temps, le « technosolutionnisme » devrait être évité et l'IA ne devrait pas être perçue comme une panacée pour résoudre le problème de la discrimination. Il est essentiel de se rappeler que les questions sociales nécessitent une approche sociale, et non une approche purement technologique. Si l'IA peut certainement être développée et utilisée pour promouvoir l'égalité, y compris l'égalité de genre, il est important de la considérer comme un outil complémentaire dans le cadre de politiques d'égalité financées comme il se doit et mûrement réfléchies. Ce changement de paradigme impose de réfléchir à l'adoption d'une nouvelle approche.

215. Voir par exemple les effets sexistes et racistes de la modération automatique du contenu : Gerrard Y et Thornham H, 'Content moderation : Social media's sexist assemblages' (2020) 22 1266.

Section 4

Recommandations

Recommandations relatives à la politique des pouvoirs publics : vers une approche de l'IA fondée sur les droits humains

Compte tenu des vides juridiques mis en évidence dans cette étude et de la complexité de la discrimination engendrée par les algorithmes et l'IA, la résolution du problème nécessite une approche multidimensionnelle fondée sur les droits humains. Outre son travail d'élaboration d'une convention générale sur l'intelligence artificielle, le Conseil de l'Europe devrait chercher à jouer un rôle de premier plan en matière de travail normatif dans le domaine particulier de l'égalité, en préparant une recommandation plus spécifique du Comité des Ministres sur l'IA, l'égalité – y compris l'égalité de genre – et la discrimination, qui serait rédigée par un comité d'expert-es placé sous l'égide de la Commission pour l'égalité de genre (GEC) et du Comité directeur sur l'anti-discrimination, la diversité et l'inclusion (CDADI).

Quatre axes complémentaires d'intervention réglementaire et d'action des pouvoirs publics ont été recensés et sont présentés ci-après. Les recommandations de cette étude doivent être lues en parallèle avec les dispositions de la convention générale en cours d'élaboration, qu'elles viennent compléter. Elles s'appuient sur la Recommandation CM/Rec(2020)1 sur les impacts des systèmes algorithmiques sur les droits de l'homme, adoptée en 2020 par le Comité des Ministres, et s'inscrivent dans le droit fil de ce texte²¹⁶.

Les recommandations adressées aux États membres sont principalement destinées à être mises en œuvre dans le secteur public – par exemple en ce qui concerne l'accès à la justice, les voies de recours, la participation démocratique, la sensibilisation du public et le renforcement des capacités –, mais un grand nombre d'entre elles devraient également s'appliquer au secteur privé, notamment les mesures en matière de prévention, de transparence et de responsabilité ainsi que les propositions visant à améliorer la diversité, l'inclusion, la représentation et la participation. En outre,

216. Recommandation CM/Rec(2020)1 du Comité des Ministres aux États membres sur les impacts des systèmes algorithmiques sur les droits de l'homme (adoptée par le Comité des Ministres le 8 avril 2020 lors de la 1373^e réunion des Délégués des Ministres) : https://search.coe.int/cm/pages/result_details.aspx?ObjectId=09000016809e1124.

les recommandations énoncées ci-après ne se portent pas sur une cause spécifique de biais, mais traitent plutôt des risques de discrimination liés à l'utilisation d'algorithmes. Elles concernent donc l'ensemble du cycle de vie des systèmes d'IA, de la conception et de la modélisation jusqu'à la formation (y compris la qualité et la représentativité de la collecte, de la conservation et du traitement des données) et au déploiement des systèmes algorithmiques dans des domaines spécifiques.

Une démarche globale prenant en compte ces quatre volets différents permettra de mettre en place une approche solide et ancrée dans les droits humains. Dans le prolongement de cette étude, il est également proposé que la GEC et le CDADI élaborent, sur la base de ces quatre axes et par le biais d'un comité d'expert-es spécifique, une recommandation du Comité des Ministres sur l'impact des systèmes d'intelligence artificielle, les risques de discrimination qu'ils engendrent et les possibilités qu'ils offrent de promouvoir l'égalité, y compris l'égalité de genre. Cette recommandation sera destinée à compléter le cadre général fixé par la convention mentionnée ci-dessus pour ce qui concerne le principe d'égalité.



Premier axe d'intervention : prévention, transparence et responsabilité

*La présente étude a mis en évidence la nécessité de changer de paradigme dans la législation en vigueur sur la non-discrimination. Alors que les obligations en matière de lutte contre la discrimination sont aujourd'hui axées sur la réparation du préjudice causé, l'étude montre qu'il faut, pour combattre la discrimination algorithmique, compléter les contrôles ex post et les mécanismes de recours existant en instaurant des obligations légales prévoyant des mesures préventives et des garanties ex ante ou en renforçant les obligations déjà en place. La prévisibilité et l'étendue des biais algorithmiques et des dommages qui en découlent appellent une telle approche : **en l'absence de garanties préventives, l'IA reproduit et accentue les inégalités existantes**. Ce changement de paradigme réglementaire se justifie également au regard des nouvelles asymétries de pouvoir et d'information qui surviennent entre fournisseurs, utilisateurs et personnes faisant l'objet des décisions générées par des systèmes de prise de décision algorithmique aux dépens des victimes potentielles de discrimination algorithmique, ainsi que des nouvelles vulnérabilités qui résultent d'un tri social généralisé, alimenté par l'analyse prédictive. Les obligations de transparence et les mécanismes de contrôle peuvent atténuer ces asymétries et renforcer la position des utilisateurs finaux des systèmes algorithmiques et notamment des victimes potentielles de discrimination algorithmique et de ceux qui représentent leurs intérêts.*

1) Les États membres sont encouragés à **étendre le recours aux actions positives pour s'attaquer à la discrimination algorithmique et à mettre à profit la notion d'obligation positive** ancrée dans la jurisprudence de la Cour européenne des droits de l'homme pour soumettre les fournisseurs et les utilisateurs à **une obligation de prévenir de manière satisfaisante la discrimination algorithmique**. Ils pourraient s'inspirer des obligations positives découlant de la CEDH, du droit de l'UE et des dispositions de la Convention relative aux droits des personnes en situation de handicap sur les « aménagements raisonnables » dans le contexte de la discrimination fondée sur le handicap. La neutralité à l'égard des biais algorithmiques n'empêchera pas la discrimination algorithmique. Le principe d'aménagement raisonnable est un repère juridique intéressant pour engager une réflexion sur les mesures préventives. Il permet d'adapter le coût de ces dernières à la taille et à la puissance économique de l'utilisateur concerné, tout en plaçant l'action positive et l'approche concrète de la promotion de l'égalité, y compris l'égalité de genre, au centre des réponses juridiques à la discrimination algorithmique. Des lignes directrices interprétatives portant sur l'article 14 de la CEDH et

confirmant une obligation positive de prévenir la discrimination algorithmique devraient par ailleurs être publiées.

2) Les États membres sont encouragés à **introduire une obligation de mener des évaluations des risques de discrimination et des analyses d'impact sur l'égalité** tout au long du cycle de vie des systèmes algorithmiques, en fonction des usages spécifiques qui sont les leurs. Les obligations de transparence et de justification *ex ante* pourraient par exemple se traduire en une obligation faite aux fournisseurs et aux utilisateurs de systèmes d'IA de procéder à des évaluations préalables des risques de discrimination et de l'impact sur l'égalité, de manière indépendante ou dans le cadre d'une analyse plus vaste des incidences en matière de droits humains. La mise sur le marché et le déploiement d'un système algorithmique donné pourraient être soumis aux résultats des évaluations des risques de discrimination et des analyses d'impact sur l'égalité. Un tel système de « pré-autorisation » oblige les fournisseurs et utilisateurs de systèmes pouvant présenter des risques à prendre des mesures préventives, voire à soumettre leurs systèmes à une procédure de certification ou d'autorisation d'exercice²¹⁷. Les systèmes dont on considère qu'ils posent des risques graves de discrimination devraient faire l'objet d'un examen plus approfondi, comme expliqué ci-après. Le principe juridique de proportionnalité pourrait contribuer à adapter ces obligations à la taille, aux capacités et à la puissance économique des fournisseurs et des utilisateurs. Les évaluations des risques et analyses d'impact devraient être facilement accessibles et rendues publiques. Elles devraient analyser l'impact discriminatoire potentiel des systèmes algorithmiques tout au long de leur cycle de vie et pour l'ensemble des motifs protégés en vertu de la Convention européenne des droits de l'homme. Dans le cadre d'HUDERIA, l'évaluation d'impact sur les droits humains, la démocratie et l'État de droit proposée par le Conseil de l'Europe, les principales composantes d'une telle analyse pourraient être les suivantes :

1. *Identification des risques* : identification des risques concernant l'égalité, y compris l'égalité de genre et la non-discrimination.
2. *Analyse d'impact* : analyse de l'impact, en tenant compte de la probabilité et de la gravité des effets sur les droits en matière d'égalité.
3. *Analyse de la gouvernance* : analyse des rôles et responsabilités des détenteurs d'obligations, des titulaires de droits et des parties

217. Voir G. Malgieri et F. Pasquale, « From Transparency to Justification: Toward Ex Ante Accountability for AI » (2022) Brooklyn Law School, Legal Studies Paper n° 712.

prenantes dans la mise en œuvre et la gouvernance des mécanismes d'atténuation de l'impact.

4. *Atténuation et évaluation*: identification des mesures d'atténuation appropriées et mise en place d'une évaluation continue.

Les États membres sont encouragés à rendre ces évaluations juridiquement contraignantes ou à adopter des mesures d'incitation fortes en intégrant les analyses d'impact aux éléments pris en compte par les juges chargés d'examiner les plaintes pour discrimination algorithmique et de déterminer si les obligations positives relatives à l'égalité de traitement sont respectées. Jusqu'à présent, certaines juridictions ont déjà proposé une législation qui instaurerait des évaluations d'impact des algorithmes pour accroître la transparence des systèmes algorithmiques de plus en plus utilisés dans notre quotidien²¹⁸. Bien que les analyses d'impact suscitent une attention croissante en tant que mécanismes de gouvernance des algorithmes, aucun processus permettant une véritable reddition de comptes n'est pour l'instant normalisé²¹⁹.

3) Des mécanismes de certification pourraient être utilisés en complément pour s'assurer que les biais ont été atténués et les risques de discrimination supprimés dans la mesure du possible pour des usages bien définis. Ces mesures de prévention *ex ante* pourraient servir à établir la nature non discriminatoire d'un système donné dans des domaines spécifiques. La certification devrait bénéficier d'un certain degré de publicité²²⁰. Elle devrait par exemple être rendue accessible et compréhensible par les utilisateurs et utilisatrices finaux/finales de systèmes algorithmiques (en vue d'éventuels

218. Les Pays-Bas ont récemment rendu ces analyses d'impact sur les droits humains obligatoires pour les institutions publiques. Voir la proposition ici : <https://www.tweedekamer.nl/kamerstukken/moties/detail?id=2022Z06024&did=2022D12329>

219. Cela étant, des méthodologies ont été élaborées dans le cadre de travaux de recherche, voir également Mantelero, Alessandro et Esposito, Samantha, *An Evidence-Based Methodology for Human Rights Impact Assessment (HRIA) in the Development of AI Data-Intensive Systems* (March 22, 2021). *Computer Law & Security Review*, 2021 et Mantelero, Alessandro. "Human Rights Impact Assessment and AI" *Beyond Data*. TMC Asser Press, La Haye, 2022. 45-91. Certains exemples, comme l'analyse d'impact néerlandaise sur les droits humains et les algorithmes, peuvent être source d'inspiration, voir Voir Janneke Gerards, Mirko Tobias Schäfer, Arthur Vankan, Iris Muis, "Impact Assessment Fundamental rights and algorithms", Ministry of the Interior and Kingdom Relations (2022), à l'adresse : <https://www.government.nl/binaries/government/documenten/reports/2022/03/31/impact-assessment-fundamental-rights-and-algorithms/Fundamental+Rights+and+Algorithms+Impact+Assessment.pdf> (dernière consultation le 22 juillet 2022).

220. La publicité devrait par ailleurs permettre de corriger dans une certaine mesure les failles de la certification, par exemple en qui concerne les conflits d'intérêts.

recours juridiques et d'évaluations des risques pour certaines applications), les organismes publics (p. ex. organismes de promotion de l'égalité, institutions nationales des droits humains, délégués à la protection des données, etc.) et les organisations de la société civile ayant un intérêt légitime (telles que définies dans les directives de l'UE sur l'égalité et la législation nationale en matière d'égalité). Il pourrait être utile à cet égard de s'inspirer de la Recommandation CM/Rec(2020)1 sur les impacts des systèmes algorithmiques sur les droits de l'homme publiée par le Conseil de l'Europe en 2020. Elle prévoit que « [d]es systèmes de certification fondés sur des normes régionales et internationales devraient être mis au point et appliqués pour garantir la provenance et la qualité des jeux de données et des modèles », et qu'ils « devraient également faire partie des processus d'approvisionnement, reposer sur les cadres réglementaires qui interdisent certaines utilisations de systèmes algorithmiques et s'y conformer »²²¹.

4) Les États membres sont invités à étudier le rapport entre responsabilité, transparence et droit du secret des affaires dans le domaine de l'IA. Pour être explicable, un algorithme doit présenter un certain degré d'accessibilité, que ce soit par des auditeurs internes ou externes, une autorité de régulation ou un tribunal. Cela dit, un algorithme propre à une entreprise peut également être couvert par la législation sur les secrets commerciaux. L'apparition du calcul multipartite sécurisé, qui peut permettre d'interroger une IA sans avoir accès au code réel, apporte des éléments intéressants en la matière. Il convient d'encourager la poursuite des recherches dans ce domaine, afin que le contrôle requis s'agissant des plaintes pour discrimination puisse être adapté de sorte qu'il soit tenu compte de la nécessité de protéger le secret des affaires.

5) Les États membres sont incités à examiner la possibilité de mettre en place des obligations juridiques imposant aux utilisateurs de systèmes d'IA de publier des données statistiques pour permettre aux parties intéressées d'évaluer les effets discriminatoires d'un système donné, notamment des statistiques ventilées concernant l'impact du système sur les groupes protégés par les dispositions anti-discrimination de la CEDH, y compris le risque de discrimination intersectionnelle. Il conviendrait aussi de publier des données sur le fonctionnement des systèmes algorithmiques, notamment (s'il se peut) les critères de décision utilisés et (s'il y a lieu) des informations sur les données d'entraînement et de validation utilisées ainsi que leur traitement. Les obligations de transparence doivent être mises en balance avec

221. Recommandation CM/Rec(2020)1 du Comité des Ministres aux États membres sur les impacts des systèmes algorithmiques sur les droits de l'homme.

les règles relatives aux secrets d'affaires et à la protection des données. En particulier, les États membres sont encouragés à introduire des obligations juridiques pour que les utilisateurs de systèmes algorithmiques fournissent des explications satisfaisantes aux consommateur-ices et aux organismes de contrôle public, à la demande de ces derniers, notamment dans le cadre des procédures judiciaires en cours. Ils pourraient par exemple garantir un droit effectif à une information accessible et intelligible.

6) Les États membres sont encouragés à **mettre en place des mécanismes de transparence**, parmi lesquels peuvent figurer l'établissement de rapports annuels sur le recours à l'IA au sein des collectivités locales et du gouvernement, une obligation faite aux sociétés de rendre compte de l'utilisation responsable de l'IA dans leurs rapports annuels, notamment dans le cadre des obligations en matière environnementale, sociale et de gouvernance (ESG), ou encore l'utilisation de registres des algorithmes. La municipalité d'Amsterdam, par exemple, a créé un tel registre, afin de disposer d'une vue d'ensemble des systèmes d'intelligence artificielle et des algorithmes utilisés par la ville²²².

7) Les États membres devraient regrouper les mesures de prévention, de transparence et de responsabilité dans un **plan d'action global sur l'IA et l'égalité** pour informer le public des initiatives en cours et orienter les interventions concrètes des parties prenantes. Le plan d'action des États membres devrait contribuer à élaborer une stratégie d'égalité « dès la conception » pour intégrer les questions d'égalité, la lutte contre la discrimination et une perspective de genre dans le développement des systèmes algorithmiques et d'IAs. Par ailleurs, les États membres sont encouragés à adopter des politiques visant à faciliter la collecte de données sur l'égalité à l'appui de l'analyse de l'impact discriminatoire des systèmes algorithmiques. Les données sur l'égalité devraient également tenir compte de la discrimination intersectionnelle.

Deuxième axe d'intervention : accès à la justice et mécanismes de recours

Comme le montre cette étude, le déploiement de l'IA et des systèmes algorithmiques ébranle une partie des équilibres de pouvoirs sur lesquels reposent habituellement les éléments centraux du droit relatif à la non-discrimination. En particulier, les nouvelles asymétries d'information mettent à mal certains des principes les plus fondamentaux concernant l'accès des victimes aux preuves,

222. Registre d'algorithmes d'Amsterdam, voir: <https://algorithmeregister.amsterdam.nl/en/ai-register>.

la capacité des victimes à établir, et des défenseurs à réfuter, une présomption de discrimination, l'attribution des responsabilités, l'identification des rapports de causalité en cas de discrimination et l'évaluation de la proportionnalité et des justifications par les juges. Cela compromet l'accès à la justice ainsi que l'intégrité et l'efficacité des recours possibles contre une discrimination algorithmique. En l'absence de réponse claire à la question de savoir comment les citoyens et citoyennes et les consommateurs et consommatrices pourront porter plainte et de quels moyens ils et elles disposeront à cet effet, les étapes suivantes sont encouragées :

1) Les États membres sont encouragés à **faciliter l'accès à la justice en mettant en place des mécanismes de contrôle public et des moyens d'action collective permettant de mener des enquêtes et d'obtenir réparation en cas de discrimination algorithmique**. Les organismes publics comme les organismes de promotion de l'égalité, les institutions nationales de défense des droits humains, les médiateurs et les agences de protection des données devraient être chargés expressément du suivi de la discrimination algorithmique dans les États membres (p. ex. au moyen d'audits des systèmes algorithmiques), de la diffusion d'informations et de la sensibilisation du public, des enquêtes sur les cas potentiels de discrimination algorithmique (évaluation du respect des lois en vigueur par des procédures de test), du soutien aux victimes, notamment par des services gratuits de conseil juridique et d'aide juridictionnelle, ainsi que de la réparation de la discrimination algorithmique par l'émission d'avis contraignants, la représentation des victimes et/ou l'intervention dans les procédures judiciaires. Cela nécessite de former le personnel, d'accroître les capacités et d'assurer un financement adéquat. Il conviendrait à cette fin de renforcer les structures existantes (p. ex. les organismes de promotion de l'égalité en vertu du droit de l'UE sur l'égalité) et de miser sur la coopération et les synergies pour mettre à profit les savoir-faire et compétences à disposition. À cet égard, il sera important de tenir compte des deux propositions de directives relatives aux normes applicables aux organismes pour l'égalité de traitement qui sont en cours de négociation dans l'UE²²³. Si elles sont adoptées, elles viendront renforcer la mission des organismes de promotion de l'égalité en matière de surveillance et d'enquête sur les cas de discrimination algorithmique, ainsi que pour le

223. Proposition de Directive du Parlement européen et du Conseil relative aux normes applicables aux organismes pour l'égalité de traitement dans le domaine de l'égalité de traitement et de l'égalité des chances entre les femmes et les hommes en matière d'emploi et de travail COM(2022) 688 final et Proposition de Directive du Conseil relative aux normes applicables aux organismes pour l'égalité de traitement dans les domaines de l'égalité de traitement entre les personnes en matière de sécurité sociale ainsi que dans l'accès à des biens et services et la fourniture de biens et services COM(2022) 689 final.

soutien aux victimes. Il est par exemple prévu dans les directives proposées de confier des pouvoirs d'enquête supplémentaires aux organismes de promotion de l'égalité, qui pourront également émettre des avis ou des décisions contraignantes, agir en justice et proposer un mécanisme alternatif de résolution des litiges pour les plaintes en matière de discrimination.

2) **Les États membres sont encouragés à ajuster, compléter et consolider les règles de preuve et d'en renforcer l'efficacité pour une répartition plus juste et plus équitable de la charge de la preuve.** Plusieurs éléments peuvent servir d'inspiration :

- ▶ la possibilité d'utiliser la présomption réfragable concernant le respect des obligations en matière d'égalité et de non-discrimination peut être envisagée. Un renversement provisoire de la présomption de légalité des systèmes algorithmiques, par exemple, reflèterait mieux la prévisibilité et l'omniprésence des biais algorithmiques ainsi que le glissement des rapports de force décrit précédemment. Cette proposition va dans le sens des travaux menés par l'UE dans le cadre de la législation sur l'IA et concorderait avec la proposition de l'UE de créer une présomption réfragable de causalité à l'égard des défendeurs en cas de dommages liés à l'IA²²⁴. La mise en place d'une telle présomption et, parallèlement, d'obligations de divulgation, entraînerait un allègement de la charge de la preuve de la victime pour ce qui est de démontrer la présomption de discrimination. Cela rejoint la proposition d'une « défiance dès la conception²²⁵ ». On peut soutenir que la prévisibilité des dommages liés aux algorithmes et les déséquilibres de pouvoir actuels justifient l'existence d'une présomption réfragable de biais algorithmique ;
- ▶ les fournisseurs et utilisateurs de systèmes algorithmiques pourraient éviter de telles présomptions en montrant qu'ils ont pris des mesures préventives efficaces et suffisantes. Ces mesures comprendraient, par exemple, une évaluation du risque de discrimination et une analyse de l'impact sur l'égalité de traitement, ainsi qu'une certification (voir ci-dessus) ;
- ▶ en l'absence de telles mesures de prévention, il serait possible d'adopter une présomption provisoire de biais algorithmique ayant pour effet

224. Comme expliqué à la section 2.III, la récente proposition de directive sur la responsabilité en matière d'IA, et notamment les éléments concernant la présomption réfragable et les obligations de divulgation, pourrait servir de base à des dispositions en vue d'alléger la charge de la preuve dans les cas de discrimination algorithmique.

225. Malgieri G and Pasquale F, 'From Transparency to Justification: Toward Ex Ante Accountability for AI' (2022) Brooklyn Law School, Legal Studies Paper N. 712.

de déplacer la charge de la preuve vers le défendeur, en s'appuyant sur les règles existantes en matière de déplacement de la charge de la preuve. Les utilisateurs et/ou fournisseurs de systèmes algorithmiques devraient alors montrer que le système est conforme aux exigences des lois anti-discrimination ;

- ▶ de tels mécanismes de responsabilité *ex ante* aideraient également les juges et les organismes de promotion de l'égalité exerçant une fonction décisionnelle dans les affaires de discrimination à appliquer le test de proportionnalité et permettraient de mieux évaluer le caractère adéquat des garanties mises en place pour la prévention. Cela inciterait les utilisateurs et fournisseurs d'IA à traduire les choix techniques en informations compréhensibles sur le plan juridique, informations qui pourront ensuite être utilisées pour déterminer si un système donné répond à un objectif légitime et si les moyens employés sont raisonnablement proportionnés au but poursuivi ;
- ▶ s'agissant des responsabilités, les États membres pourraient explorer le concept juridique de « négligence » pour concevoir des réponses juridiques dans le cas où les fournisseurs et/ou utilisateurs d'un système discriminatoire n'auraient pas pris de mesures de prévention des biais algorithmiques. En l'absence de personnalité juridique des systèmes d'IA, les utilisateurs et fournisseurs devraient être stratégiquement tenus responsables en cas de discrimination algorithmique, même lorsque celle-ci est produite de manière « autonome » par le système discriminatoire. Le manquement à l'obligation de prévenir raisonnablement la discrimination algorithmique pourrait donner lieu à une présomption réfragable, comme indiqué précédemment ;
- ▶ la responsabilité en cas de discrimination algorithmique devrait déboucher sur une obligation claire de supprimer tout effet discriminatoire pour certains types d'usages et de réparer les dommages (matériels et moraux) subis du fait de la discrimination algorithmique. Il conviendrait d'adopter des délais précis pour l'élimination des impacts discriminatoires dans les systèmes d'IA. Une obligation de publication pourrait également contribuer à l'application des règles de responsabilité, par exemple en rendant obligatoire la communication d'informations au public par toute entreprise privée jugée responsable de discrimination algorithmique.

3) **Encourager la coopération entre les autorités de régulation.** À titre d'exemple, les autorités de contrôle des services financiers devraient pouvoir travailler avec les autorités de protection des données, notamment en ce qui concerne l'utilisation de données à caractère personnel dans le contexte de

l'attribution de prêts ou pour des enquêtes sur d'éventuels biais. Les régulateurs devraient également mener des recherches sur l'utilisation de technologies renforçant la protection de la vie privée et de mécanismes de vérification standardisés. Dans le même esprit, les États membres sont encouragés à faciliter la coopération entre les autorités de protection des données, les organismes de promotion de l'égalité, les institutions nationales des droits humains et les agences de protection des consommateur-ices.

4) Conformément au projet de cadre du CAI, qui prête attention à la discrimination mais aussi à la vulnérabilité, les États membres sont invités à **étudier les nouvelles formes de vulnérabilité « algorithmique » liées à l'utilisation des systèmes d'IA et à promouvoir la recherche sur les méthodes de protection juridique** contre cette vulnérabilité. L'utilisation généralisée des systèmes d'IA produit des formes extrêmes de tri social et de différenciation sociale qui remettent en cause les limites du droit relatif à la discrimination. Ce tri social génère de nouvelles formes de vulnérabilité sociale, par exemple en appliquant systématiquement à certains groupes algorithmiques de moins bonnes conditions économiques, ou en les excluant de certains biens et services. Ces nouvelles formes de différenciation sociale risquent de créer une discrimination « émergente » et de profondes inégalités socio-économiques.²²⁶ Il importe aussi de reconnaître les nouvelles situations d'infériorité, de dépendance et d'assujettissement auxquelles sont confrontées les personnes vulnérables dans le contexte du traitement des données.

5) Les États membres sont incités à **affirmer sans ambiguïté que l'interdiction de la discrimination énoncée à l'article 14 concerne la discrimination intersectionnelle et la discrimination par procuration**, deux formes d'atteintes discriminatoires que les systèmes algorithmiques sont particulièrement susceptibles de provoquer. Au niveau du Conseil de l'Europe, une assistance pourrait être apportée en élaborant des recommandations et des lignes directrices sur l'interprétation de l'article 14 de la CEDH.

6) Dans le droit-fil des travaux du CAI, les États membres sont encouragés à étudier les possibilités **d'utiliser le droit de la protection du consommateur pour compléter la lutte contre la discrimination**, par exemple en

226. Voir Matthias Leese, *The new profiling: Algorithms, black boxes, and the fail of anti-discriminatory safeguards in the European Union*, 45 SECURITY DIALOGUE 494–511, 501 (2014) ; Monique Mann & Tobias Matzner, *Challenging algorithmic profiling: The limits of data protection and anti-discrimination in responding to Emerging discrimination*, 6 BIG DATA & SOCIETY, 5–6 (2019).

facilitant l'accès à l'information, en interdisant certaines caractéristiques dans les systèmes algorithmiques conformément à la notion de clause abusive, etc.

Troisième axe d'intervention : diversité, inclusion, représentation et participation

La présente étude souligne la nécessité d'éviter le technocentrisme et le solutionnisme et de se focaliser sur la discrimination, ses composantes sociales et ses origines plutôt que sur les biais techniques²²⁷. L'un des aspects centraux de l'intégration de l'égalité dans l'IA consiste à assurer la diversité et l'inclusion par la représentation et la participation des femmes et des groupes touchés par la discrimination au sein des communautés professionnelles concernées, ainsi que par une formation spécialisée. En particulier, les établissements d'enseignement et les entreprises devraient avoir l'obligation de promouvoir une culture de l'ouverture, de l'inclusion et de la diversité par des mesures d'action positive visant à étendre l'accès aux secteurs professionnels liés à l'IA et à favoriser la réussite dans ces derniers. Par conséquent, nous formulons les recommandations suivantes :

1) Les États membres devraient **identifier et soutenir des mesures d'action positive**, notamment d'augmentation de la participation des femmes et de diversification des communautés professionnelles, et appliquer activement les obligations positives de promotion de l'égalité dans l'IA. Ils devraient également envisager de **rendre certaines formes d'action positive juridiquement contraignantes** dans le cadre de la lutte contre la discrimination algorithmique. Les États membres sont encouragés en particulier à **donner une place centrale à l'action positive dans les secteurs professionnels participant au développement et au déploiement non discriminatoires des systèmes d'IA et à l'analyse des risques y relatifs**. Les mesures d'action positive devraient s'appliquer à la communauté professionnelle de l'IA, et concourir par exemple à l'intégration durable des femmes et des filles (mise en œuvre de mesures en faveur de l'équilibre entre vie professionnelle et vie privée, entre autres) et des personnes d'origines ethniques, religieuses et linguistiques diverses, des personnes LGBTI et d'autres groupes protégés ou défavorisés dans les programmes de formation et les secteurs professionnels

227. Mantelero et Fanucci appellent par exemple à « adopter une perspective plus large sur le secteur de l'IA en considérant la totalité de sa chaîne d'approvisionnement sous l'angle des droits humains ». Voir Mantelero, Alessandro et Fanucci, Francesca, *Great Ambitions. The International Debate on AI Regulation and Human Rights in the Prism of the Council of Europe's CAHAI* (4 avril 2022). Philip Czech et coll. (sous la dir. de), *European Yearbook on Human Rights 2022* (Intersentia, à paraître).

concernés²²⁸. Plusieurs types d'actions peuvent être envisagées à cette fin, de la sensibilisation au soutien financier et à la formation spécialisée en passant par les quotas. Au nombre de ces mesures pourraient par exemple figurer :

- ▶ des politiques d'embauche et de promotion visant à atteindre une représentation équilibrée des personnes exposées à la discrimination aux postes les plus élevés et des programmes de soutien pour les aider à se positionner pour réussir, notamment une obligation de recruter et de promouvoir des femmes à au moins 50 % des postes les plus élevés dans les secteurs de la technologie ;
- ▶ des quotas et des programmes de bourse pour soutenir la représentation des femmes et des filles et des personnes exposées à la discrimination dans les études de sciences, de technologie, d'ingénierie et de mathématiques (STIM) ;
- ▶ des politiques sectorielles et/ou d'entreprise sur la lutte contre les stéréotypes, la discrimination, le harcèlement et la violence à l'égard des femmes et des groupes protégés, notamment des politiques sur le lieu de travail favorisant l'inclusion et la réintégration à la suite d'une interruption de la carrière professionnelle ;
- ▶ des programmes de formation visant à sensibiliser la main-d'œuvre d'un secteur aux effets discriminatoires de l'IA et aux stratégies pour les prévenir.

2) **Les obligations positives de promouvoir l'égalité devraient donner une base juridique pour faire en sorte que la promotion de l'égalité soit au cœur de l'élaboration des systèmes d'IA et des systèmes algorithmiques.**

Une surveillance plus active du respect de ces obligations pourrait entraîner un changement de paradigme dans les communautés professionnelles concernées et permettrait de veiller à ce que les produits algorithmiques et l'IA soient conçus et développés en gardant l'égalité présente à l'esprit. Par exemple, les États membres pourraient établir des **règles de passation de marchés** qui incluent comme critère d'admissibilité une exigence de diversité dans les équipes professionnelles chargées du développement des systèmes d'IA, ainsi que l'obligation de démontrer comment le respect de la

228. Pour des exemples d'études de référence, voir notamment Institut européen pour l'égalité entre les hommes et les femmes (EIGE), *Study and work in the EU: Set apart by gender* (2018) : <https://eige.europa.eu/publications/study-and-work-eu-set-apart-gender-report> ; EIGE, *Gender Equality Index 2020 Digitalisation and the future of work* (2020) : <https://eige.europa.eu/publications/gender-equality-index-2020-digitalisation-and-future-work> ; et EIGE, *Artificial intelligence, platform work and gender equality* (2020) : <https://eige.europa.eu/publications/artificial-intelligence-platform-work-and-gender-equality>.

règle de droit a été incorporé dans le système d'IA utilisé par les pouvoirs publics.

3) Les obligations positives de promouvoir l'égalité pourraient également se traduire par l'obligation pour les entreprises du secteur de l'IA **de développer et de mettre en œuvre une stratégie d'égalité** couvrant les groupes protégés par l'article 14 de la CEDH et l'article 1 du Protocole 12 à la CEDH. Les États membres pourraient également encourager ou exiger des entreprises du secteur de l'IA qu'elles **désignent un responsable chargé** :

- ▶ **de superviser l'application de cette stratégie** ;
- ▶ **de nouer le dialogue entre les équipes juridiques et techniques** afin de faciliter l'introduction d'exigences légales dans la conception des systèmes d'IA ; et
- ▶ **de coopérer avec les autorités de réglementation et d'exécution pour démontrer la conformité** des systèmes développés par l'entreprise avec la législation sur la non-discrimination ou, le cas échéant, apporter des modifications pour assurer la conformité ou retirer ces systèmes du marché²²⁹.

Quatrième axe d'intervention : participation démocratique, sensibilisation de l'opinion publique et renforcement des capacités

L'IA transforme indéniablement nos sociétés. C'est pourquoi, en parallèle des exigences ex ante, du contrôle du respect de la réglementation et des évaluations ex post, il est crucial que les États membres investissent dans l'éducation des citoyen·nes et des consommateur·ices et consommatrices afin que chacun et chacune dispose de tous les moyens d'action nécessaires pour exercer leur citoyenneté numérique. Par conséquent :

1) Les États membres sont encouragés à instituer un **droit à l'information sur la médiation algorithmique** dans le contexte de plaintes ou de réclamations pour discrimination. Toutes les organisations devraient avoir l'obligation d'indiquer aux utilisateurs et utilisatrices s'ils interagissent avec un être humain ou une machine. Les utilisateurs devraient également être informés de la manière dont les décisions sont prises et des moyens de les contester.

229. Dans ce sens, voir Yeung, Karen and Harkens, Adam, How Do 'Technical' Design-Choices Made When Building Algorithmic Decisionmaking Tools for Criminal Justice Authorities Create Constitutional Dangers? (Part 1) (December 7, 2022). Public Law, Forthcoming, p. 3.

2) Les États membres devraient promouvoir la **mise en place de programmes d'éducation au numérique**, en veillant notamment à tenir compte du contexte, pour renforcer la connaissance des droits à l'égalité, notamment l'égalité de genre, et à la non-discrimination dans le contexte des applications de l'IA²³⁰. Les États membres devraient également encourager **une culture favorisant la négociation collective en ce qui concerne la numérisation du lieu de travail et la participation des acteurs concernés aux décisions en matière de gestion algorithmique**. La Confédération générale italienne du travail (CGIL) a ainsi proposé que la manière dont les systèmes algorithmiques interviennent dans l'organisation du travail et dans les processus de travail soit examinée et négociée dans le cadre des négociations syndicales²³¹.

3) Les États membres sont encouragés à renforcer les exigences légales applicables à la participation démocratique dans l'élaboration des normes. Les normes joueront un rôle crucial dans l'exposition des systèmes d'AI aux forces du marché. Cela dit, le processus normatif n'est guère ouvert à l'examen public et les organisations et individus, notamment les plus vulnérables, y participent peu. Par conséquent, les États membres sont encouragés à prendre les mesures suivantes :

- ▶ **identifier les bonnes pratiques** en matière de démocratisation du processus d'élaboration de normes relatives à l'IA ;
- ▶ **établir des obligations légales de participation démocratique** aux processus d'élaboration de normes relatives à l'IA. La normalisation dans le domaine de l'IA jouant un rôle de plus en plus important, y compris pour l'égalité et la non-discrimination, il existe un risque intrinsèque que les normes soient utilisées pour définir et interpréter des obligations légales ayant un impact sur les droits humains, d'autant plus que cela passe par des organismes de normalisation privés auxquels la société civile ne participe que de manière limitée. Par conséquent, afin d'assurer la représentation des femmes et des groupes touchés par la discrimination et des spécialistes des questions d'égalité, les États membres sont encouragés à faciliter la participation de ces groupes, et les ONG ayant un intérêt légitime, aux processus d'élaboration des normes. Des consultations et une surveillance

230. La question de l'accès aux applications d'IA se pose également, en particulier si un mur payant en limite l'utilisation.

231. Voir Daniele Carchidi, 'Contrattare per governare gli impatti della digitalizzazione sul mondo del lavoro: il caso afiniti' (2022) SLC-CGIL disponible à l'adresse: <https://www.slc-cgil.it/notizie-tlc-ed-emittenza/3791-afiniti-un-caso-riuscito-di-contrattazione-dell-algoritmo.html>.

publiques s'imposent car le processus normatif risque de s'orienter vers les secteurs de l'action publique et du droit qui peuvent nécessiter un certain degré d'interprétation, par exemple en ce qui concerne les biais dans les données.

4) Le renforcement des capacités devrait passer par un investissement dans des travaux de recherche interdisciplinaire sur les algorithmes non-discriminatoires et dans des stratégies visant à protéger l'égalité dans l'utilisation des systèmes algorithmiques. Cependant, de récentes études ont confirmé qu'il n'est guère envisageable de prévenir la discrimination algorithmique en supprimant les biais algorithmiques, par exemple par le débiaisage des ensembles de données, à la fois du fait de la complexité de ces biais et de l'évolution des systèmes d'apprentissage-machine, notamment lorsqu'ils sont mis en œuvre dans des contextes sociaux dynamiques²³². Bien que l'existence de biais soit prévisible, on ne peut prédire et contrôler tous les biais possibles et la discrimination qui en découle²³³. Le débiaisage ne peut donc constituer qu'une partie des mesures préventives et devrait être **complété par des mesures s'attaquant aux racines sociétales de la discrimination.**

232. Voir Balayn A. et Gürses S., *Beyond Debiasing : Regulating AI and its inequalities* (European Digital Rights 2021).

233. *Ibid.*

www.coe.int

Le Conseil de l'Europe est la principale organisation de défense des droits de l'homme du continent. Il comprend 46 États membres, dont l'ensemble des membres de l'Union européenne. Tous les États membres du Conseil de l'Europe ont signé la Convention européenne des droits de l'homme, un traité visant à protéger les droits de l'homme, la démocratie et l'État de droit. La Cour européenne des droits de l'homme contrôle la mise en œuvre de la Convention dans les États membres.

COUNCIL OF EUROPE



CONSEIL DE L'EUROPE