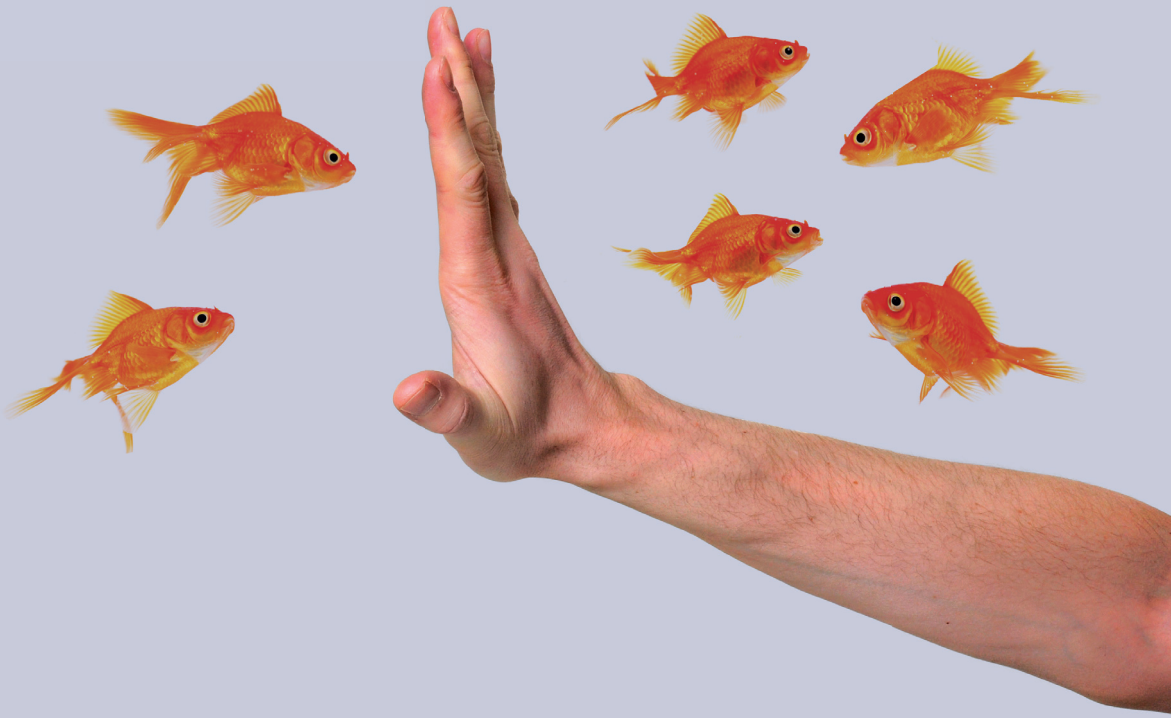
INSTITUT
MONTAIGNE

Algorithmes :
contrôle des biais S.V.P.



RAPPORT MARS 2020

Think tank indépendant créé en 2000, l'Institut Montaigne est une plateforme de réflexion, de propositions et d'expérimentations consacrée aux politiques publiques en France et en Europe. À travers ses publications et les événements qu'il organise, il souhaite jouer pleinement son rôle d'acteur du débat démocratique avec une approche transpartisane. Ses travaux sont le fruit d'une méthode d'analyse et de recherche rigoureuse et critique, ouverte sur les comparaisons internationales. Association à but non lucratif, l'Institut Montaigne réunit des chefs d'entreprise, des hauts fonctionnaires, des universitaires et des personnalités issues d'horizons divers. Ses financements sont exclusivement privés, aucune contribution n'excédant 1,5% d'un budget annuel de 6,5 millions d'euros.

INSTITUT
MONTAIGNE

Algorithmes : contrôle des biais S.V.P.

RAPPORT – MARS 2020

Il n'est désir plus naturel
que le désir de connaissance

SOMMAIRE

INTRODUCTION

En 2016, une enquête révélait que le logiciel COMPAS de prédiction de récidive pour les criminels discriminait les populations afro-américaines alors même qu'il était utilisé au quotidien par les juges américains pour décider d'accorder ou non des libérations sous caution. Début 2019, des chercheurs accusaient l'algorithme de Facebook de recommandation d'offres d'emploi de promouvoir moins fréquemment les femmes. Fin 2019, un algorithme d'Apple Card était dénoncé pour sa discrimination envers les femmes en limitant automatiquement leurs plafonds de carte bancaire. À l'occasion de ces polémiques et d'autres, les biais des algorithmes sont de plus en plus pointés du doigt aux États-Unis.

Ce rapport tente de donner une perspective française à cette problématique aujourd'hui essentiellement traitée sous un prisme américain. Il existe en effet peu d'exemples publics de biais algorithmiques en France ou en Europe continentale, notamment de biais conduisant à des discriminations relatives aux 25 critères protégés par la loi française (âge, genre, religion…[1]). Les rares cas avérés impliquent des acteurs américains. Le numérique et l'intelligence artificielle ne pourront pourtant pas se développer en France s'ils sont porteurs de discriminations massives. Une telle automatisation à grande échelle de décisions inéquitables serait en effet inacceptable pour la société.

Notre travail poursuit l'étude de Télécom Paris et de la Fondation Abeona, *Algorithmes : biais, discrimination et équité* publiée en 2019. Sur la base de ce constat technique, nous avons voulu, à travers la quarantaine d'entretiens réalisés et plusieurs réunions d'un groupe de travail de personnalités qualifiées, apporter des solutions concrètes pour que chacun puisse s'assurer en France d'un déploiement d'algorithmes à la fois éthiques et utiles.

1 Origine, sexe, âge, situation familiale, grossesse, apparence physique, situation économique, patronyme, état de santé, perte d'autonomie, handicap, moeurs, caractéristiques génétiques, orientation sexuelle, genre, opinions politiques ou philosophiques, langues, appartenance réelle ou supposée à une ethnie, nation ou prétendue race, lieu de résidence, domiciliation bancaire.

RÉSUMÉ

LES CONSTATS ET ENJEUX

Malgré le risque de biais dans certains cas, les algorithmes sont, à bien des égards, un progrès en matière de lutte contre les discriminations. Les hommes et les femmes portent régulièrement, bien que parfois inconsciemment, des jugements biaisés. Ils sont inconstants dans leurs décisions. **Utiliser un algorithme revient à formaliser des règles applicables à tous, à en mesurer les résultats et donc à se donner les moyens d'assurer l'absence de biais.**

Les biais algorithmiques conduisant à des discriminations sont rarement dus à un code erroné de l'algorithme. Les données, incomplètes, de mauvaise qualité, ou reflétant les biais présents dans la société, sont bien plus souvent à l'origine de ces biais. Le combat des biais algorithmiques est donc avant tout un combat contre des discriminations déjà existantes au quotidien. L'enjeu n'est pas seulement de produire des algorithmes équitables mais aussi de réduire les discriminations dans la société.

Ce combat est difficile à plusieurs égards. Tout d'abord, **définir ce que serait un algorithme sans biais est complexe.** Certains biais sont volontaires, comme le fait de promouvoir les boursiers dans le cursus scolaire. D'autres biais sont inconscients, et conduisent à discriminer certains groupes.

Un algorithme traitant les individus de manière équitable s'approche d'un algorithme sans biais indésirable, sans pour autant le garantir. **Le détour par l'équité ne clôt pourtant pas le débat, car l'équité d'un algorithme peut prendre des formes multiples.** L'appréciation de ce qui est juste comporte une dimension culturelle et morale. L'attitude équitable ne sera pas la même selon qu'on parle d'un algorithme d'analyse de radiographie des poumons ou de recommandation de publicités politiques. L'équité totale entre individus et l'équité totale entre groupes sont mathématiquement incompatibles. **Il y aura toujours des choix à faire, des choix de société, des choix politiques.**

Ensuite, **corriger un algorithme pour le rendre équitable, c'est souvent réduire sa performance.** Lorsqu'on développe un algorithme, on choisit une ou plusieurs métriques qui permettent de l'optimiser et d'évaluer s'il remplit bien sa tâche. Ces métriques définissent sa performance. Ajouter une contrainte, c'est limiter la capacité d'optimiser l'algorithme vis-à-vis de son critère de performance initial. Il est toujours plus difficile de poursuivre plusieurs buts à la fois plutôt qu'un seul. Il sera donc complexe et coûteux pour de nombreux acteurs de produire des algorithmes performants et équitables.

Enfin, **lutter contre les biais des algorithmes consiste à réaliser une synthèse entre la protection des citoyens contre les discriminations et la possibilité d'expérimenter, cruciale dans l'économie numérique.** Restreindre fortement l'usage des algorithmes sur la suspicion qu'ils pourraient avoir des biais, c'est se priver de nouveaux outils pouvant objectiver nos décisions, c'est brider l'industrie française du numérique et subir à long terme une domination technologique américaine et chinoise. Laisser faire, c'est ignorer le potentiel de destruction de telles innovations sur notre tissu social.

RÉSUMÉ

LES RECOMMANDATIONS

Face à ces enjeux, il faut être clair : nous ne recommandons ni une loi contre les biais des algorithmes commune à tous les secteurs d'activités, ni un contrôle systématique par l'État de l'absence de biais dans les algorithmes.

Il existe d'ores et déjà de nombreux textes s'attaquant aux discriminations. Ceux-ci s'appliquent au monde physique comme au monde numérique et sont de nature à limiter le risque de biais,. Compte tenu du faible recul dont nous disposons, une loi spécifique aux biais risquerait de bloquer toute innovation sans même résoudre le problème de fond.

Le règlement général sur la protection des données (RGPD) a montré que l'usage des données personnelles est bien trop répandu pour qu'une autorité administrative puisse en contrôler l'intégralité avant leur utilisation. Nous pensons qu'il en sera de même pour les algorithmes et qu'il est illusoire d'attendre de l'État qu'il contrôle chacun des algorithmes pour s'assurer de leur caractère éthique avant leur utilisation.

Nous nous sommes attachés à formuler des recommandations les plus réalistes possible afin de permettre un développement rapide des nouvelles technologies dans un cadre respectueux de nos modes de vie et de nos valeurs.

> *Tester la présence de biais dans les algorithmes comme l'on teste les effets secondaires des médicaments.*

À l'instar des nouveaux médicaments, il est difficile de comprendre le fonctionnement de tous les algorithmes, notamment ceux utilisant l'intelligence artificielle. Par ailleurs, comprendre leur fonctionnement ne garantit pas qu'il n'aura pas de biais algorithmiques. C'est *in fine* par le test de l'absence de biais qu'il est possible de créer la confiance dans le caractère équitable des algorithmes.

Tester l'équité d'un algorithme a un coût et nécessite des données de test qui incluent spécifiquement certaines données sensibles (genre, origine sociale). Les développeurs et acheteurs d'algorithmes devront intégrer cette contrainte, et recourir à des tests fonctionnels ou de performance pour s'assurer de l'absence de biais. Dans certains cas où la création de ces bases de données est difficile ou problématique, l'État pourrait la prendre en charge.

Promouvoir une équité active, plutôt que d'espérer l'équité en ne mesurant pas la diversité

Pour lutter contre les discriminations, la France a longtemps fait le choix de ne reconnaître que des citoyens, égaux en droit, plutôt que des individus divers. En ce qui concerne les algorithmes, cette approche n'est pas pertinente. Un algorithme peut introduire des biais contre les femmes, même si l'on a explicitement exclu le genre des variables utilisées — il est en effet facile de déduire cette information à partir d'autres informations, comme le fait d'acheter en ligne des soins pour femme . Pour lutter contre les discriminations, il faut donc pouvoir les détecter.

Cela suppose de passer d'une approche qui attend l'équité par l'ignorance de la diversité à une équité active, c'est-à-dire accepter que l'équité d'un algorithme ne s'obtient pas en excluant toutes les variables protégées comme le sexe, l'âge ou encore la religion mais au contraire en les incluant et en testant l'indépendance du résultat vis-à-vis de ces variables. Pour ce faire, il est nécessaire de disposer de ces informations protégées. Mais si ces informations sont protégées, c'est justement parce qu'elles peuvent être source de discrimination. La collecte et l'utilisation de ces données doivent donc être strictement encadrées! Afin d'éviter les dérives, une telle collecte serait limitée aux tests de la présence de biais et elle serait restreinte à un échantillon des utilisateurs concernés. Par ailleurs, une telle approche devrait faire l'objet d'une analyse d'impact déclarée à la CNIL de manière préalable. Enfin, la nature des algorithmes testés devra justifier la collecte de telles données.

Être plus exigeant pour les algorithmes ayant un fort impact sur les personnes (droits fondamentaux, sécurité, accès aux services essentiels)

La sensibilité d'un algorithme vis-à-vis de la société dépend certes de son secteur d'activité mais surtout de son impact potentiel sur les citoyens. Celui-ci est fort dès lors que l'algorithme peut restreindre l'accès à des services essentiels comme un compte bancaire ou la recherche d'un emploi, mettre en danger la sécurité (santé, police), ou bafouer des droits fondamentaux. Ces domaines font déjà l'objet d'obligations fortes en matière de discrimination. Lorsqu'un algorithme y est introduit, cela ne peut être au prix d'une diminution des exigences.

Pour ces algorithmes, nous recommandons un cadre ad hoc prévoyant des obligations de transparence en ce qui concerne les données utilisées et les objectifs fixés à l'algorithme ainsi qu'un droit de recours contre la décision prise. La création d'un tel cadre ne nécessite pas une nouvelle loi sur les biais des algorithmes mais plutôt la mise en œuvre de bonnes pratiques dans les entreprises et administrations, l'usage de dispositifs juridiques existants et l'ajout au cas par cas de dispositions dans des législations sectorielles.

Assurer la diversité des équipes de conception et de déploiement des algorithmes

Les algorithmes transforment les business models des entreprises. Pour cette raison, il importe que les responsables et utilisateurs soient de plus en plus impliqués dans leur conception.

Définir le comportement équitable de l'algorithme permet d'en modifier profondément les impacts économiques et sociétaux. Il est plus que jamais nécessaire d'intégrer une diversité de perspectives dans la prise de ce type de décision qui ne peut être le seul fait d'experts techniques. Au-delà de la diversité professionnelle, il est désormais clair que des équipes socialement diverses sont mieux armées pour prévenir les biais, pour éviter de reproduire des discriminations.

Intégrer une diversité de profils, de compétences, d'expériences, d'âges, de genres dans les équipes de conception, de production et de pilotage des algorithmes doit devenir une norme pour prévenir les biais algorithmiques.

Et pour aller plus loin…

Au-delà de ces quatre recommandations principales, nous sommes convaincus qu'un important travail demeure nécessaire en matière de formation. Cela doit concerner les chercheurs et développeurs bien sûr, notamment pour les biais algorithmiques, mais également les dirigeants et citoyens sur le cadre plus général de l'intelligence artificielle, afin que chacun puisse s'approprier les opportunités et les dangers de cette technologie.

La vigilance sera également renforcée dans les organisations qui décideront de mettre en œuvre des chartes et bonnes pratiques. Il faut encourager ces initiatives qui, nous avons pu le constater lors de nos entretiens, génèrent une prise de conscience collective sur les dangers des biais algorithmiques, en complément des mesures techniques et opérationnelles.

Enfin, la vigilance doit être extérieure aux organisations. Il semble utile, dans le cas d'algorithmes à fort impact, de renforcer les exigences. Il faut d'abord soutenir l'émergence de labels qui garantissent la qualité des données utilisées et de l'organisation qui développe l'algorithme, l'existence de procédures de contrôle ou encore l'auditabilité de l'algorithme. Les acteurs économiques auront besoin de telles garanties pour s'emparer des algorithmes, et tirer pleinement parti de la révolution qu'ils représentent. Pour les algorithmes à fort impact, une capacité d'audit et de contrôle de certaines exigences pourrait être confiée à une tierce partie ou à l'État.

I

LES ALGORITHMES SONT À LA FOIS CATALYSEURS ET INHIBITEURS DE DISCRIMINATIONS

L'intelligence artificielle (IA) – ou plutôt l'apprentissage machine – a fait de formidables progrès depuis quelques années. Tous les domaines de nos vies contemporaines laissent des traces numériques, génèrent des données, qui peuvent servir à développer des algorithmes. Ceux-ci nous aident à prédire des défauts de pièces mécaniques, à reconnaître des tumeurs, ou à proposer un taux d'emprunt bancaire.

La décennie 2010 a été celle de la montée en puissance de l'industrie de la tech et de ses services renforcés par l'intelligence artificielle. Malgré toutes ses promesses, malgré la victoire des geeks, la décennie s'est achevée sur deux années de *techlash*, de contestation contre l'industrie de la tech, et particulièrement ces géants qui la dominent. Les inquiétudes liées aux conséquences de l'intelligence artificielle se multiplient : pour notre travail, nos médias, nos démocraties.

L'IA toute puissante, dépassant l'humain dans son intelligence, n'est pas pour demain. Avant de la redouter, il est urgent de regarder en face la façon dont les algorithmes s'insèrent d'ores et déjà dans notre quotidien.

Les algorithmes sont parés des atouts de la neutralité et de l'objectivité. Mais ils sont intégrés dans nos sociétés, et en prolongent les défauts et les inégalités structurelles. Nombreux sont les exemples qui documentent la façon dont les algorithmes trient, classent et excluent certains groupes[2]. Comme les bases de données digitales, comme leurs ancêtres de papier avant eux, les algorithmes sont susceptibles d'aggraver certaines inégalités.

Dans nos sociétés occidentales, la diffusion des algorithmes est concomitante d'une plus grande prise en compte des discriminations et plus largement des désavantages structurels entre les groupes d'individus. Que l'on parle de genre, d'origine ethnique, d'orientation sexuelle ou d'âge, notre sensibilité aux inégalités entre groupes grandit.

2 Voir l'article de Bertail P., Bounie D., Clémençon S. et Waelbroeck P., Algorithmes: biais, discrimination et équité, 2019.

L'étude des « biais algorithmiques » – de l'impact des algorithmes sur les discriminations – est largement dominée par des voix et des cas anglo-saxons. Qu'il s'agisse de la justice, du recrutement, du crédit, de la reconnaissance faciale, les polémiques les plus vives ont tendance à venir des États-Unis.[3]

Le débat est-il pour autant identique en France ? Les algorithmes ne sont pas développés de la même façon, nous disposons d'une protection plus stricte sur la collecte des données personnelles, et certaines pratiques sont formellement interdites, comme la segmentation et un traitement différencié entre hommes et femmes dans l'assurance.

Nous sommes convaincus que les algorithmes peuvent apporter de la transparence et de l'objectivité à nos systèmes trop souvent minés par les discriminations. Nous avons, par nos auditions, cherché à cerner les risques posés par les algorithmes pour les discriminations. Nous avons ensuite essayé de dresser quelques recommandations pour que les peurs – parfois exagérées, parfois légitimes – soient surmontées, et que la France avance vers une « IA de confiance ».

A. Les algorithmes sont parfois un remède utile contre les discriminations

Des organisations de nombreux secteurs s'essayent à déployer des algorithmes pour améliorer leurs décisions, les objectiver, et dans certains cas prévenir des discriminations :

Recrutement : les leaders mondiaux de l'intérim mettent en place des algorithmes pour proposer à leurs conseillers de recrutement les profils qui paraissent les plus adaptés à une offre d'emploi. L'algorithme intègre généralement les compétences des candidats et la satisfaction de leurs employeurs passés. Il est capable de se détacher des « profils types » vers lesquels les employeurs se tournent généralement et de réduire les discriminations à l'embauche. Cela fait longtemps que l'on sait que s'appeler Rachid ou Mariam a malheureusement un impact sur les chances d'être recruté, et que le CV anonyme n'y change pas grand chose[4]. On sait aussi que lorsqu'on demande certains prérequis (plusieurs stages dans l'industrie par exemple), on exclut de fait des candidats qui viennent souvent de catégories défavorisées. Des algorithmes qui partent des compétences et non plus seulement des expériences peuvent aider à surmonter ces problèmes.

3 Impliquant respectivement COMPAS, Amazon, Apple Card ou Microsoft.
4 Behagel L., Crepou B., Le Barbauchon T., *Évaluation de l'impact du CV anonyme, PSE* ; voir aussi, Institut Montaigne, *Discriminations religieuse à l'embauche, une réalité*, Octobre 2015.

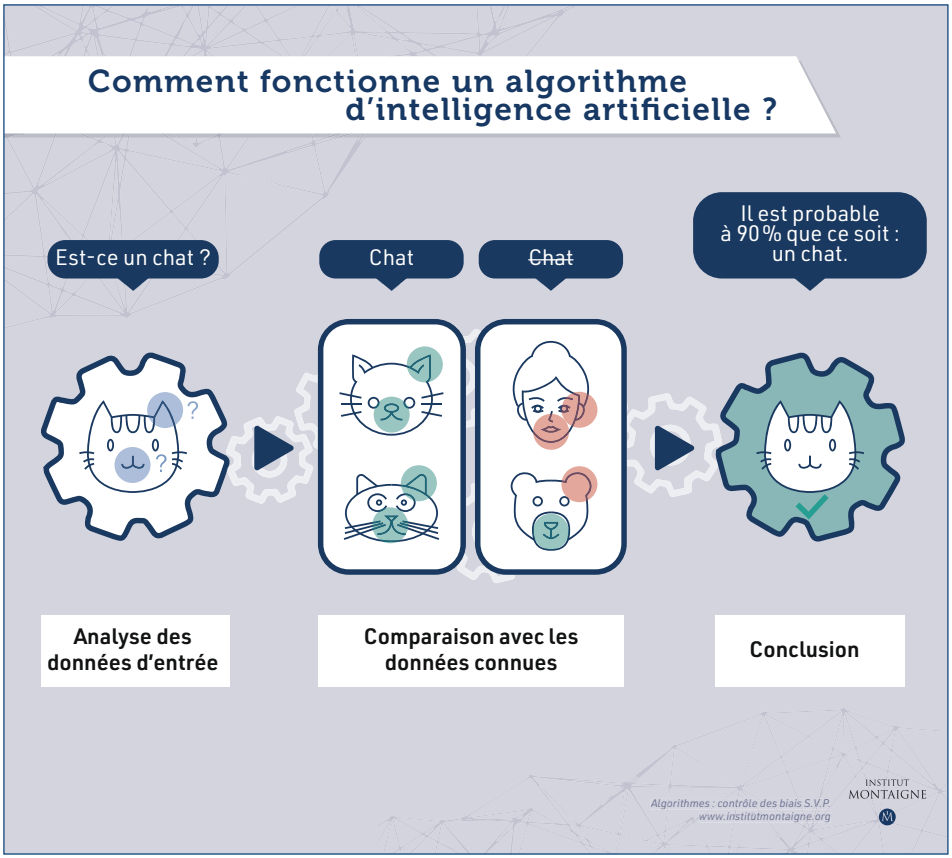
Immobilier : accéder à un logement est parfois une course de longue haleine, particulièrement dans des marchés tendus des grandes villes. La course n'en est que plus longue pour certains. À profil identique, un candidat d'origine maghrébine qui se déclare fonctionnaire aura une réponse du propriétaire dans 15,5 % des cas, contre 42,9 % pour un candidat d'origine française qui envoie le même signal de stabilité professionnelle[5]. Certaines mesures peuvent être efficaces pour réduire l'écart de taux de réponse, comme un rappel à la loi par le Défenseur des Droits, mais leurs effets s'atténuent après 9 mois, et disparaissent après 15 mois[6]. Des algorithmes de recommandation automatique de candidats pour des offres immobilières pourraient permettre de réduire cette discrimination, en évaluant de la même façon deux personnes au profil identique d'un point de vue de critères objectifs.

Justice : la justice américaine est connue pour son passé de discrimination envers des personnes afro-américaines. Certains accusent les algorithmes de reproduire les mêmes discriminations quand d'autres y voient l'opportunité d'en sortir. La libération sous caution est un cas d'école depuis la polémique autour de l'algorithme COMPAS[7]. Actuellement, des cliniciens sont chargés d'évaluer le risque de récidive d'un suspect, pour aider le juge à décider si le suspect peut être libéré sous caution. Ces évaluations cliniques surévaluent les risques qu'un suspect afro-américain récidive sous caution et amènent donc les juges à accorder moins de libérations sous caution aux afro-américains qu'au reste de la population. Elles restent pourtant plus précises que le seul jugement des policiers ou des juges. Des algorithmes tentent d'évaluer ce risque sur la base des antécédents et proposent cette évaluation aux juges. Selon certaines mesures, ces évaluations algorithmiques sont toujours au désavantage des suspects afro-américains. Néanmoins, elles sont plus précises que les études cliniques et permettent aux juges d'être plus constants et plus précis dans leurs décisions. Ces algorithmes ne sont pas parfaits mais leur utilisation est un pas dans la bonne direction.

Ces applications n'existent pas pour l'heure en France. On sait pourtant que, toutes choses égales par ailleurs, les personnes sans emploi ont 1,8 fois plus de risques d'être placées en comparution immédiate que les personnes avec un emploi stable[8]. Certains juges le font de manière raisonnée (pour éviter qu'il soit impossible pour la

5 Bunel M., l'Horty Y., Du Parquet L., Petit P., *Les discriminations dans l'accès au logement à Paris ; une expérience contrôlée*, ffhalshs-01521995f, Mai 2017.
6 Défenseur des droits, *Test de discrimination dans l'accès au logement selon l'origine*, octobre 2019.
7 Goel S., Shroff R., L. Skeem J., Slobogin C., *The Accuracy, Equity, and Jurisprudence of Criminal Risk Assessment SSRN*, Décembre 2018.
8 Gautron V., Retière J-N, *La justice pénale est-elle discriminatoire ? Une étude empirique des pratiques décisionnelles dans cinq tribunaux correctionnels*, Colloque "Discriminations : état de la recherche", Alliance de Recherche sur les Discriminations (ARDIS), Université Paris Est Marne-la-Vallée, France. ffhalshs-01075666, Décembre 2013.

justice de les retrouver et de les juger). Il demeure que des algorithmes permettraient d'expliciter ces choix et d'aider les juges à réduire la part aléatoire de leurs décisions, tout en préservant l'individualisation des peines.



Qu'est-ce que l'IA ?

Un algorithme est une suite d'opérations ou d'instructions permettant d'obtenir un résultat. Une recette de cuisine, par bien des aspects, est un algorithme. Les algorithmes existent sous de nombreuses formes. L'une d'entre elles, l'apprentissage machine (machine learning), a connu un développement significatif ces dernières années, lié en partie à la croissance des quantités de données disponibles.

.../...

Comment fonctionne un algorithme de *machine learning* ?

Un exemple éclairant est celui des algorithmes de reconnaissance de forme. Dans ce type de problème, l'algorithme doit accomplir la tâche suivante : à partir de données d'entrée X, il doit reconnaître automatiquement la catégorie Y associée à chaque objet/individu X, avec un risque d'erreur minimal. La catégorie Y est d'un type donné, spécifié à l'avance.

De très nombreuses applications correspondent à cette formulation, de la biométrie au diagnostic et au pronostic médical assisté en passant par la gestion du risque de crédit. Dans le cas de la vision par ordinateur, par exemple, X correspondra à une image pixélisée et la sortie Y à une « étiquette » associée à l'image indiquant la présence éventuelle d'un objet spécifique dans celle-ci (une tumeur, une fêlure de côte).

La règle de décision qui permet de décider quelle « étiquette » accoler à chaque image est déterminée par un algorithme d'apprentissage. Celui-ci opère sur une base de données déjà étiquetées (i.e. les « données d'apprentissage ») qui associent déjà des images avec la bonne étiquette.

L'importance des grandes bases de données pour l'apprentissage machine

L'objectif de l'algorithme d'apprentissage est de découvrir dans les données d'apprentissage les régularités qui lui permettront de trouver l'étiquette associée à de nouvelles données encore non observées. Face à une image radio qu'il n'a jamais vue, l'algorithme est capable de déterminer avec assez de prédiction la présence d'une fêlure. On cherche à minimiser la probabilité pour l'algorithme de se tromper d'étiquette Y pour une donnée X aléatoire.

Pour être efficace face à de nouvelles données, un algorithme doit avoir été entraîné sur des cas assez variés. D'où l'importance d'une grande quantité de données et des capacités de calcul importantes. Les mégadonnées du web, les immenses bibliothèques d'images, de sons ou de textes « étiquetés » – souvent par des humains – sont donc cruciales.

B. Les algorithmes font peser de nouveaux risques sur les discriminations, et le débat français reste influencé par des exemples américains

a. Faut-il avoir peur des algorithmes ?

Le lancement de l'iPhone en 2007 a initié une période d'optimisme technologique vis à vis de la révolution numérique. Une décennie plus tard, le *techlash* est là. Il ne passe plus un mois sans qu'une nouvelle enquête ne révèle des pratiques invasives dans notre vie privée, ou des algorithmes mis au service d'objectifs douteux. Pour la mathématicienne américaine Cathy O'Neil, les algorithmes et le big data sont même des « armes de destruction mathématique »[9] : des outils qui, sous couvert de formules mathématiques objectives, renforcent les inégalités et les discriminations, amplifient les effets des inégalités.

En 2015, Amazon a mis en place un algorithme pour faciliter le recrutement des talents[10]. Entraîné sur des centaines de milliers de CV reçus par Amazon pendant dix ans, l'algorithme attribuait une note allant de 1 à 5 étoiles. Mais l'algorithme a été rapidement suspendu en raison de son incapacité à sélectionner les meilleurs candidats et de son biais à l'encontre des femmes. L'algorithme attribuait fréquemment des mauvaises notes à des profils de femmes qualifiées et proposait systématiquement des candidats homme sous-qualifiés. Il défavorisait les CV contenant les mots « *women's* », y compris « *women's chess club captain* », et favorisait les CV contenant « *executed* » ou « *captured* », plus fréquents dans les CV masculins. C'est la qualité des données d'entraînement qui a été mise en cause, les hommes constituant l'écrasante majorité des cadres recrutés dans le passé, tandis que de nombreuses femmes étaient surqualifiées pour leur poste. L'algorithme avait ainsi appris à sous-estimer le CV des femmes.

Malgré les bénéfices que des algorithmes pourraient apporter à la lutte contre les discriminations, c'est la méfiance et la peur de leurs effets qui dominent, en tout cas dans le débat public.

L'exclusion ou le traitement défavorable à certains groupes préexiste pourtant aux algorithmes. Le racisme au crédit, le sexisme à l'embauche n'ont pas attendu le troisième millénaire. Les algorithmes ne sont pas plus nés dans les années 2010. Les essais de police prédictive précèdent l'essor de l'apprentissage machine, tandis que la banque et l'assurance ont depuis longtemps intégré des modèles statistiques et des algorithmes dans leurs modes de fonctionnement.

9 O'Neil C., *Weapons of Math Destruction*, Penguin Books, Juin 2017.
10 Reuters, *Amazon scraps secret AI recruiting tool that showed bias against women*, Octobre 2018.

Qu'ils atténuent les discriminations, qu'ils les maintiennent ou qu'ils les aggravent, les algorithmes sont tout autant condamnés. Cela s'explique de plusieurs façons.

Un algorithme qui réduirait une discrimination sans la faire totalement disparaître serait probablement jugé insuffisant au regard de nos idéaux. Et plus nous faisons de progrès vers l'égalité, plus l'inégalité nous paraît intolérable. C'est le paradoxe de Tocqueville :

> *« Chez les peuples démocratiques, les hommes obtiennent aisément une certaine égalité; ils ne sauraient atteindre celle qu'ils désirent. Celle-ci recule chaque jour devant eux, mais sans jamais se dérober à leurs regards, et, en se retirant, elle les attire à sa poursuite. Sans cesse ils croient qu'ils vont la saisir, et elle échappe sans cesse à leurs étreintes. Ils la voient d'assez près pour connaître ses charmes, ils ne l'approchent pas assez pour en jouir, et ils meurent avant d'avoir savouré pleinement ses douceurs. »*
>
> Tocqueville, *De la démocratie en Amérique*

Même à supposer que les algorithmes ne soient que des outils neutres, ils pourraient toujours perpétuer ou consolider une discrimination antérieure. Ils permettraient aussi de mesurer avec précision un biais qui pouvait être connu, mais dont l'ampleur restait cachée. Chaque nouvel algorithme est ainsi l'occasion de révéler un peu plus les formes que prennent les inégalités dans nos sociétés.

Enfin, les algorithmes d'apprentissage ont bien le potentiel d'aggraver la situation en matière de discriminations :

- **L'implémentation à grande échelle** de ces algorithmes intervient simultanément à la numérisation de nos transactions et de nos comportements qui génère des données dans tous les domaines de nos sociétés. Les champs d'application vont bien au-delà de la banque, de l'assurance et des grandes marques de consommateurs.
- Les algorithmes d'apprentissage machine apparaissent comme des **boîtes noires**, dont le fonctionnement reste parfois inscrutable même à leur concepteur. Ils sont plus complexes que leurs prédécesseurs. À l'ère du big data, des dizaines de variables peuvent être fournies à un algorithme, charge à lui de trouver les combinaisons et les pondérations les plus appropriées. Néanmoins il n'existe pas toujours de solution pour rendre intelligible le résultat donné : pourquoi l'algorithme a-t-il abouti à ce résultat-là plutôt qu'à un autre ?
- Lorsque les algorithmes ne font « que » reproduire des inégalités ou des discriminations existantes, ils peuvent **généraliser des discriminations** auparavant circonscrites. L'algorithme de COMPAS est entraîné sur un seul jeu de données judiciaires, celui de Broward County en Floride. Un juge dans l'Oregon qui utilise COMPAS peut

récupérer ainsi les biais historiques du système judiciaire de Broward County. L'utilisation massive d'un tel algorithme aux États-Unis pourrait donc multiplier les effets d'une discrimination limitée à un simple comté de Floride.

b. Reconstruire le débat français sur l'impact des algorithmes

Dans les médias comme dans la recherche académique, l'omniprésence des exemples anglo-saxons, et particulièrement des cas américains, est frappante. Le déploiement des algorithmes aux États-Unis est bien plus avancé qu'en France, conséquence d'une réglementation moins stricte et d'une industrie numérique plus développée. Certaines des applications les plus polémiques aux États Unis n'auraient en effet pas pu voir le jour en France. L'article 47 de la loi informatique et libertés de 1978 modifiée en 2018 stipule ainsi « aucune décision de justice impliquant une appréciation sur le comportement d'une personne ne peut avoir pour fondement un traitement automatisé de données à caractère personnel destiné à évaluer certains aspects de la personnalité de cette personne ».

Ce décalage entre la France (et l'Europe continentale plus largement) avec les États-Unis est à la fois une bonne et une mauvaise nouvelle. Nous avons l'opportunité d'imaginer un système d'encadrement qui évite les dérives de certains systèmes déployés outre-Atlantique, et qui corresponde à notre système de valeurs. Mais nous accusons également un retard dans la maîtrise de ces outils. Le risque est grand d'entraver la transformation numérique de l'économie française, faute de bien comprendre la source des biais, de bien évaluer le risque qu'ils posent, et d'identifier les moyens efficaces de les combattre.

Si les organisations installées en France sont bien moins nombreuses à déployer des algorithmes d'apprentissage machine qu'aux États-Unis, les Français utilisent quotidiennement des services numériques des GAFA (Google, Apple, Facebook, Amazon), et leurs fonctionnalités améliorées par des algorithmes.

Selon l'IFOP, 80 % des Français jugent la présence des algorithmes déjà massive dans leur vie mais un peu plus de la moitié avoue ne pas savoir précisément ce que sont les algorithmes[11]. En dehors de quelques domaines (banque et assurance, ciblage publicitaire, services purement numériques des GAFA), au-delà des effets d'affichage, les décisions assistées ou prises par des algorithmes restent peu nombreuses en France.

11 CNIL, *Éthique et numérique : les algorithmes en débat,* Janvier 2017.

II

BIAIS DES ALGORITHMES : UN PROBLÈME ANCIEN ET COMPLEXE

A. Les biais préexistent aux algorithmes et résident principalement dans les données qu'ils utilisent

Les algorithmes de *machine learning* apprennent et décident à partir de données produites par les humains et converties dans des formats numériques. Il peut s'agir de données relationnelles provenant de réseaux sociaux, des comportements d'achats utilisés à des fins marketing, des préférences musicales extraites des plateformes de streaming, des vidéos et photos postées sur internet, des SMS échangés, des historiques de recherches sur Google, des décisions en matière de recrutement ou d'octroi de crédit etc.

Ces données sont un miroir digital des comportements humains. Elles reflètent strictement nos habitudes et donc nos biais, lorsque nous en avons. Avec le développement des capteurs et de l'internet des objets, de plus en plus de données fines sont collectées (maison et enceintes connectées, applications mobiles de santé, de vacances, de loisirs, etc). Les algorithmes qui apprendront sur ces données et leurs biais auront une capacité à standardiser et amplifier des discriminations si nous ne sommes pas vigilants.

Discrimination et biais, de quoi parle-t-on ?

Dans le langage technique, le biais d'un algorithme est l'écart moyen entre sa prédiction et la valeur que l'on cherchait à prédire. Concrètement, cela peut être l'écart entre le nombre d'images radios étiquetées par l'algorithme comme « comportant une tumeur » et le nombre d'images radios comportant réellement une tumeur. Un fort biais signifie que l'algorithme manque de relations pertinentes entre données d'entrée et de sortie pour effectuer des prédictions correctes.

.../...

Ces **« biais techniques »** sont pris en compte par les ingénieurs, car ils viennent directement diminuer la performance de l'algorithme. Ils relèvent de la discussion publique autour des « biais des algorithmes » lorsqu'ils désavantagent un groupe spécifique sur des critères illicites. Ce qui n'est pas systématique : un algorithme qui reconnaît moins souvent les chênes que les bouleaux sur des images est biaisé techniquement, sans préjudice en matière de discrimination.

Les biais des algorithmes vont au-delà des biais techniques. L'algorithme peut être très performant techniquement tout en étant « biaisé » d'un point de vue social. Ces biais sont condamnés comme des discriminations car ils sélectionnent et arbitrent souvent en défaveur de populations déjà défavorisées. Ces **« biais de société »** sont en fait la reproduction via l'algorithme de biais déjà présents au sein de la société.

La distinction entre biais technique et biais de société est importante car **ceux-ci n'attirent pas la même attention de la part des développeurs.** Les biais techniques réduisent la performance de l'algorithme, entravent la réalisation de son objectif. Réduire les biais techniques a un coût, mais également souvent un bénéfice clair pour le développeur. *A contrario*, suivre les biais de société permet à l'algorithme d'être plus performant. En matière de publicité ou d'offre d'emplois, coller aux stéréotypes peut permettre de maximiser le nombre de clics sur les annonces.

Les biais des algorithmes peuvent aussi être des décisions conscientes visant à soutenir une stratégie d'entreprise. Google a été condamné à une amende de 2,4 milliards d'euros pour avoir favorisé ses propres produits dans les résultats de recherche de Google Shopping au détriment de ses concurrents. L'algorithme était donc volontairement biaisé.

Dans le cadre de ce rapport le terme de biais algorithmiques désigne aussi bien les biais techniques, bien connus des statisticiens, que les biais de société moins bien définis à leurs yeux.

Les biais techniques trouvent souvent leur source dans les données

Inspiré de l'article de Patrice Bertail, David Bounie, Stéphan Clémençon et Patrick Waelbroeck, *Algorithmes : biais, discrimination et équité*, 2019.

Il existe plusieurs types de biais techniques, à savoir de biais qui contreviennent à la performance de l'algorithme. Leur source peut être un défaut dans le développement mais, bien plus souvent, il s'agit d'un défaut dans la qualité ou dans la représentativité des données.

Certaines données sont compliquées, voire impossible, à collecter. Si on tente parfois d'en trouver une approximation, on peut aussi se contenter d'autres variables. On ouvre alors la voie à un possible **biais de variable omise.** Certaines compétences telles que le leadership ou l'intelligence émotionnelle sont par exemple difficiles à mesurer, et peuvent être négativement corrélées aux résultats scolaires. Un algorithme de sélection prenant uniquement ces derniers en compte aurait un biais qui l'amènerait à ne pas repérer certaines personnes à haut potentiel, alors même qu'il aurait dû les repérer pour atteindre ses objectifs.

Au-delà du choix des variables, des **biais de sélection** peuvent apparaître lorsque l'échantillon d'apprentissage n'est pas représentatif de l'environnement d'application. Par exemple en médecine, lorsqu'on entraîne l'algorithme sur des données de populations à majorité blanche. C'est le cas aussi bien connu de la reconnaissance faciale, qui fonctionne mieux sur les personnes blanches[12] que sur les personnes de couleur du fait des bases de données utilisées pour l'entraînement. Une évaluation[13] réalisée par le National Institute for Science and Technology américain sur 189 algorithmes de 99 développeurs a ainsi établi que les taux de faux positifs (de personnes reconnues de façon erronées par leur téléphone par exemple) étaient 10 à 100 fois plus élevés pour les personnes d'Afrique et d'Asie de l'Est (sauf pour ces derniers en cas d'algorithme développé en Chine).

La source de biais techniques la plus commune reste certainement la qualité des données d'apprentissage, qui peut induire des **biais dans les bases de données.** Si les étiquettes des images d'entraînement sont erronées (de nombreuses images étiquetées « mouton » alors que l'image comporte en fait un bouc), le résultat final sera inévitablement biaisé, au sens technique qu'utilisent les statisticiens. C'est ce biais qui est à l'œuvre quand on entraîne l'algorithme sur des données dites « biaisées » (par exemple des avis sexistes sur la performance des employés). L'algorithme peut aussi sembler fonctionner alors qu'il n'a trouvé que des indices dans les données d'apprentissage : au lieu de repérer des radios qui présentent des tumeurs, un algorithme pourrait être entraîné à repérer des radios stockées par le *data scientist* dans le dossier C://Images/Photos_tumeurs. Ces indices sont factices, car utilisés dans

12 Buolamwini, J. et Gebru, T. *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification* in Proceedings of the 1st Conference on Fairness, Accountability and Transparency, 2018.
13 NIST, *Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects*, 2019.

un autre environnement. L'algorithme sera alors incapable de repérer les images présentant des tumeurs.

Ces biais sont dits techniques car ils sont pris en compte par les statisticiens. Mais ils ne sont pas sans conséquences sur la société. L'usage par Kodak d'un standard technique calibré uniquement pour les personnes blanches, la carte *Shirley*, a conduit à l'impossibilité pour les personnes noires d'être représentées correctement *via* l'image - d'être visibles, au sens propre comme au figuré.

Les cartes *Shirley* illustrent qu'en matière de biais techniques, le problème se situe souvent du côté des données, notamment de leur qualité ou de leur représentativité. Si les développeurs de Kodak avaient constitué une base de cartes *Shirley* représentative de la diversité des couleurs de peau, le système n'aurait pas discriminé les personnes non caucasiennes. Cet exemple nous rappelle également que les standards techniques ne sont pas bénins, un rappel utile quand on voit aujourd'hui la bataille menée par les entreprises chinoises pour définir les standards techniques pour les algorithmes de reconnaissance faciale.

Les cartes Shirley de Kodak, des données à la sources de biais techniques

Les biais dans les technologies émergentes existaient bien avant l'intelligence artificielle. Un exemple notable est celui des films de couleur de Kodak qui ont standardisé une dégradation de la qualité des photographies pour les personnes dont la couleur de peau n'était pas blanche.

Au début des années 1940, Kodak était l'un des seuls fournisseurs de photographies de couleur. Pour calibrer les couleurs de peau, les ombres et les lumières pendant le développement et l'impression des photos, l'entreprise créa une carte *Shirley*, portrait d'une femme à la peau blanche et aux cheveux noirs. Cette carte servait de modèle et était inspirée d'une employée de Kodak. Les années passant, les modèles *Shirley* changèrent mais les images modèles successives se conformaient toutes aux standards de beauté en vigueur. Ces cartes étaient distribuées dans tous les laboratoires du monde avec des kits comportant des négatifs originaux, permettant de calibrer les appareils. Les intrants chimiques étaient optimisés pour magnifier les couleurs de tonalités légères, et capturaient ainsi mal les tonalités sombres. Celles absorbant plus de lumière nécessitaient un développement différent pour offrir un rendu optimal. Ainsi, les images de peaux noires étaient peu visibles, seuls les dents et les yeux apparaissant avec un fort contraste.

.../...

Les mouvements américains contre les discriminations n'ont jamais questionné Kodak. À l'époque, l'hypothèse dominante était que le problème était d'ordre scientifique, qu'il n'existait pas de solutions techniques pour mieux représenter les peaux noires. Pourtant, dans les années 1970, Jean-Luc Godard refusa de réaliser un film au Mozambique avec des pellicules Kodak, les accusant d'être racistes.

Une solution partielle vint du monde de l'entreprise. Deux clients importants de Kodak étaient des fabricants de confiseries et de meubles. Ils se sont ainsi plaints que les publicités pour leurs chocolats et meubles sombres n'avaient pas de bon rendus avec les pellicules Kodak. L'entreprise développa alors de nouveaux films et fabriqua de nouvelles *Shirley*, 3 femmes de types caucasien, afro-américain et asiatique.

Les personnes afro-américaines ont ainsi vu leur identité visuelle limitée pendant 30 ans, tandis que les personnes caucasiennes devenaient des standards de beauté à travers l'usage d'une technologie novatrice mais hautement biaisée.

Plusieurs biais de nos sociétés peuvent être encodés dans les données utilisées par les algorithmes.

La psychologie distingue deux types de biais venant distordre nos décisions : les **biais émotionnels** (ou affectifs), et les **biais cognitifs** (notamment des stéréotypes). Tandis que les biais émotionnels nous amènent à refuser de croire en des réalités désagréables, les stéréotypes reviennent à traiter une personne selon le groupe auquel elle appartient (et les traits que l'on associe avec ce groupe), plutôt que sur ses caractéristiques individuelles.

Ces deux types de biais ont imprégné nos sociétés avant même l'arrivée des algorithmes et de l'intelligence artificielle. Pourtant, au XXIe siècle encore, rares sont les biais qui sont ouvertement reconnus. Seulement 10 % de la population[14] reconnaît explicitement avoir des stéréotypes, et les biais inconscients sont fortement répandus, affectant notre jugement et pouvant parfois entraîner des discriminations.

14 Fiske S. and Taylor S. *Social Cognition: From Brains to Culture, 2nd edn*, Sage: London,1984.
15 Hall, E. T., Beyond Culture, Anchor, Décembre 1976.

Les recherches de l'anthropologue Edward T. Hall ont mis en évidence les liens, involontaires, entre la façon dont s'échangent information et stéréotypes[15]. Dans les sociétés qu'il désigne comme « fortement contextuelles », l'échange d'information est assuré via des éléments implicites comme le langage corporel ou le comportement social. Ce type d'information est facilement appréhendé par un membre de la communauté mais est plus complexe d'accès pour un étranger, instaurant ainsi des barrières à l'intégration. Le Japon et la France sont des exemples de sociétés fortement contextuelles. Dans ces sociétés, les traditions sont présentes au quotidien, et le jugement passe par l'application de normes. Ces environnements sont exposés à la force des stéréotypes, qui sont une forme de lecture du contexte.

Les algorithmes peuvent répandre dans la société ces types de biais, en se contentant de répliquer les décisions humaines déjà biaisées, ou parce que les développeurs sont eux même l'objet de ces biais.

Les développeurs peuvent ainsi parfois laisser des biais cognitifs troubler la façon dont ils conçoivent ou interprètent leurs modèles. Cela peut les amener à orienter des modèles selon leur vision du monde. Michal Kosinski et Yilun Wang, deux chercheurs de Stanford, ont ainsi proposé en 2018 de détecter l'orientation sexuelle d'un individu selon sa morphologie faciale[16]. La physiognomonie — l'idée que l'apparence physique donne un aperçu du caractère — a une histoire longue, et troublée. Au-delà, de nombreux chercheurs ont souligné que l'algorithme détectait bien plus des styles (de barbe, de maquillage, de port de lunettes) qu'une quelconque « physionomie homosexuelle »[17]. Or ces éléments de styles sont eux-même en partie des signaux que nous envoyons pour nous inclure dans certains groupes. Loin de détecter l'orientation sexuelle à partir des traits du visage, et de renforcer ainsi une certaine théorie sur l'origine de l'orientation sexuelle, les travaux de Kosinski et Wang sont sans doute plus révélateurs de leurs présupposés que d'une quelconque réalité.

Les algorithmes ont également été à plusieurs reprises accusés de répandre des stéréotypes, notamment à l'encontre de l'égalité entre les hommes et les femmes. Par exemple, les femmes ont tendance à répondre seulement à des offres d'emploi qu'elles pensent avoir une très forte probabilité de décrocher, et orientent ainsi les offres que les algorithmes leur présentent. C'est également ce biais de stéréotype qui se retrouve dans les co-occurrences de mots : le mot « femme » est ainsi associé dans les corpus à « styliste » ou « coiffeuse », tandis que le mot « homme » est associé

16 Kosinski M., Wang Y. *Deep Neural Networks Are More Accurate Than Humans at Detecting Sexual Orientation From Facial Images* in Journal of Personality and Social Psychology, Février 2018.
17 Aguera y Arcas,B. *Do algorithms reveal orientations or just expose our stereotypes?* Medium Equality, Janvier 2018.

à « capitaine », « chef », ou « financier »[18]. Ces stéréotypes viennent nourrir et biaiser les algorithmes, notamment ceux qui recommandent des offres d'emploi.

Au-delà des biais cognitifs et affectifs, les biais économiques forment un autre type de biais présent dans la société. Un algorithme peut contenir un biais volontairement ou involontairement pour des raisons de stratégie commerciale. Un algorithme qui optimise simplement le rapport coût-efficacité de la diffusion d'offres d'emploi affiche moins d'annonces destinées aux femmes jeunes qu'aux hommes jeunes[19]. En effet, on observe que les espaces publicitaires à destination des jeunes femmes sont plus chers que ces mêmes espaces à destination des jeunes hommes. Ainsi, il sera moins coûteux pour l'algorithme de préférer les hommes pour ces annonces d'emploi. La stratégie commerciale visant à recruter en minimisant les coûts de recrutement conduit à une discrimination à l'encontre des femmes.

On pourrait considérer que les biais de société ne sont pas l'affaire des algorithmes qui n'en sont pas à l'origine. Néanmoins, un algorithme peut multiplier ce biais, ce qui n'est pas sans risque. En effet, les biais de société peuvent avoir des conséquences en profondeur sur les individus.

Les biais et discriminations ont par exemple un effet auto-réalisateur, les groupes discriminés étant amenés à se conformer aux stéréotypes en vigueur. Une enquête dans les supermarchés français[20] a ainsi montré qu'être exposé à des managers ayant des biais importants affecte négativement les performances au travail du personnel appartenant à des minorités[21]. Lorsque les horaires de travail de personnes issues de minorités coïncidaient avec ceux des managers ayant des stéréotypes, les caissiers validaient les articles moins rapidement, étaient plus souvent absents et quittaient le travail plus tôt, conduisant à des salaires plus réduits dans un environnement où la norme est le salaire horaire. Les managers exprimant des stéréotypes ne traitent cependant pas différemment les employés issus de minorités, mais semblent simplement passer moins de temps avec eux. Ainsi, les stéréotypes n'ont même pas besoin de guider l'action explicite des managers pour avoir un impact sur les écarts salariaux et sur les habitudes de recrutement.

18 Bolukbasi T., Chang K-W., Zou J., Saligrama V., Kalai, A. *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings*, 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, Juillet 2016.

19 Lambrecht, A. and Tucker, C. *Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads*. Mars 2018.

20 Glover D., Pallais A., Pariente W., *Discrimination as a Self-Fulfilling Prophecy: Evidence from French Grocery Stores*, The Quarterly Journal of Economics, Volume 132, Issue 3, Août 2017.

21 Personnes ayant un nom à consonance d'Afrique du Nord ou d'Afrique sub-saharienne.

Face à ce type de risque, il n'est pas possible de limiter l'analyse des biais des algorithmes aux biais techniques : la réflexion doit intégrer pleinement le risque de transmission de biais de sociétés qui sont de nature multiple.

B. L'algorithme équitable a de multiples définitions contradictoires ; est-ce aux organisations de choisir laquelle appliquer ?

Que l'on parle d'un biais technique ou d'un biais de société, la dimension culturelle des biais algorithmiques ne peut pas être ignorée. Au delà de son efficacité, la recherche académique définit usuellement trois qualités dont peuvent être dotés les algorithmes : l'équité, bien sûr, mais également la neutralité et la loyauté[22]. Chacune peut constituer une garantie partielle contre les biais algorithmiques. Ces trois qualités font toutefois l'objet de définitions multiples selon le contexte dans lequel l'algorithme va être utilisé, mais également selon l'environnement culturel (américain ou européen par exemple). L'algorithme sans biais n'existe pas dans l'absolu. Des choix doivent être faits - et sont quotidiennement faits - par leurs concepteurs.

a. Pourquoi faut-il parler des équités des algorithmes au pluriel ?

L'équité est littéralement la « qualité consistant à attribuer à chacun ce qui lui est dû par référence aux principes de la justice naturelle »[23]. Mais la « justice naturelle » n'est ni écrite, ni consensuelle. Chaque théorie de ce qui est juste entraîne sa propre définition de l'équité. La définition d'une décision équitable dépend ainsi de la culture, du secteur d'application, des objectifs que l'on se donne. Difficulté additionnelle pour des algorithmes entraînés sur des données historiques, cette définition varie au cours du temps dans chaque culture. À titre d'exemple, on distingue deux grandes catégories dans les définitions de l'équité : l'équité de groupe et l'équité individuelle.

22 Un algorithme est neutre lorsqu'il représente fidèlement la réalité. L'algorithme est loyal lorsqu'il répond fidèlement aux attentes de son concepteur. L'algorithme équitable est quand à lui celui qui est juste dans ses décisions.
23 Définition « équité », Larousse.

Les différentes formes d'équité

L'équité peut être conçue de deux façons différentes :
- l'équité individuelle, qui assure que des individus aux profils proches seront traités de la même façon ;
- l'équité de groupe, qui assure que le processus de décision ne défavorise pas arbitrairement un certain groupe (celle généralement utilisée dans les tests d'algorithmes).

La définition de l'équité que l'on souhaite appliquer dépend de l'hypothèse que l'on fait sur notre capacité à mesurer le monde. Les décisions d'admission universitaire en sont un bon exemple. Avec ou sans algorithme, l'équité d'une décision demande de mettre en regard :
- la décision en elle-même (d'admettre ou non un lycéen dans une formation) ;
- les critères jugés pertinents pour la décision (la motivation, la curiosité, la force de travail potentiellement).

L'équité peut être rompue entre individus lorsque la décision n'est pas prise sur les mêmes critères pour tout le monde. Mais elle peut également être rompue lorsque la façon de mesurer les critères est au désavantage de certains groupes.

Les lettres de motivation font par exemple office d'une approximation de la motivation des candidats. Cette approximation n'est pas neutre quand elle favorise ceux qui peuvent solliciter de multiples conseils sur comment « se vendre », et qui sont plus souvent enfants de cadres et de professeurs que d'employés ou d'ouvriers.

S'il n'est pas possible de mesurer directement les critères pertinents, et que cette mesure n'est pas aussi fiable pour certains groupes que pour d'autres, alors on pourra vouloir corriger le processus par des contraintes d'équité de groupe.

Les deux types d'équité sont intrinsèquement incompatibles. En effet, il n'est mathématiquement pas possible de traiter tous les individus présentant des caractéristiques identiques de la même manière, tout en assurant un traitement équitable entre des groupes différents. Face à un résultat déséquilibré en faveur d'un groupe, la façon la plus simple de rétablir l'équité de groupe sera d'être plus strict pour les individus du groupe favorisé.

Le recours à l'équité individuelle et à l'équité de groupe varie significativement selon les cultures, et entre les champs d'applications. Il est communément accepté aux États-Unis de pratiquer de la discrimination positive pour les admissions aux universités en fonction des origines ethniques. Un traitement équitable est alors défini comme une équité de groupe. À l'inverse, en France, le recours aux concours, aux épreuves standardisées, se réfère à l'idéal d'exigences identiques pour tous les élèves, sans tenir compte de leurs origines. Les deux approches sont le reflet d'équilibres politiques et historiques. Elles font figure de sagesse conventionnelle dans leur pays.

Ces choix ne sont par ailleurs pas intemporels et font régulièrement l'objet de contestations. Un groupe d'étudiants d'origine asiatique a ainsi porté plainte en 2019 contre l'université d'Harvard lui reprochant de favoriser dans ses processus de recrutement les étudiants d'origine afro-américaine et hispanique[24] au nom de la diversité. Le juge a finalement donné raison à l'université, affirmant que pour atteindre une réelle diversité, il était encore trop tôt pour ne pas prendre en compte l'origine ethnique comme un des critères de recrutement. Comme le rappelle Raja Chatila, directeur de l'Institut des systèmes intelligent et de la robotique, « penser l'éthique des systèmes autonomes renvoie à notre façon de voir le monde »[25].

Il est illusoire de penser définir un type d'équité qui s'appliquerait en toute circonstance et sans prise en compte du contexte. L'appel au législateur restera vain, tant la question est ardue, et politiquement brûlante. Tout au plus peut-on l'espérer dans certains domaines précis, comme l'éducation ou la santé. Il sera *in fine* de la responsabilité des développeurs et des managers d'une entreprise ou de l'administration utilisant un algorithme de décider le type d'équité à mettre en œuvre, dans le cadre de la loi.

Face à cette diversité de définitions de l'équité, il existe communément trois approches alternatives pour l'algorithme : l'anti-classification, la parité de classification et la logique de calibration. Ces trois logiques privilégient tantôt une équité individuelle, tantôt une équité de groupe.

24 Hartocollis,A. *Harvard Does Not Discriminate Against Asian-Americans in Admissions, Judge Rules*,The New-York Times, 1er Octobre 2019, mise à jour le 5 Novembre 2019.
25 Garreau M, et Gateaud P. Entretien avec *Raja Chatila : Penser l'éthique des systèmes autonomes renvoie à notre façon de voir le monde*, L'Usine Nouvelle, Décembre 2018.

Les définitions formelles de l'équité pour un algorithme

Exemple d'un algorithme d'admission universitaire et de son équité au regard d'une variable protégée, le genre.

Anti-classification : l'algorithme ignore la variable de genre, et n'inclut que des variables non liées au genre (résultats écrits au baccalauréat, contrôle continu, résultats oraux au baccalauréat, en supposant bien sûr que ces variables ne sont pas corrélées au genre). Cette méthode poursuit l'équité individuelle entre tous les candidats, quel que soit leur genre. Il s'agit de l'équité par ignorance des variables sensibles.

Parité de classification : l'algorithme est contraint de telle façon que la proportion de faux positifs (personnes admises et dont on réalise qu'elles n'ont pas le niveau requis) est identique pour les lycéens et les lycéennes. L'algorithme sera ajusté de telle façon qu'il fera autant d'erreurs sur chaque groupe. C'est une autre forme d'équité de groupe.

Logique de calibration : l'algorithme est contraint de telle façon que pour des lycéens et lycéennes aux résultats scolaires et aptitudes similaires, les résultats d'admissions sont complètement indépendants des variables protégées. Contrairement à l'anti-classification, l'algorithme est programmé pour respecter cette indépendance entre variables protégées et évaluation des performances des élèves, même si cela se fait au détriment de certains candidats.

b. Équité des algorithmes, un idéal difficile à atteindre

Une fois qu'une définition de l'équité est retenue et appliquée, affirmer qu'un algorithme est équitable reste complexe. En effet, l'objectivité des critères utilisés par l'algorithme dans son fonctionnement est difficile à garantir.

Un algorithme cherche à prendre (ou à aider) des décisions sur la base de critères pertinents. Mais ces critères ne sont pas toujours directement mesurables : on ne peut directement mesurer l'intelligence, le potentiel ou la curiosité d'un lycéen ou d'une lycéenne. On se repose donc sur des caractéristiques mesurables, qui sont forcément des approximations des critères que l'on souhaite réellement prendre en compte. Par exemple ses notes au baccalauréat ou sa lettre de motivation. Ces approximations comportent en elles-mêmes, certains biais et stéréotypes.

Cette problématique est particulièrement visible dans le domaine de la police et du contrôle préventif d'identité. Comment décider quelles personnes contrôler dans un aéroport ou à l'entrée d'un stade ? À défaut de pouvoir mesurer de manière objective le niveau de dangerosité d'un passant, les forces de sécurité vont généralement utiliser des critères nécessairement imparfaits. Certains pourraient utiliser l'intuition, d'autres se focaliser sur les jeunes hommes seuls, ou sur une impression de dangerosité. Probablement par peur de biais, d'autres préfèreront le hasard.

C'est notamment le cas pour la douane de certains aéroports au Mexique où les personnes contrôlées sont choisies pour partie au hasard.

Au-delà de la définition de l'équité, il est donc bien nécessaire de questionner les approximations réalisées dans le choix des critères utilisés par l'algorithme.

c. Un algorithme à la fois équitable et performant, un équilibre difficile

Un algorithme de prise de décision ou d'aide à la décision optimise un résultat en fonction de données d'entrée et d'un objectif. La définition de cet objectif est le cœur de l'algorithme.

Cette optimisation est rarement faite sans imposer des contraintes à l'algorithme, des « barrières » à ne pas dépasser. Dans le cas de Facebook par exemple, l'algorithme de recommandation publicitaire a pour objectif principal de maximiser le nombre de clics sur les publicités affichées à l'utilisateur. On peut néanmoins penser que d'autres contraintes pourraient être prises en compte : un maintien d'une forme de diversité dans les publicités affichées, une interdiction de certains types de contenus (par exemple, les publicités de sites de rencontres pour les enfants ou jeunes adolescents), une limitation des publicités politiques, etc.

Il existe alors un conflit entre l'objectif principal et les contraintes secondaires. C'est notamment le cas lorsque l'on impose à l'algorithme des contraintes en matière d'équité : celles-ci se traduisent toutes choses égales par ailleurs par une moindre performance (au regard de l'objectif initial). L'équité d'un algorithme se fait souvent au détriment de sa performance.

Ainsi, refuser de proposer des publicités de sites de rencontres à de jeunes adolescents sur Facebook revient à accepter de diminuer le nombre de clics et donc à réduire les revenus pour Facebook du fait d'une performance dégradée de l'algorithme.

Le combat pour un algorithme de Youtube plus éthique… et moins performant

L'ONG AlgoTransparency s'est donné pour mission d'informer le public sur le fonctionnement des algorithmes influençant notre accès à l'information. Elle a notamment activement travaillé sur les algorithmes de recommandations de vidéos de Youtube et a questionné publiquement les impacts négatifs des objectifs assignés à cet algorithme.

L'algorithme publicitaire de Youtube, lorsqu'il recommande des contenus clivants et extrêmes, suit la logique d'optimisation qui lui a été donnée lors de sa conception : maximiser le nombre de recommandations qui seront suivies d'un clic de la part de l'utilisateur, et d'un visionnage de la vidéo, qui génèrera des contenus publicitaires.

Il s'avère que les contenus clivants ont une plus forte propension à attirer l'œil et l'attention des humains en général, et notamment des utilisateurs de Youtube. Le biais de l'algorithme de Youtube envers certains contenus est donc tout à fait volontaire.

Réduire ce biais en limitant la diffusion des contenus clivants implique donc, si les objectifs de l'algorithme restent les mêmes, une réduction de sa performance.

Si l'équité n'est pas prise en compte en amont, alors toute correction aura un coût. Elle exigera un compromis entre la performance d'un algorithme au regard de ses objectifs initiaux et le respect de critères additionnels comme l'éthique ou l'équité.

Cependant, ce compromis n'est pas figé, et ses critères n'ont rien d'automatique. Un algorithme de recommandation de publicité ciblée pour l'achat de rasoirs pour homme, optimisé pour cibler uniquement des hommes, aurait mécaniquement un taux de faux positif plus élevé pour les femmes que pour les hommes. L'algorithme de recommandation repose dans ce cas sur un critère de reconnaissance du genre. Celui-ci va être calibré pour reconnaître toutes les femmes et éviter ainsi de leur montrer la publicité, ce qui serait une dépense « inutile ». Il sera par contre moins strict dans la reconnaissance des hommes. Ne pas reconnaître le genre d'un homme implique de ne pas lui montrer la publicité, ce qui est une occasion manquée, mais a des conséquences moins graves en cas de contrainte de budget.

Les faux positifs pour les hommes (l'algorithme pense que c'est un homme alors que c'est une femme) seront donc évités le plus possible, tandis que les faux positifs pour les femmes (l'algorithme pense que c'est une femme alors que c'est un homme) seront plus facilement tolérés.

La plus simple manière d'obtenir l'égalité entre les hommes et les femmes serait de dégrader la précision de l'algorithme chez les hommes. Mais cela serait contre productif avec son objectif : maximiser l'efficacité des dépenses publicitaires pour vendre des rasoirs pour hommes.

Forcer *ex post* un logiciel à respecter des critères d'équité conduit nécessairement à une réduction de sa performance au regard de son critère principal, car on ajoute des contraintes dans sa fonction d'optimisation. Si cela peut paraître nécessaire dans le domaine de la justice, la réponse est moins évidente pour un algorithme détectant les cancers à partir de radiographies : doit-on privilégier son équité, au détriment de sa performance de prédiction des cancers sur certains patients ?

d. Loyauté et neutralité, deux approches complémentaires à l'équité, mais imparfaites

Loyauté et neutralité de l'algorithme sont souvent brandis comme des concepts permettant d'obtenir des algorithmes éthiques. Tout comme l'équité, ces concepts souffrent de limites et ne sauraient être suffisants pour permettre d'atteindre des algorithmes sans biais.

La neutralité d'un algorithme consiste à assurer que celui-ci donne une représentation fidèle de la réalité, identique à celle-ci : les décisions de l'algorithme doivent correspondre à la réalité. S'il s'agit de sélectionner des CV de candidats, le système devrait alors proposer la même proportion de candidats hommes et femmes que celle existant aujourd'hui dans la base de données de candidats.

Un algorithme neutre pourrait donc comporter, par construction, tous les biais présents dans notre société. Certains défendront que c'est après tout normal, car ce n'est pas le rôle de l'algorithme de corriger lui même les problèmes de la société. D'autres affirmeront au contraire que compte tenu de sa capacité à répliquer à l'échelle et diffuser un biais préexistant, l'algorithme a un rôle, voire une responsabilité en la matière.

Le concept de loyauté se réfère, lui, à l'utilisateur. Il s'agit pour un algorithme de respecter non pas la réalité mais les attentes des utilisateurs et consommateurs de

l'algorithme (différentes de celles du concepteur). Cette approche pose immédiatement la question de l'identité des utilisateurs et de la définition de leurs attentes. L'utilisateur d'un logiciel de reconnaissance faciale est-il le policier en faisant usage pour retrouver un criminel, ou le passant filmé et identifié dans une gare en prenant son train ? Quelles sont ses attentes et comment le développeur d'un logiciel peut-il les anticiper ?

La loyauté d'un algorithme dépend fondamentalement des attentes de personnes aux cultures très distinctes. En France, on pourrait attendre d'un algorithme proposant des offres d'emploi qu'il s'assure d'une promotion équilibrée auprès des hommes et des femmes. Un autre pays pourrait considérer que cette attitude empêcherait l'algorithme de choisir uniquement sur des critères professionnels.

e. Cas d'usage réel d'un algorithme sans biais pour le recrutement

Une grande entreprise du domaine de l'intérim utilise un algorithme pour recommander des CV à ses clients cherchant à recruter du personnel. L'entreprise est sensible à la problématique des biais algorithmiques. À ce titre, elle teste son algorithme pour s'assurer que celui-ci ne discrimine pas suivant quelques critères élémentaires, notamment l'âge, le sexe[26].

L'algorithme mesure donc, pour chaque profession, le nombre de CV d'hommes et de femmes recommandés. À partir de là, si un déséquilibre est constaté, trois solutions sont possibles :

- redresser le résultat pour atteindre l'équilibre entre hommes et femmes constaté historiquement pour ce type d'emploi dans l'entreprise ;
- redresser le résultat pour atteindre l'équilibre entre hommes et femmes constaté dans cette profession à un échelon plus grand que l'entreprise (département, région, pays…) ;
- redresser le résultat pour atteindre un équilibre entre hommes et femmes fixé de manière arbitraire (40-60, 50-50, 60-40, etc.).

La décision du type de réponses à adopter relève de l'entreprise et de ses dirigeants. Elle doit être explicable, pas nécessairement au grand public, mais à minima à une autorité tierce. Il n'existe pas nécessairement de « bonne réponse », puisque chacune des approches pourrait sembler équitable.

26 Aux États-Unis où la collecte de données sensibles est autorisée, de nombreux autres paramètres sont testés : religion, opinion politique, genre, etc.

S'il y a un redressement des résultats, cela se fait nécessairement au détriment du score d'appariement des candidats proposés et cela est accepté ouvertement par l'entreprise.

En l'absence d'obligation faite aux entreprises, il est évident que la concurrence incitera les entreprises à limiter le redressements des résultats. Cependant, l'opinion publique étant de plus en plus sensible à ces sujets, il est également probable que le risque réputationnel les incitera à redresser des résultats manifestement biaisés.

III

DE NOMBREUSES LOIS EXISTENT DÉJÀ CONTRE LES DISCRIMINATIONS, PRIVILÉGIONS LEUR APPLICATION PLUTÔT QUE D'ENVISAGER DE NOUVEAUX TEXTES SPÉCIFIQUES AUX ALGORITHMES

'Don't use a sledgehammer to kill a fly'

Cette maxime anglaise - « il est inutile de se munir d'un marteau pour tuer une mouche » - est à garder à l'esprit lorsque l'on traite du sujet des biais algorithmiques. Il est en effet tentant, à la suite des scandales américains, d'appeler à une loi pour une réglementation des algorithmes. Néanmoins, l'incertitude sur l'étendue et la forme de leur développement futur et le faible nombre de cas de biais algorithmiques avérés à l'heure actuelle en Europe et en France appellent à la prudence.

Par ailleurs, la vigueur de la compétition internationale dans l'économie numérique fait peser le risque d'un déclassement de la France et de l'Europe en cas de réglementation inadaptée des algorithmes, avec des conséquences importantes pour l'innovation.

Enfin, une législation conséquente existe déjà en matière de discrimination et de numérique. Celle-ci donne de nombreux moyens pour limiter les risques de biais algorithmiques, notamment grâce aux obligations de transparence et aux recours possibles. Examiner dans le détail leur mise en oeuvre et les obstacles à leur efficacité serait essentiel avant d'envisager une nouvelle initiative législative.

A. Les lois contre les discriminations s'appliquent aux algorithmes

Les discriminations existaient bien avant le numérique et de nombreux textes réglementaires en Europe et en France les interdisent. Ainsi, la loi française[27] explicite 25 critères comme l'âge ou le sexe dont l'usage est considéré comme discriminatoire dans des situations d'accès à l'emploi, aux services publics et privés, ou encore à la protection sociale.

Au niveau européen, pas moins de cinq directives européennes définissent les garde-fous en matière de discrimination, que ce soit dans le monde du travail ou dans l'accès aux services[28].

Lorsque ces textes prohibent l'usage de critères comme le sexe en matière d'accès à l'emploi, cela s'applique aux recruteurs, qu'ils soient humains, ou algorithmiques. L'arsenal juridique actuellement en vigueur est de ce point de vue transposable aux algorithmes et à leurs éventuels biais.

Les 25 critères de discrimination reconnus par la loi du 27 mai 2008

La loi interdisant les discriminations définit 25 critères dont l'usage est prohibé dans 7 situations distinctes.

Les 25 critères : 16 d'entre eux sont reconnu au niveau européen, à savoir l'âge, le sexe, l'origine, l'appartenance réelle ou supposée à une ethnie, l'état de santé, la grossesse, le handicap, les caractéristiques génétiques, l'orientation sexuelle, l'identité de genre, les opinions politiques, les activités syndicales, les opinions philosophiques et la religion.

Neuf critères complémentaires sont reconnus dans le droit français : la situation de famille, l'apparence physique, le nom, les mœurs, le lieu de résidence, la perte d'autonomie, la vulnérabilité économique, la domiciliation bancaire, la capacité à s'exprimer en français.

.../...

27 Loi 2008-496 du 27 mai 2008 portant diverses dispositions d'adaptation au droit communautaire dans le domaine de la lutte contre les discriminations.

28 Directive 2000/43/CE du 29 juin 2000 relative à l'égalité de traitement entre personne sans distinction d'origine ethnique ; Directive 2000/78/CE du 27 novembre 2000 relative à l'égalité générale en matière d'emploi ; Directive 2002/73/CE du 23 septembre 2002 relative à l'égalité entre hommes et femmes au travail ; Directive 2004/113/CE du 13 décembre 2004 relative à l'égalité entre les femmes et les hommes dans l'accès aux biens et services ; Directive 2006/54/CE du 5 juillet 2006 relative à l'égalité des chances entre hommes et femmes en matière d'emploi.

Les situations : sept situations sont reconnues, c'est-à-dire l'accès à l'emploi, la rémunération, l'accès aux biens et services publics et privés, l'accès à un lieu accueillant du public, l'accès à la protection sociale, l'éducation et la formation.

Pour caractériser un traitement discriminatoire, deux conditions sont nécessaires :

- le traitement doit être fondé sur l'un des 25 critères protégés par la loi, et
- le traitement doit concerner une des sept situations mentionnées dans la loi.

B. Les lois du numérique existantes limitent la possibilité de biais

I. Au niveau européen…

Au-delà de la lutte contre les discriminations, les lois relatives au numérique contribuent à encadrer les pratiques. Les biais algorithmiques ne font pas l'objet d'un corpus juridique à proprement parler ; néanmoins, les lois encadrant la manipulation des données personnelles limitent les risques pour les citoyens européens d'être victimes de biais algorithmiques.

L'Union Européenne a rénové son cadre de protection des données personnelles en 2016 avec le règlement général sur la circulation et la protection des données[29] (RGPD). Il remplace la directive sur la protection des données personnelles de 1995[30]. Le RGPD, première législation au monde adaptée au *Big Data* et à l'IA, définit un certain nombre de droits et de devoirs relatifs à la manipulation de données personnelles. Il s'agit notamment de garantir qu'un traitement de données personnelles (a fortiori par un algorithme) soit licite, loyal et transparent. Le RGPD institue par ailleurs de nouveaux droits pour le citoyen, notamment d'objection, d'explication et d'accès.

29 Règlement 2016/679 du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données.
30 Directive 95/46/CE du 25 octobre 1995, abrogée le 24 mai 2018.

RGPD et nouveaux droits numériques

Le RGPD est un règlement européen qui vise à protéger les données personnelles des résidents européens. Entré en vigueur en 2016, il introduit trois droits fondamentaux pour le citoyen.

Droit d'objection : l'article 22 du règlement, intitulé "décision individuelle automatisée, y compris le profilage", définit le droit à l'objection de la manière suivante :

« La personne concernée a le droit de ne pas faire l'objet d'une décision fondée exclusivement sur un traitement automatisé, y compris le profilage, produisant des effets juridiques la concernant ou l'affectant de manière significative de façon similaire. »

Tout citoyen européen peut donc s'opposer à une décision algorithmique, à condition qu'elle soit totalement automatisée et qu'elle ait des conséquences juridiques ou qui l'affectent de manière significative. À titre d'exemple, est considéré comme significatif des décisions en matière de demande de crédit, d'offre d'emploi ou encore de justice.

Droit d'explication : le droit à l'explication oblige l'utilisateur de l'algorithme employant des données personnelles à informer le citoyen objet d'une décision algorithmique. Ce droit est affirmé dans l'article 13, paragraphe 2(f) du règlement qui définit les informations à fournir et notamment :
« *L'existence d'une prise de décision automatisée, […] et, […] des informations utiles concernant la logique sous-jacente, ainsi que l'importance et les conséquences prévues de ce traitement pour la personne concernée.* »

Le caractère explicable des décisions de l'algorithme est donc indispensable depuis l'entrée en vigueur du RGPD.

Droit d'accès : le droit d'accès défini dans l'article 15 permet à chaque citoyen de demander l'accès, la correction ou l'effacement de données personnelles le concernant. Chacun peut donc agir pour obtenir des informations précises sur les données utilisées pour une décision automatisée lorsqu'il suspecte la présence d'un biais algorithmique.

Ces nouveaux droits fournissent au citoyen quelques moyens de recours contre les décisions algorithmiques qui seraient biaisées. Ils nécessitent cependant une attitude proactive de la part de la personne concernée pour obtenir plus d'informations et contester un algorithme qui serait biaisé. Dans le monde numérique, obtenir cette proactivité peut être difficile. L'absence de lecture par les internautes des conditions d'utilisation des données personnelles qui sont à présent affichées lors de l'entrée sur un nouveau site web est un exemple de la réticence à se plonger dans ces sujets arides et complexes.

Si le RGPD pose bien les bases d'une législation relative aux biais algorithmiques, il souffre néanmoins de plusieurs limites.

Premièrement, il reste un texte assez général, aux possibilités d'exemption multiples. Les jurisprudences des juridictions nationales et de la Cour de Justice de l'Union Européenne en définiront la portée réelle. Ainsi, la législation nationale peut limiter ces nouveaux droits dans les domaines de la sécurité nationale, de la défense nationale, de la sécurité publique, de la justice et pour des objectifs d'intérêt général dans les domaines monétaire, budgétaire, fiscal, de la santé et de la sécurité sociale[31]. Il existe ainsi une directive spécifique concernant la protection des données personnelles pour les activités de police et de justice[32].

Deuxièmement, le règlement ne précise rien concernant des algorithmes reproduisant à grande échelle des biais de société, qui ne seraient probablement pas contraires au règlement en matière de données personnelles.

Par ailleurs, en limitant la collecte et la conservation de données personnelles et sensibles, le RGPD restreint la capacité à tester l'absence de biais d'un algorithme.

Enfin, le droit européen cible uniquement les algorithmes quand ceux-ci utilisent des données personnelles, ou prennent des décisions portant sur des individus. Or, certains algorithmes peuvent avoir des impacts sur les citoyens, sans utiliser de données personnelles ou sans prendre de décisions les concernant directement.

31 Article 23 concernant les limitations du règlement.
32 Directive 2016/680 du 27 avril 2016.

Quand la CNIL met en garde contre les algorithmes sans données personnelles

La CNIL a été chargée de se prononcer sur les enjeux éthiques liés aux technologies numériques. Dans son document de synthèse sur le débat public concernant l'éthique des technologies numériques publié en 2017, elle alerte sur le vide juridique concernant les algorithmes n'utilisant pas de données personnelles. Ceux-ci peuvent, malgré l'absence de décision directe sur les citoyens, représenter des risques.

Par exemple, on peut considérer un algorithme définissant pour une région les plats consommés dans les cantines en fonction de données non personnelles comme les historiques des menus servis par les cantines. Cet algorithme pourrait comporter des biais, comme celui de favoriser certains types d'aliments au-delà de ce qui est souhaitable, ou encore celui d'attribuer certains repas plus régulièrement à un type de lycée qu'à un autre, sans motif explicable. Il n'est pas aujourd'hui certain que ce biais, de nature collective, rentrerait dans le champ de la législation sur les données personnelles, malgré son impact sur des individus.

Au-delà de cette mise en garde, la CNIL fait émerger, tant pour les algorithmes du secteur privé que du secteur public, les principes de loyauté et de vigilance de l'algorithme. Si le premier est à présent évoqué dans plusieurs textes juridiques, le second, plus récent, consiste à organiser, par des moyens techniques et humains, l'évaluation régulière du fonctionnement et des résultats d'un algorithme, que celui-ci soit prédictif ou non.

Cette recommandation éclaire sur les priorités que souhaite se donner l'écosystème français en matière d'algorithmes : une loyauté de l'algorithme, pour la personne visée mais également vis-à-vis de la société, et une vigilance de tous les instants avant et surtout pendant l'usage d'un algorithme.

Malgré ses limites, le RGPD est un outil qui possède des leviers puissants pour garantir sa mise en œuvre. Il a par nature une portée extraterritoriale. En effet, il s'applique à toute entité ou personne, manipulant des données à caractère personnel de citoyens européens. Une entreprise américaine du numérique ayant pour utilisateurs des citoyens européens est ainsi pleinement concernée par ce règlement. Les pénalités en cas de non-respect de ses dispositions sont conséquentes et peuvent s'élever à 20 millions d'euros ou 4% du

chiffre d'affaire mondial. Il existe donc des moyens conséquents prévus par ce règlement pour contraindre les acteurs à respecter les nouveaux droits numériques des citoyens.

II. … et au niveau national

Au-delà de cette approche européenne, dont l'essentiel est défini dans le RGPD, le droit national français en matière numérique apporte également de nombreuses réponses au sujet des biais algorithmiques.

La loi pour une République numérique de 2016[33] a autorisé la décision automatisée dans le domaine public[34] sous deux conditions. Le citoyen doit être informé que la décision le concernant a été prise de manière automatisée et il peut, à sa demande, obtenir des informations sur l'algorithme et, notamment, « le degré et le mode de contribution du traitement algorithmique à la prise de décision, les données traitées et leurs sources, les paramètres de traitement et, le cas échéant, leur pondération, appliqués à la situation de l'intéressé, les opérations effectuées par le traitement ».

Cette loi a un impact fondamental pour les décisions individuelles automatisées : celles-ci ne sont pas légales si l'algorithme n'est pas techniquement explicable ou si certains éléments de son fonctionnement ne peuvent pas être communiqués pour des raisons légales (secret fiscal ou secret défense par exemple). La transparence totale de l'algorithme est donc obligatoire, dès lors que l'algorithme prend seul une décision individuelle, ce qui va au-delà du règlement européen.

33 Loi 2016-1321 du 7 octobre 2016 (lien) dont le rôle a notamment été de mettre à jour la loi informatique et libertés de 1978, fondatrice en matière de protection des droits dans le domaine numérique, apporte certaines spécificités en matière de biais algorithmiques.

34 Ce type d'algorithmes existe déjà, pour le calcul du montant des impôts à payer par exemple.

Quand le Conseil Constitutionnel censure le *machine learning*

La loi pour une République numérique rend peu probable de voir émerger des algorithmes de *machine learning* pour des décisions publiques à caractère individuel. Il serait en effet compliqué de garantir une explicabilité totale pour ces technologies dont les arbres de décision sont par construction difficiles à interpréter.

Néanmoins, le Conseil Constitutionnel a restreint davantage les possibilités d'emploi du *machine learning* pour des décisions publiques.

Dans son avis rendu le 12 juin 2018 sur cette même loi, il précise qu'un algorithme auto-apprenant, c'est-à-dire révisant lui même ses règles de fonctionnement, sans l'avis et le contrôle du responsable de traitement, ne peut conduire à une décision individuelle automatisée.

Autrement dit, les algorithmes auto-apprenants ne sont pas constitutionnellement autorisés à prendre des décisions pour le compte de la puissance publique. Le risque de biais en la matière est donc plus limité.

Si la loi pour une République numérique renforce les exigences de transparence en matière d'algorithmes publics et restreint l'usage de décisions automatisées, elle ne contraint pas davantage le cadre pour les algorithmes privés, qui restent bien entendu soumis au RGPD.

Enfin, la France adopte progressivement, par des législations sectorielles, de nouvelles règles à respecter en matière d'usage des algorithmes. Sont notamment visés les secteurs de la santé, du transport, ou de l'information.

Le projet de loi relatif à la bioéthique et déposé à l'Assemblée Nationale le 28 juillet 2019[35] instaure dans son article 11 des obligations lors de l'usage d'algorithmes de traitement de données massives à visée de prévention, de diagnostic ou thérapeutique. En particulier, le professionnel de santé devra informer le patient de l'usage d'un algorithme et de ses modalités de fonctionnement. Il conservera par ailleurs la possibilité de modifier les paramètres de l'algorithme, dont les comportements seront enregistrés.

35 Projet de loi n°2187 relatif à la bioéthique, 24 juillet 2019. Promulgation attendue à l'été 2020.

La loi d'orientation des mobilités[36] du 24 décembre 2019 définit des obligations de transparence pour les plateformes, notamment sur les critères utilisés par les algorithmes aboutissant à des décisions sur la rémunération des chauffeurs ou encore la localisation des missions qui leurs sont assignées.

Enfin, la proposition de loi visant à lutter contre la haine sur internet[37] devrait également renforcer les règles s'appliquant aux algorithmes dans le domaine des fake news. Le texte impose aux plates-formes, suivant une logique d'obligation de résultats, de retirer, dans les 24 heures de leur publication, tout contenu haineux. Le contrôle *ex post* des erreurs éventuelles d'un algorithme de tri des contenus est une disposition qui a également été adoptée dans d'autres pays comme l'Allemagne[38] et qui permet de s'assurer du respect de la loi sans avoir à expliquer ou maîtriser le fonctionnement de l'algorithme.

Ces différents textes de lois ont néanmoins une limite en matière de biais algorithmiques : ils portent davantage sur des mesures intervenant après l'usage d'un algorithme biaisé, notamment en instaurant droits de recours et transparence. Ils ne proposent donc rien en matière préventive, c'est à dire avant qu'un algorithme éventuellement biaisé entre en production.

C. Face aux États-Unis, le droit européen et le droit français développent leurs spécificités en matière d'algorithmes

Les protections contre les biais algorithmiques sont nombreuses, qu'elles se situent dans le champ de la lutte contre les discriminations ou dans celui des lois du numérique (RGPD, loi pour une République numérique, etc.).

Malgré le peu d'exemples de biais algorithmiques en France et en Europe, les lois diffèrent des modèles anglo-saxons sur plusieurs points.

Tout d'abord, l'usage de données concernant l'appartenance ethnique des individus est un point important de divergence. Contrairement au Royaume-Uni ou aux États-Unis, leur collecte et leur usage sont jugés anticonstitutionnels en France. Si des

36 Projet de loi et rapport annexe de la loi n° 2019-1428 du 24 décembre 2019 d'orientation des mobilités.
37 Proposition de loi n° 1 785 visant à lutter contre la haine sur internet du 20 mars 2019.
38 Orsini,A. *Discours haineux : les réseaux sociaux risquent 50 millions d'euros d'amende en Allemagne* dans Numerama, 2 janvier 2018.

exceptions à la collecte statistique sont prévues par la loi informatique et liberté de 1978[39] pour autoriser la collecte de données sur la santé, l'opinion politique ou religieuse ou encore l'orientation sexuelle, la collecte de données relatives à l'origine ethnique est spécifiquement exclue. Une telle collecte est pourtant la pierre angulaire de la lutte contre les biais et de la discrimination positive aux États-Unis.

La tradition universaliste française se refuse en effet à identifier des communautés ethniques au sein de la République. En 2007, la loi relative à la maîtrise de l'immigration, à l'intégration et à l'asile[40] contenait une disposition visant à autoriser la collecte de telles données. Cette disposition a été censurée par le Conseil Constitutionnel[41], qui a invoqué l'article 1er de la Constitution, qui prévoit « l'égalité de tous les citoyens devant la loi, sans distinction d'origine, de race ou de religion ». Dans le commentaire de sa décision, le Conseil Constitutionnel considère que des études peuvent être menées sur des données objectives, comme le lieu de naissance, le nom ou la nationalité, ou encore subjectives, comme le ressenti d'appartenance. Assigner aux individus une identité ethnique est en revanche formellement interdit.

Ensuite, les États-Unis n'ont pas encore de législation fédérale concernant la protection des données personnelles ou encore les algorithmes. Si la Federal Trade Commission (FTC) émet bien quelques recommandations à partir de lois sectorielles (marchés financiers, protection de la jeunesse, etc.), ce sont plutôt les États fédérés qui se saisissent de ces sujets.

L'Illinois a récemment adopté une loi qui oblige les entreprises utilisant des algorithmes d'aide à la décision à des fins de recrutement à informer les candidats de l'usage d'un algorithme, à expliquer le fonctionnement et les critères utilisés par l'algorithme et enfin à solliciter le consentement des postulants.

Le Massachussetts ou la Californie ont par ailleurs mis en œuvre des lois proches du RGPD avec l'instauration d'obligations de transparence, et de droit d'accès pour le citoyen. Ces textes se limitent néanmoins au traitement de données personnelles et ne concernent pas d'éventuelles décisions automatisées.

Au niveau fédéral, l'idée d'une régulation spécifique des données personnelles n'est pas consensuelle. Internet est en effet considéré - notamment par la Cour Suprême - comme un espace public, et les informations que l'on y publie sont, par conséquent,

39 Des exceptions à cette interdiction existent dans le domaine médical, de la sécurité, ou encore pour des données qui seraient anonymisées.
40 Loi n°2007-1631 du 20 novembre 2007.
41 Décision du Conseil Constitutionnel, 2007-557 DC.

considérées comme publiques par de nombreux responsables. L'usage des données personnelles et des algorithmes est ainsi encadré davantage dans des textes sur le droit du consommateur ou de lutte contre les discriminations.

L'*Algorithm Accountability Act*, un projet de loi sur les biais algorithmiques aux États-Unis

Alors que la prétention à la vie privée sur internet est remise en cause, les discriminations causées par les biais algorithmiques sont un sujet fréquent de controverse aux États Unis. Les scandales concernant les annonces immobilières de Facebook et l'algorithme d'évaluation de CV d'Amazon ont poussé les démocrates à proposer au Sénat, en avril 2019, une loi d'encadrement des algorithmes intitulée *Algorithm Accountability Act*.

Si ce texte était voté en l'état, il obligerait l'administration fédérale, les États et les entreprises réalisant plus de 50 millions de dollars de chiffre d'affaires et contenant des informations sur plus d'un million de personnes, à réaliser des études d'impact concernant leurs algorithmes de décision ou d'aide à la décision automatisée dite à risque *("high risk automated decision")*.

Les algorithmes à risque sont ceux affectant des éléments sensibles de la vie privée comme la performance au travail, la santé, la vie personnelle, la localisation ou utilisant des données comme la race, le genre, l'opinion religieuse ou encore les opinions politiques. Ils incluent également la surveillance d'espaces publics importants, et donc les potentiels usages de reconnaissance faciale.

Ces études d'impact devraient vérifier le comportement de l'algorithme à l'aune des critères suivants : *accuracy, fairness, bias, discrimination, privacy and security*. En cas de manquement, l'entité devrait y remédier et la FTC serait dotée d'un pouvoir de sanction.

Cette loi pourrait recueillir un soutien des républicains, qui manifestent leur mécontentement face à Facebook, Google et Twitter, accusés de biais politiques en faveur des démocrates dans leurs algorithmes de recommandations.

D. L'application aux algorithmes du droit existant est aujourd'hui imparfaite et difficile

Le droit, qu'il porte sur les discriminations ou sur le numérique, est d'ores et déjà développé. Les entités chargées de son application sont à la fois des organismes spécialisés sur ces thématiques, comme le Défenseur des droits ou la CNIL, et des régulateurs sectoriels comme l'ARCEP pour les télécoms, le CSA pour les médias ou l'ACPR pour la banque et l'assurance. Dans les deux cas, il est actuellement difficile pour ces entités d'appliquer pleinement le droit existant.

Pour les régulateurs sectoriels, des conflits d'intérêt peuvent limiter la capacité à lutter contre les biais algorithmiques. En effet, ces entités ont souvent un objectif principal éloigné de ce type de sujets.

L'assurance, la mutualité et la banque disposent par exemple d'un régulateur dédié, l'ACPR, qui ne s'intéresse pas uniquement aux variables utilisées dans les modèles et algorithmes d'évaluation des risques et de tarification, mais également à leur qualité, car la stabilité financière dépend de la bonne évaluation des risques. L'ACPR préfèrera peut-être des algorithmes biaisés mais précis pour garantir la stabilité financière, à l'absence de biais dans les algorithmes. Le conflit entre ces deux objectifs oblige donc probablement à réfléchir au partage des responsabilités pour la régulation des biais dans les algorithmes d'assurance.

En ce qui concerne les régulateurs transverses comme le Défenseur des droits et la CNIL, les difficultés résident aujourd'hui dans leur capacité à réguler dans les faits un très grand nombre d'entités. Il est difficile pour la CNIL, qui dispose d'un budget de 18 millions d'euros et 210 agents, d'auditer pleinement les 80 000 organismes qui doivent aujourd'hui désigner des délégués aux données personnelles[42]. En 2018, 310 contrôles seulement ont ainsi été réalisés.

La CNIL est aujourd'hui confrontée à de réelles difficultés de moyens pour mettre en œuvre le RGPD. Ajouter de nouvelles prérogatives en matière de biais algorithmiques via de nouvelles réglementations ne ferait qu'aggraver cette situation.

C'est pourquoi il est essentiel aujourd'hui que la France se focalise sur la pleine application de la réglementation existante en matière de discrimination dans le domaine numérique. Il existe de nombreuses dispositions qui permettent de prévenir les effets négatifs de biais, à condition de les appliquer progressivement aux nouveaux algorithmes.

42 Sénat, Projet de loi de finances pour 2019 : Direction de l'action du gouvernement, publication officielle et information administrative.

IV

RECOMMANDATIONS

Cela fait longtemps que les informaticiens et statisticiens connaissent les biais, puisque les biais techniques sont inhérents à tout algorithme. En revanche, les managers, fonctionnaires, juges et citoyens qui y sont maintenant confrontés découvrent, parfois avec effroi, ce problème, souvent à l'occasion de scandales américains. Il existe désormais un constat partagé sur la présence croissante des algorithmes dans nos vies et sur leurs biais possibles. Face à cela, deux approches s'affrontent dans le débat public.

Certains défendent que si l'algorithme n'est pas capable de montrer son absence de biais, il ne doit pas être utilisé. Les outils adéquats sont alors une loi des algorithmes qui interdit ou restreint leur usage dans de nombreux cas. Le corollaire est un régulateur fort, dont la mission est de vérifier que les algorithmes en service ne sont pas biaisés. Les partisans de cette approche n'hésitent plus à qualifier les algorithmes d'arme de destruction mathématique[43]. Les bénéfices apportés par les algorithmes n'en vaudraient pas les risques.

L'autre camp, défend au contraire qu'il est plus que jamais nécessaire de profiter des innovations apportées par le numérique. Même si les algorithmes sont biaisés, ils le seront toujours moins que les humains. Par ailleurs, si nous ne développons pas ces technologies aujourd'hui, les États-Unis et la Chine s'en chargeront pour nous. En plus d'être biaisés, ils seront contrôlés pas des entités hors d'Europe, comme c'est déjà le cas pour tous les services algorithmiques produits par les GAFA. L'impact économique du retard pris en matière numérique sera alors impossible à rattraper et fatal à la prospérité européenne. Il faudrait donc ne surtout pas proposer de nouvelles réglementation et laisser les acteurs publics et privés développer comme bon leur semble des algorithmes, en facilitant autant que possible l'innovation. Après tout, les avantages valent bien quelques biais.

Ces deux approches sont, à plusieurs égards, excessives. S'il est clair que les biais des algorithmes présentent un danger, il faut les mettre en regard des opportunités que la révolution numérique propose. Un algorithme est avant tout un formidable moyen de découvrir et de réduire les biais existants chez les humains. Il oblige à formaliser les choix que nous effectuons en matière d'équité. Un biais algorithmique est-il néfaste, s'il est bien moindre qu'un biais existant? L'algorithme, c'est aussi le

42 O'Neil C., *Weapons of Math Destruction*, Penguin Books, Juin 2017.

passage à grande échelle de comportements observés localement. Si les bases de données d'entraînement sont biaisées, qui acceptera de voir se répandre à grande échelle des biais dans notre pays ? Les exemples de la justice américaine sont, à ce titre, inquiétants.

Tout le défi réside donc dans l'équilibre entre un accompagnement de l'innovation et une bonne prise en charge des risques de discrimination. Nous sommes convaincus qu'une approche pas à pas est nécessaire.

Les usages des algorithmes et nos connaissances sur leurs biais évoluent en permanence. L'enjeu est donc de développer notre capacité à comprendre, à détecter et à réagir à ces biais. Avec un objectif : renforcer la confiance dans le fonctionnement et l'équité des algorithmes pour accélérer leur diffusion.

Nos recommandations tiennent en trois axes : prévenir l'introduction de biais algorithmiques, notamment via la formation des équipes, détecter leur présence via des tests internes aux organisations, et faire évaluer par des tiers les algorithmes à fort impact.

A. Les propositions que ce rapport a choisi de ne pas retenir

Les algorithmes et leurs défauts sont un sujet anxiogène pour de nombreux européens, particulièrement en France. Les biais algorithmiques sont, à raison, vus comme les prémices d'une société où la technologie aurait fait disparaître l'égalité, sur l'autel de l'optimisation, de la statistique et du progrès technologique. La perte de contrôle vis-à-vis des programmeurs est la première chose qui vient en tête des européens lorsqu'il est question d'algorithme, selon une enquête[44] de la fondation Bertelsmann de 2018, les Français étant les plus inquiets. Ceux-ci sont d'ailleurs 72 % à estimer qu'un recrutement aidé d'un algorithme serait d'abord une menace, selon une enquête[45] citée par la CNIL en 2017.

Face à cela, de nombreuses voix s'élèvent pour des décisions fortes, rapides et emblématiques. Nous pensons qu'un tel sujet ne doit pas donner lieu à des initiatives hâtives. Il existe donc un certain nombre de propositions que vous ne verrez pas dans ce rapport.

• Non proposition 1 : une loi portant sur les biais des algorithmes

L'ambition de la nouvelle Commission européenne de proposer des initiatives pour favoriser l'émergence d'une IA éthique, responsable et sans biais nous paraît tout à fait positive. Nous sommes par ailleurs complètement en ligne avec l'approche consistant à concentrer les efforts sur les algorithmes risquant d'avoir un fort impact. Néanmoins, nous pensons prématuré à ce stade de travailler à une directive sur l'IA et l'éthique des algorithmes[46] en ce qui concerne les biais des algorithmes.

Comme nous l'expliquions dans la première partie de ce rapport, les cas avérés de biais algorithmiques en Europe et en France sont encore très limités. Il est dès lors difficile de produire des réglementations équilibrées sur des problèmes pour lesquels il n'y a pas de recul suffisant. Légiférer sans avoir pris le temps d'observer et d'analyser finement le phénomène, c'est prendre le risque d'une surréglementation nuisible à la numérisation de l'économie et à l'innovation. À l'inverse, sous-réglementer risquerait de laisser se produire des atteintes aux libertés et à l'égalité, entraînant nécessairement des initiatives législatives soudaines suite à un scandale.

44 Grzymek, V. et Puntschuh, M. *What Europe Knows and Thinks About Algorithms*, Bertelsmann Stiftung, Février 2019.
45 Cité dans *Comment permettre à l'homme de garder la main?*, synthèse du débat public confié à la CNIL par la loi pour une République numérique, Décembre 2017.
46 European Parliament, legislation train Schedule, *Communication on artificial intelligence for Europe*, Mai 2018.

Le deuxième obstacle majeur à une loi *biais des algorithmes* tient dans l'extraordinaire diversité des usages d'algorithmes. La banque, l'assurance, l'automobile, la santé, le recrutement, la publicité, la justice ou la police sont quelques-uns des domaines où le déploiement d'algorithmes promet d'être massif. À ce titre, comment définir des règles qui s'appliqueront à tous, sans tenir compte des spécificités sectorielles ? La définition d'un biais et de ce qu'est l'équité sera immanquablement différente selon que l'on discute d'un algorithme de conduite autonome, de mise au point d'un protocole de chimiothérapie ou de ciblage publicitaire.

Promouvoir une « loi pour prévenir et combattre les biais des algorithmes » en 2020 serait donc à nos yeux une approche inadaptée. La mobilisation des acteurs de la société civile et des entreprises nous paraît, à ce stade, bien plus propice, en attendant d'y voir plus clair. Cette approche n'altère cependant en rien la capacité de l'État à prendre des initiatives réglementaires pour les cas d'usages les plus critiques qui sont, bien entendu, nombreux.

• Non proposition 2 : un contrôle des algorithmes par l'État

Le RGPD et sa mise en œuvre en France par la CNIL ont montré les limites de ce modèle de régulation. Attendre de la CNIL qu'elle contrôle et autorise l'ensemble des traitements de données personnelles qui se développent dans la société française est une illusion. L'obligation pour le responsable du traitement de données personnelles de réaliser une étude d'impact à la place d'une déclaration auprès de la CNIL est la conséquence d'une réalité brutale : le monde numérique est bien trop vaste et mouvant pour être contrôlé *ex ante* par une autorité indépendante, même dotée de moyens considérables.

Croire que ces difficultés à réaliser un contrôle *ex ante* du numérique par un régulateur ne concernent que les données personnelles et que la situation serait différente pour les algorithmes est tout aussi illusoire. C'est pourquoi nous pensons que demander à l'État de contrôler l'absence de biais dans les algorithmes n'est ni faisable ni souhaitable. Il faut dès à présent envisager que ce rôle de contrôle des biais algorithmiques soit au moins partiellement transféré aux entreprises mettant en œuvre des algorithmes, à des laboratoires de recherche et de certification, ou à des tierces parties sur le modèle des sociétés d'audit et de contrôle des comptes financiers des entreprises par exemple. Ce contrôle décentralisé devra s'accompagner en parallèle d'une responsabilisation des acteurs face aux conséquences des biais des algorithmes.

B. Prévenir les biais en répandant des bonnes pratiques et des efforts de formation pour tous ceux qui produisent ou utilisent des algorithmes

Afin que les technologies d'IA soient intégrées avec succès et soient bénéfiques pour notre société, nous recommandons que toute la chaîne des acteurs impliqués dans la production ou affectés par les décisions des algorithmes soit correctement formée aux risques de biais et discriminations et comprenne les risques et les avantages du déploiement des algorithmes. Le déploiement des meilleures pratiques dans les entreprises et administrations, notamment en matière de diversité des équipes, est également essentiel à cette fin.

Certains algorithmes sont développés en interne et ensuite déployés au sein de la même organisation (par exemple l'algorithme de recrutement d'Amazon). D'autres sont achetés à des fournisseurs puis déployés par des organisations qui ne les ont pas conçus (par exemple l'algorithme de prédiction de la criminalité PredPol utilisé par la police de Los Angeles).

Il existe toute une chaîne de production des algorithmes :

Management : les techniciens sont encadrés par des managers, qui font des choix sur les ressources à allouer au projet, les objectifs à optimiser. Ils peuvent décider que la détection de potentiels biais n'est pas prioritaire au vu des risques estimés, et peuvent fixer un objectif stratégique pour l'algorithme qui reproduit des biais de société.

Développement : les développeurs informatiques ou *data scientists* qui codent les algorithmes en sont les architectes. Les développeurs ne travaillent en général pas de façon isolée : le *machine learning*, dans l'esprit du développement informatique, est largement publié en *open source*, sur internet. De grandes entreprises développent leurs propres outils mais les mettent ensuite à disposition de tous les développeurs. En 2015, Google a ainsi publié TensorFlow, une bibliothèque qui permet à chacun de construire son réseau de neurones.
Lorsqu'en 2018 une étude du MIT Media Lab prouve que les algorithmes d'IBM et de Microsoft reconnaissent le genre de 99 % des hommes à la peau pâle mais de 65 % seulement des femmes à la peau foncée, le New York Times interroge la diversité des équipes qui ont conçu ces algorithmes.

.../...

Récolte des données : les données d'entraînement, cruciales avant même d'utiliser l'algorithme sur ses propres données, sont rarement collectées en chambre. Des bases de données sont publiées par des chercheurs. ImageNet, la base de données développée à Princeton, a donné accès à 14 millions d'images annotées sur 20 000 catégories, et permis de les utiliser pour entraîner des algorithmes de reconnaissance visuelle.
Néanmoins tout biais dans une base de données de cette taille peut créer des biais dans de nombreux algorithmes. Le projet ImageNet Roulette d'AI Now, a ainsi révélé qu'un entraînement sur ImageNet pouvait entraîner de multiples biais (des personnes noires aux *selfies* labellisés « criminel » ou « condamné »). ImageNet a ainsi retiré plus de 60 000 images de sa base pour tenter d'y remédier.

Étiquetage des données : Il existe par ailleurs toute une industrie de l'étiquetage des données, qui intervient de la reconnaissance d'image à la modération de contenus. Les décisions de ces travailleurs du clic peuvent ne pas correspondre aux attentes des utilisateurs, a fortiori quand une grande partie de l'industrie est délocalisée dans des pays à bas salaires, dont les repères culturels peuvent être différents. C'est toute la difficulté de la tâche demandée aux modérateurs de Facebook, qui créent des bases de données d'entraînement pour les algorithmes de modération, et y incluent leurs propres biais.

Utilisation : l'algorithme n'est souvent qu'une aide à la décision. Ce sont par exemple les juges qui décident de la remise en liberté d'un détenu ou les recruteurs d'Amazon qui envoient une offre d'emploi. Les biais des algorithmes n'ont d'effets que lorsqu'ils ne sont pas filtrés ou rectifiés par ceux qui prennent des décisions. Certaines banques, plutôt que d'imposer des exigences d'équité dans leurs algorithmes, décident par exemple de confier la prédiction du risque de crédit à une personne qui aura la charge de rectifier la recommandation de l'algorithme avec sa propre expérience.

.../...

Feedback («renforcement») : les algorithmes sont insérés dans des systèmes qui collectent des données d'utilisation et permettent d'en affiner le fonctionnement et d'en optimiser la performance. Quand j'échange avec un agent conversationnel (un *chatbot*), j'oriente le système. Tay, un agent conversationnel lancé par Microsoft en 2016 sur Twitter a ainsi rapidement proféré des insanités, des injures raciales, ne faisant que reproduire le comportement des internautes qui « testaient » le système en lui envoyant des injures. Son compte a été fermé en moins de vingt-quatre heures. Depuis le web 2.0, où l'interaction est la règle, toute boucle de rétroaction sur les systèmes intelligents est susceptible de créer un biais dans l'algorithme.

Avec tant d'acteurs, la question de la responsabilité en cas de biais algorithmique est cruciale. Si la police prédictive est biaisée racialement, qui est responsable ? Est-ce l'utilisateur (le policier de Los Angeles), est-ce le manager/acheteur (le responsable informatique de la police de Los Angeles), est-ce le développeur, est-ce le fournisseur de données biaisées ?

• Proposition 1 : déployer des bonnes pratiques pour prévenir la diffusion de biais algorithmiques (chartes internes, diversité des équipes)

Les biais algorithmiques sont un danger réel pour les citoyens mais également pour les entreprises et administrations qui les conçoivent ou les déploient. Prévenir les biais sera bien moins coûteux que de les corriger. Car, au-delà du risque juridique, l'enjeu réputationnel est considérable. Il est donc essentiel que chaque maillon de la chaine mette en place des bonnes pratiques pour prévenir, détecter et alerter sur de possibles biais.

Sans grand renfort de publicité, plusieurs entreprises françaises mettent peu à peu en place des pratiques et des chartes pour faire face au risque de biais. Nous recommandons à tous les acteurs souhaitant construire et utiliser des algorithmes de suivre leur exemple en mettant en œuvre une charte de développement et de déploiement d'algorithmes équitables (cela ne s'appliquerait bien entendu pas aux algorithmes les plus anodins, comme ceux qui déterminent le volume idéal de la musique dans une voiture).

Sans être exhaustifs, quelques points méritent d'être inclus dans ces chartes internes :
- des exigences de méthodologie pour assurer la qualité des algorithmes ;
- les propriétés que doivent présenter les algorithmes développés (notamment si l'on

veut pouvoir les auditer plus tard dans le déploiement) ;
- les mécanismes internes pour gérer les tensions entre différents objectifs, définir des exigences d'équité pour les algorithmes, et préciser leur formalisation informatique ;
- les analyses et évaluations internes à faire subir à l'algorithme.

Quelques exemples de bonnes pratiques

Masquer des variables sensibles : certaines entreprises, notamment dans les secteurs de la banque, de l'assurance et du recrutement, isolent des variables protégées de leurs clients : l'âge, le genre, l'adresse, etc. Ces variables ne sont plus accessibles aux algorithmes d'évaluation des risques ou de tarification. Une telle pratique vise à réduire le risque qu'un algorithme utilise ces éléments comme facteur discriminant. Cette approche « d'équité par l'ignorance » a ses limites car des algorithmes de *machine learning* peuvent très bien déduire les variables protégées (par exemple le genre via le modèle et la couleur de voiture). Sur ce point, voir notre proposition 4.

Comparer les taux de faux positifs : dans le recrutement, de grandes agences vérifient de manière systématique que les taux de faux-positifs, c'est-à-dire d'erreurs de classification, dans leurs algorithmes, sont identiques pour différents sous-groupes de la population. Il s'agit d'éviter des scénarios où un logiciel serait bien plus efficace pour certaines personnes que pour d'autres, quitte à en niveler la performance par le bas. Cela vise à éviter des cas similaires à la reconnaissance faciale aux États-Unis.

Surveiller l'algorithme après son déploiement : les algorithmes qui apprennent et évoluent au fur et à mesure de leur utilisation sont de plus en plus nombreux. Certaines entreprises vérifient donc, à intervalle régulier, que ces algorithmes n'introduisent pas de nouveaux biais en réitérant des tests de leur équité.

Implication du management : plusieurs entreprises prévoient très précisément les cas dans lesquels le management ou le comité des risques doivent être sollicités (introduction d'une nouvelle variable, arbitrage à réaliser entre performance et équité, définition d'un critère d'équité à évaluer).

Cet extrait de bonnes pratiques a vocation à donner quelques idées aux entreprises déployant des algorithmes. Confrontées au risque de biais, elles pourront mettre en œuvre des chartes de développement d'algorithmes équitables dans leurs organisations. Cette étape est essentielle pour changer les comportements et prévenir les scandales de discrimination qui pourraient émerger.

Les algorithmes ne seront pas uniquement déployés dans des entreprises qui en ont une parfaite maîtrise. Certaines organisations achèteront des technologies matures pour augmenter leur productivité, pour assurer l'interaction avec les clients ou dans les ressources humaines. Or, les cadres de ces organisations-là ont généralement une connaissance bien plus limitée des risques et limites des algorithmes.

Comme pour tout achat sensible, l'acheteur doit être assez mature pour questionner son fournisseur et ses pratiques. Les bonnes pratiques se diffuseront donc d'autant plus vite que les acheteurs sauront être exigeants.

En matière de bonnes pratiques, il est nécessaire d'insister sur l'importance de mobiliser une diversité de profils, tant sociale que professionnelle, dans les projets de développement d'algorithmes. Par diversité, nous entendons des personnes avec des parcours différents et qui ont de ce fait été confrontés à une diversité de situations en lien avec le produit développé. Les cadres et les techniciens des entreprises centrées sur les données ont une très bonne compréhension des opportunités que représentent les algorithmes qu'ils développent pour des besoins externes et internes. Mais leur compréhension des défis d'équité peut parfois souffrir d'un manque de diversité, à la fois dans leurs rangs et dans les profils rencontrés dans l'élaboration d'un projet : les femmes sont plus susceptibles que les hommes de détecter le potentiel de harcèlement que représente une technologie en ligne, les plus jeunes auront un rapport très différent des seniors aux outils numériques.

Composer une équipe de développeurs réellement diverse en termes de genre, de parcours social, d'ethnie, ou autre, se heurte forcément au manque de diversité dans les formations d'informatique, de statistiques, ou de *data science*.

• Proposition 2 : former les techniciens et ingénieurs aux risques de biais et améliorer la connaissance citoyenne des risques et opportunités de l'IA.

Si les développeurs semblent en première ligne face aux biais, c'est en réalité l'intégralité des acteurs de la vie d'un algorithme qui sont concernés : scientifiques, responsables d'entreprise mais également citoyens, collaborateurs, dirigeants politiques, etc. Nous recommandons que chacun de ces groupes fasse l'objet de formations

sur le sujet des biais et de l'intelligence artificielle. Si la formation des scientifiques et développeurs pourra se focaliser sur les biais algorithmiques et notamment les biais de société, les contenus à destination du grand public devront offrir à chacun la possibilité d'appréhender les enjeux de l'intelligence artificielle.

En France, la plupart des développeurs ont été formés aux mathématiques appliquées, aux statistiques et à l'informatique, sans formation spécialisée en sciences sociales. On leur apprend à comprendre les défis techniques de la conception et de l'optimisation des algorithmes, et non les défis sociétaux.

Dans la terminologie technique, le biais représente tout ce qui dévie l'algorithme de la production du résultat optimal en vue duquel l'algorithme a été conçu. Il est limité aux biais que nous appelons « techniques » et n'inclut pas l'impact que les résultats optimaux produisent sur les groupes de personnes dont les données ont été utilisées pour « entraîner l'algorithme », ni sur ceux dont la vie sera affectée par la prédiction de l'algorithme.

Les experts des données ont besoin d'une formation spécialisée pour comprendre ce que sont les biais de société, ainsi que les différentes notions d'équité. Ils devront également apprendre à adapter ces cadres généraux à des situations et à des algorithmes spécifiques.

Ces experts et expertes devront comprendre l'importance de recueillir un échantillon de données d'apprentissage qui reflète correctement la population qui sera touchée par l'algorithme, au-delà d'une simple représentativité statistique qui pourrait ignorer des biais de société. Le problème vient souvent du fait que les données d'apprentissage représentent des populations familières aux *data scientists* qui ont conçu les algorithmes. Les algorithmes, apprenant sur des exemples limités, ne peuvent alors pas produire des décisions appropriées pour les populations qui sont exclues de l'échantillon d'apprentissage.

À titre d'exemple, les premiers algorithmes de reconnaissance faciale de Google classifiaient certaines personnes noires comme des gorilles ! L'entraînement de l'algorithme avait été fait à partir d'un ensemble limité de photos qui reflétaient le groupe social des concepteurs - des hommes blancs. La base de données ne contenait pas assez de visages des personnes de couleur pour permettre d'identifier correctement ces personnes dans de nouvelles images présentées à l'algorithme.

La formation de spécialistes de données et d'ingénieurs à l'assemblage d'un échantillon d'apprentissage qui représente correctement la population est l'un des prérequis les plus importants pour accroître l'équité dans la prise de décision algorithmique.

Cela passe avant tout par de la formation continue à travers les outils du XXIe siècle (discussions sur les plateformes de développement, MOOC), et pas uniquement par la formation en cycle universitaire. Cette dernière est importante, mais n'est pas suffisante, notamment car il s'agit de former également les personnes déjà en activité. Au-delà des scientifiques, les responsables d'entreprises prenant des décisions de mise en œuvre d'algorithmes doivent également être formés, mais sur des enjeux différents. Ils doivent avoir en tête les priorités que sont la mixité des équipes de développement et la mise en œuvre de processus pour évaluer le risque de biais dans les algorithmes. À cet égard, les bonnes pratiques mentionnées dans la première proposition du rapport et tirées d'exemples réels que nous avons rencontrés doivent servir d'exemples pédagogiques pour déployer plus largement ce type d'initiative.

Enfin, les citoyens, collaborateurs et dirigeants politiques doivent avoir conscience des enjeux liés au développement des algorithmes et de l'IA pour pouvoir répondre aux peurs et exercer un devoir de vigilance.

En effet, l'adoption de l'IA ne se fera pas en France sans répondre aux peurs qu'elle suscite, parfois dignes d'histoires de science-fiction, où les machines seraient sur le point de prendre le contrôle sur les humainsElles cohabitent avec de sérieuses inquiétudes sur l'impact de la technologie IA sur le marché de l'emploi et des préoccupations sur la façon dont les données privées sont utilisées sur Internet et sur les réseaux sociaux. Il est important de séparer le mythe de la réalité - le grand public veut et a le droit de comprendre comment les algorithmes sont conçus, quelles données sont utilisées, comment les biais seront traités et comment cette technologie aura un impact sur leur vie. Chacun et chacune doit avoir les moyens de cette compréhension qui permettra de créer une confiance mais aussi une vigilance éclairée en matière de technologies.

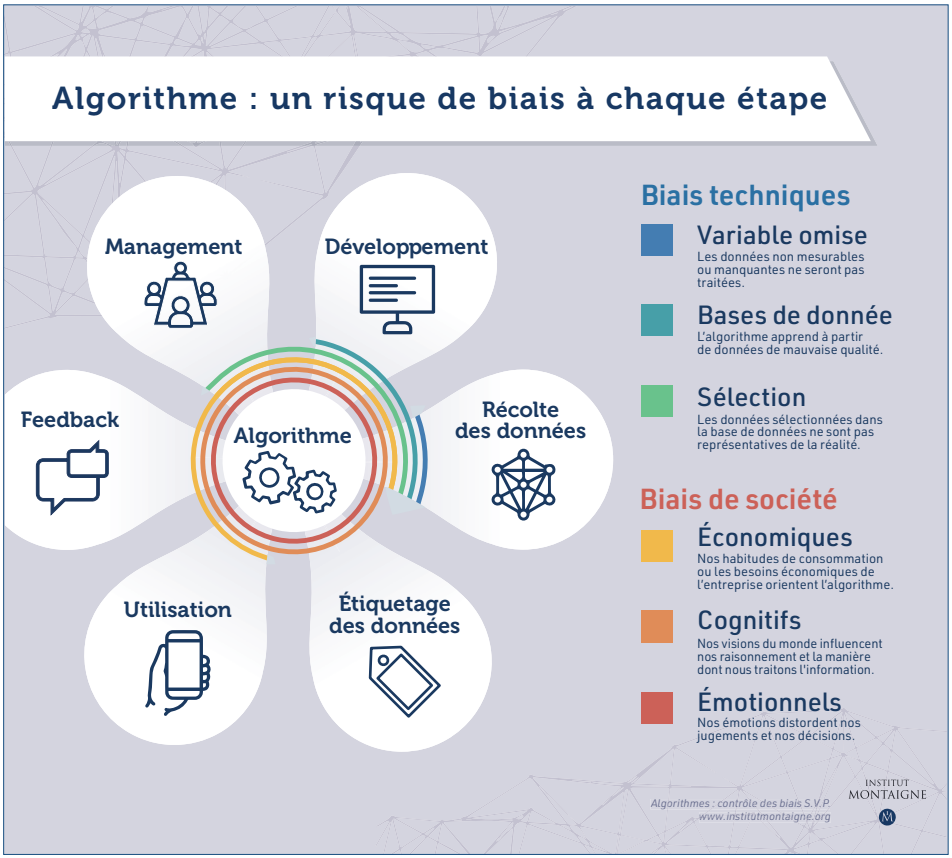
Les connaissances de base sur les algorithmes machine devraient faire partie de l'éducation dans les écoles, mais aussi être accessibles aux adultes tout au long de leur vie. En Finlande, 3,5 % de la population a déjà reçu une formation à l'IA, un cours de formation ayant été conçu localement pour la population.

En France, l'Institut Montaigne et OpenClassrooms, en partenariat avec la Fondation Abeona, se sont associés pour créer Objectif IA, une formation en ligne qui sera disponible au premier trimestre 2020. À travers une série de vidéos, d'exercices interactifs et d'exemples, le cours abordera plusieurs sujets fondamentaux, à la fois techniques et sociétaux, pour permettre à chaque personne de saisir les opportunités et les défis de l'IA et de développer une pensée critique qui les aidera à comprendre et à naviguer dans ce nouveau paysage technologique. Le cours sera disponible gratuitement en ligne, mais aussi distribué par de nombreuses entreprises à leurs collaborateurs.

www.institutmontaigne.org

Nous recommandons le développement de ce type d'initiatives à destination des citoyens et décideurs publics pour améliorer leur connaissance de ces sujets.

Une autre initiative en cours est nommée « 1 scientifique 1 classe : chiche ! ». Le ministère de l'Éducation nationale a ainsi signé une convention avec l'Inria afin d'initier la phase pilote de ce projet. Ce programme s'inscrit dans le nouvel enseignement « Sciences numériques et technologie » et permet aux élèves de seconde de rencontrer des chercheurs et des chercheuses en sciences du numérique. Ceux-ci viennent offrir aux élèves une meilleure compréhension d'un monde totalement transformé par le numérique, permettant ainsi de nourrir l'intérêt et d'encourager des vocations, notamment chez les jeunes femmes. Ce type d'initiatives, dont nous recommandons vivement le développement, devrait intégrer le thème des algorithmes d'IA et de leur impact sociétal.



C. Donner à chaque organisation les moyens de détecter et combattre les biais de ses algorithmes

• Proposition 3 : tester les algorithmes avant utilisation en s'inspirant des études cliniques des médicaments

De nombreuses voix se sont élevées récemment pour promouvoir une intelligence artificielle qui soit explicable. C'est notamment le cas du rapport Villani, qui a défini en 2018 la stratégie française en matière d'intelligence artificielle[47].

Nous pensons néanmoins qu'il est difficile d'expliquer simplement tous les algorithmes. Nous défendons plutôt la nécessité de tester les algorithmes, pour d'une part s'assurer qu'ils font bien ce pourquoi ils ont été développés, et d'autre part détecter la présence de biais. Ces tests se feraient à l'instar des études cliniques dans le domaine des médicaments où l'entreprise pharmaceutique vérifie l'absence de nocivité et l'efficacité du médicament sur un échantillon de patients.

Cette approche de test est plus résiliente que l'explicabilité à plusieurs égards. En effet, l'explicabilité des algorithmes souffre de nombreuses limites et il n'est pas certain qu'une IA pleinement explicable soit possible. De plus, la complexité des algorithmes croît de jour en jour. Même un algorithme « explicable » au sens technique peut rester abscons pour la plupart d'entre nous.

Les limites de l'IA explicable

L'explicabilité des algorithmes souffre de nombreuses limites. Bien que désirable, elle est techniquement difficile à obtenir, tant elle est contraire au principe même de l'apprentissage machine. L'IA optimise de façon aveugle, sans permettre à son utilisateur de donner un sens à l'enchaînement de corrélations calculées. Pire, il y a, avec les techniques actuelles, une réelle opposition entre explicabilité et performance, réduisant les perspectives d'une explicabilité systématique des algorithmes.

.../...

47 Mission parlementaire du 8 septembre 2017 au 8 mars 2018, Rapport Cédric Villani Donner un sens à l'intelligence artificielle.

Par ailleurs, expliquer un algorithme ne répond pas réellement aux besoins en matière de biais. Un algorithme contenant 50 000 règles toutes explicables restera incompréhensible au commun des mortels. De plus, il est difficile de mobiliser les utilisateurs sur ces sujets. Comme beaucoup d'entre nous ceux-ci risquent de ne pas s'y intéresser et de valider sans lire les notices détaillant le fonctionnement des algorithmes. Ils ne chercheront pas à comprendre l'algorithme mais à se rassurer sur son fonctionnement équitable. Qui cherche à savoir comment fonctionne un avion ? Nous sommes bien plus rassurés par le fait de savoir qu'il a passé les tests de sécurité.

Enfin, l'explicabilité risque de se heurter au concept du secret des affaires, qui protégera dans une certaine mesure le contenu des algorithmes (variables, poids, etc.), toujours plus stratégiques.

Sans explicabilité, il est néanmoins possible de s'assurer que les algorithmes présentent bien certaines propriétés. Des tests sont bien plus en mesure de nous rassurer sur ce qui importe vraiment : un traitement équitable des personnes.

Ces tests doivent être autant que possible réalisés par les entités déployant les algorithmes, qu'elles soient privées ou publiques. Elles sont les mieux placées en matière de compétences et de moyens pour les effectuer. Elles y ont aussi un intérêt, pour limiter les risques juridiques et réputationnels, même si de tels tests représentent un coût et des difficultés techniques importantes.

Conduire de tels tests nécessite de définir ce que l'entreprise ou l'administration utilisant l'algorithme considère comme équitable. Sauf cas rares, ce cadre ne sera probablement pas gravé dans le marbre par l'État, tant il dépend de la définition et des circonstances (secteur d'application, criticité de l'algorithme, époque, etc.). Certaines sociétés considèrent que les taux de faux positifs sur des groupes doivent être identiques avec une marge de 5 %. Dans le recrutement, des entreprises privilégient une parité stricte ou se donnent des marges de manœuvre (pas moins de 40 % de femmes). D'autres privilégient une parité en ligne avec le secteur, avec la composition des diplômés. Afin d'aider à déterminer ces seuils et les méthodologies de test adéquates, un accompagnement de la CNIL en amont et pendant le processus de test, à l'instar de ce qui est fait dans le cadre de la protection des données personnelles, serait bénéfique.

Certains biais sont volontaires, acceptables et sont le fruit de stratégies commerciales, d'autres non. C'est donc *in fine* à l'entreprise de se positionner sur ce qu'elle

considère comme la bonne définition d'un algorithme équitable. Elle en porte ensuite le risque juridique et réputationnel, mais refuser de se confronter à ces questions ne ferait qu'augmenter le risque.

• Proposition 4 : adopter une démarche d'équité active autorisant l'usage de variables sensibles dans le strict but de mesurer les biais et d'évaluer les algorithmes

À l'inverse de l'Allemagne, où l'unité de la nation allemande a fini par pousser les États allemands à s'unir, on dit qu'en France c'est l'État qui a construit la Nation, en ignorant et en effaçant les différences de dialecte, de religion, de croyance, de coutumes avec plus ou moins de succès. Dans ce mouvement la France a parfois préféré faire la promotion d'un certain universalisme en masquant sa diversité pour tenter de l'oublier. Les statistiques ethniques sont, à de rares exceptions près, interdites, et les données relatives aux 25 critères de discrimination sont extrêmement difficiles à collecter. Or, comment tester la présence de biais sur ces 25 critères si ces données ne sont pas disponibles ?

Cette disposition pose aujourd'hui des difficultés pour mesurer les discriminations. François Héran, sociologue et démographe au Collège de France, écrit dans la préface de l'enquête statistique *Trajectoires et Origines*[48] *:* « Comprendra-t-on encore [dans dix ans] que l'on ait pu soupçonner certaines de ces questions sur les origines ou les apparences de vouloir « saper les fondements de la République », alors qu'elles visaient modestement à saisir au plus près le mécanisme des discriminations qui mine le principe d'égalité ? »

Exclure les variables sensibles d'un algorithme est donc insuffisant. Dans le domaine des algorithmes publicitaires, il n'est pas nécessaire de connaître le genre d'un acheteur de chemises pour hommes ou de maillots de bains pour femme pour le deviner avec une faible marge d'erreur. L'équité par ignorance ne suffit pas car les algorithmes peuvent discriminer sur des critères auxquels ils n'ont pas accès de manière directe. Comme l'expliquent S. Corbett-Davies et S. Goel, chercheurs à l'Université de Stanford, dans un papier de recherche sur l'équité des algorithmes[49], l'exclusion des variables sensibles ou anti-classification ne répond pas efficacement aux objectifs d'algorithmes équitables. Bien souvent, une approche par calibration, c'est à dire une approche par laquelle on s'assure que les résultats obtenus sont effectivement indépendants des variables protégées, est préférable.

48 Enquête réalisée en 2009 et collectant notamment des informations sur l'origine géographique, les nationalités antérieures à la nationalité française ou encore le sentiment d'appartenance dans la population française.
49 Corbett-Davies S., Goel S., *The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning*, Stanford University, Juillet 2018.

Nous recommandons d'abandonner l'approche d'équité par ignorance et d'adopter une stratégie d'équité active. Il s'agit d'accepter de mesurer les discriminations, de tester la présence de biais, grâce à la collecte d'information sur les 25 critères protégés.

Bien plus que l'exclusion de variables sensibles dans un algorithme, c'est l'indépendance du résultat par rapport à des variables protégées qui garantit un algorithme équitable. C'est par cette approche que nous serons collectivement en mesure d'identifier et réduire les biais des algorithmes, et de la société en général.

Néanmoins, cette approche doit être strictement encadrée. Aujourd'hui, il existe déjà des cas d'exceptions pour la collecte de ces données sensibles, notamment dans le cadre d'usages statistiques. Nous proposons d'étendre ces exceptions au cas très spécifique de tests d'identification de discriminations éventuelles, en permettant une collecte réduite - sur un échantillon des données - de ces variables sensibles.

Pour les algorithme dont la nature le justifie, le développeur pourra alors vérifier que les résultats sont indépendants des variables protégées. Une telle collecte nécessiterait bien entendu le consentement de l'utilisateur. Celui-ci pourrait accepter de partager ces informations pour contribuer à constituer une base de test, permettant de vérifier l'absence de biais. La collecte nécessiterait par ailleurs la réalisation préalable d'une analyse d'impact et sa transmission à la CNIL.

Par mesure de précaution, la collecte serait limitée à un échantillon restreint d'utilisateurs, dont la définition reste à établir. Cela pourrait concerner , par exemple une fraction d'entre eux (20% des utilisateurs par exemple), ou un échantillon suffisamment large pour garantir une représentativité statistique (en général quelques milliers d'utilisateurs). Cette limite permettrait d'éviter une dérive généralisée tout en garantissant une collecte suffisamment représentative. Dans certains cas, il restera difficile d'acquérir la confiance des individus et de les convaincre que ces données ne seront utilisées qu'à des fins de test. Par exemple dans le cas d'un recrutement : un candidat pourrait être méfiant face à un recruteur qui voudrait collecter des variables protégées, malgré l'assurance que ce ne serait utilisé qu'à des fins de test. La collecte de données pourrait donc également être effectuée par des tierces parties pour apporter des garanties contre l'utilisation à d'autres fins.

Une approche d'équité active ne veut cependant pas dire que la France s'engagerait dans une politique de quotas pour chacun des critères de discrimination. Ce n'est pas aux algorithmes de définir le bon équilibre entre chaque groupe. Le risque de cristalliser la société et de la réduire à une compétition entre différents groupes serait trop grand. Il s'agit plutôt de donner les moyens aux acteurs publics et privés de détecter des écarts et discriminations manifestes.

• Proposition 5 : mettre à disposition des bases de données de test publiques pour permettre aux entreprises d'évaluer les biais de leur méthodologie

Une approche d'équité active nécessite une base de données comportant des variables normalement protégées au titre des 25 critères de discrimination. Dans certains cas il ne sera pas possible pour les entreprises de constituer ces bases de données de test : parce que les individus n'accepteront pas de voir leurs données collectées malgré les garde-fous (analyse d'impact, collecte sur un échantillon restreint, finalité de test uniquement), parce que l'on ne souhaite pas que les entreprises collectent ce type d'informations. Dans ces cas-là, l'État ou une autorité indépendante pourrait prendre en charge la constitution de ces bases de données.

Un cas d'usage mentionné par de nombreux spécialistes concerne la reconnaissance faciale. Les développeurs ne disposent pas d'une base de données représentative de la population française pour tester l'absence de biais de leur algorithme. Il n'y a pas de moyen de vérifier que les algorithmes fonctionneront correctement sur toute la population résidant en France. Cela vaut pour des développeurs français, mais également pour ceux qui importent en France des algorithmes entraînés aux États-Unis ou en Chine.

Nous recommandons de mettre à disposition des bases de données publiques avec des informations sur certains des 25 critères protégés pour des cas d'usages précis. Cela doit être limité à des cas spécifiques comme la reconnaissance faciale (genre, etc.), ou l'évaluation de risque de crédit par les banques (historique de revenus, genre, classe socio-professionnelle).

De telles bases de données seraient utilisées pour tester la méthodologie, avant ou après entraînement sur les données propres. Elles permettront de vérifier qu'une variable protégée n'a pas été « redécouverte » par l'algorithme comme on aurait pu le craindre pour l'assurance automobile[50].

Aux États-Unis, le travail du NIST (National Institute of Standards and Technology) a établi une référence. Il met à disposition des bases de données pour évaluer les technologies de biométrie tant en matière de performance que d'absence de biais. Le gouvernement américain a octroyé à cet institut public un accès à de grandes quantités de données confidentielles. Sa rigueur scientifique est reconnue tant dans les métriques que dans les protocoles. En France, le Laboratoire national de métrologie

50 Un algorithme d'assurance automobile qui n'a pas accès aux données de genre pourrait délivrer des primes plus élevées pour les voitures rouges. Il s'agit en effet d'une bonne mesure approximative du genre, puisque les véhicules de cette couleur sont plus souvent la propriété d'hommes (risque d'accident est plus élevé).

et d'essais (LNE) pourrait endosser cette mission. Le LNE, qui dépend du ministère de l'Economie, a en effet pour mission d'assister les pouvoirs publics et les acteurs économiques dans l'élaboration de méthodes d'essais et intervient déjà dans des domaines d'application de l'IA tels que les véhicules autonomes, les dispositifs médicaux intelligents (notamment d'aide au diagnostic), les robots industriels collaboratifs ou les drones. Il est responsable d'un défi destiné à évaluer des robots agricoles, et a déjà noué des partenariats pour développer des méthodes d'évaluation des algorithmes de transcription automatique de la parole.

L'objectif n'est pas de publier des bases de données contenant des informations très confidentielles sur des milliers de Français. D'autant plus que l'anonymisation complète est de plus en plus vue comme impossible. L'accès à ces bases pourra être fortement contrôlé, comme l'est l'accès aux bases de données statistiques de l'INSEE. La piste de bases de données « synthétiques » doit également être explorée : elle permettrait de bruiter les données, c'est à dire de les modifier pour publier des bases de données ne comportant aucun profil réel, tout en gardant les propriétés statistiques initiales intactes.

Être la base de données de référence implique de grandes responsabilités, comme l'a montré l'exemple d'ImageNet. Il sera impératif de scruter ces bases de données, et d'avoir la plus haute exigence à leur égard.

D. Évaluer les algorithmes à fort impact pour limiter leurs risques

Nous pensons nécessaire de ne pas considérer tous les algorithmes de la même manière. De même qu'il n'existe pas de régulation pour toutes les machines (seules les plus dangereuses sont encadrées), ce sont les algorithmes à fort impact qui doivent attirer l'attention des pouvoirs publics, des entreprises et de la société civile.

Une focalisation de l'attention sur les algorithmes à fort impact est également l'approche suggérée par la Commission Européenne dans son projet de livre blanc sur l'intelligence artificielle dévoilé en janvier 2020[51]. Elle suggère en effet que les régulateurs et développeurs d'algorithmes devraient concentrer leurs efforts sur les algorithmes à impact, ou *high risk*.

51 Projet de livre blanc sur l'intelligence artificielle de la Commission Européenne.

Comment définir un fort impact ? La Commission Européenne mentionne deux critères cumulatifs : un algorithme doit (I) concerner un secteur défini comme à risque (santé, transport, police…) et (II) présenter des effets juridiques ou un risque d'atteinte physique ou matérielle pour la personne ou l'entité visée. Au-delà de cette définition, qui se limite aux impacts juridiques et matériels, trois marqueurs sont importants à nos yeux : le déni d'accès à des services essentiels[52], l'atteinte à la sécurité des personnes[53], la restriction des droits fondamentaux[54]. Les algorithmes dont les décisions génèrent un ou plusieurs de ces types d'impact peuvent être qualifiés de critiques.

Exemples d'algorithmes à fort impact

Dans l'automobile, certains algorithmes, anodins, sont employés pour ajuster la température du véhicule ou l'inclinaison du siège conducteur. À l'inverse, d'autres algorithmes définissent le comportement d'un véhicule autonome. Un biais dans le second aurait un très fort impact, à la différence du premier. Les industriels et régulateurs du secteur ne s'y sont pas trompés. La norme ISO26262 sur les systèmes critiques, définie par l'Organisation internationale de normalisation, encadre spécifiquement les algorithmes qui ont un impact sur la sécurité des passagers. Pas les autres. Cette norme fait d'ailleurs l'objet de travaux de modernisation pour l'adapter à l'intelligence artificielle. Dans ce domaine, les algorithmes à fort impact sont avant tout ceux qui touchent à la sécurité des passagers et passants.

En matière bancaire, certains algorithmes vont réaliser du ciblage publicitaire tandis que d'autres vont évaluer pour chaque personne son risque et donc son éligibilité à un crédit bancaire. Ces algorithmes sont naturellement optimisés pour réduire le risque global du portefeuille. Une amélioration du code sera appliquée dès lors qu'elle améliorera le profil de risque global, même si cela implique une moindre précision sur certaines personnes. Les conséquences sur la vie de ces individus sont potentiellement importantes puisque l'on peut alors réduire leur accès à des services essentiels.

…/…

52 Domiciliation bancaire, sécurité sociale, emploi, etc.
53 Atteinte physique comme morale
54 Droit de libre circulation, droit à manifester, etc.

Dans le secteur de la santé, des algorithmes auront pour mission de détecter une fracture bénigne sur votre radio, d'autres de déterminer à partir de vos données biologiques si vous êtes atteints d'une récidive de cancer. L'un des deux engage le pronostic vital du patient. Un biais significatif serait, à ce titre, bien plus critique. C'est d'ailleurs la distinction qu'a opérée l'autorité américaine de contrôle des médicaments (Federal Drug Administration). Dans son discussion paper sur le cadre réglementaire à appliquer aux dispositifs médicaux constitués de logiciels auto-apprenants (SaMD), elle différencie les algorithmes en fonction de leur risque (criticité pour le patient et impact de l'algorithme sur la décision clinique). Si l'auto-apprentissage modifie le risque pour le patient, alors le dispositif devra être capable de prouver qu'il est toujours aussi sûr. Autrement, il aura la possibilité d'évoluer à sa guise. La grille d'analyse proposée par la Haute Autorité de Santé pour évaluer les dispositifs médicaux embarquant de l'intelligence artificielle intègre elle aussi une question sur l'impact que pourraient avoir des biais dans la collecte de données.

En matière de police, le déploiement d'algorithmes de reconnaissance faciale risque, en présence de biais, de restreindre les droits fondamentaux de groupes de citoyens. Le droit de manifester, mais aussi la présomption d'innocence seraient concernés et limités par rapport à d'autres citoyens. À l'inverse, les algorithmes permettant de traduire de manière automatisée les langues étrangères sont a priori moins menaçants, en cas de biais avérés.

Nous recommandons à chaque acteur développant des algorithmes de mettre en place des exigences spécifiques pour ces algorithmes à fort impact. L'État aura également un rôle en stimulant le développement de labels et de certifications pour ces algorithmes afin de renforcer la confiance de tous envers leur fonctionnement, et en développant une capacité propre pour les auditer en dernier recours.

a. Proposition 6 : être plus exigeant pour les algorithmes à fort impact

La nécessité de concentrer les réponses aux biais algorithmiques sur les logiciels ayant le plus d'impact nous amène à leur définir un cadre particulier. Lorsqu'un algorithme présente un risque d'impact important, deux mesures deviennent essentielles : la transparence et le droit de recours. Nous recommandons que les acteurs publics et privés mettant en œuvre des algorithmes à fort impact rendent ces deux mesures effectives.

La transparence consiste à dévoiler aux utilisateurs des algorithmes ou à un tiers de confiance, des informations sur son but, sa méthodologie et le type de données qu'il traite. Le besoin de transparence est proportionnel à l'impact : les exigences sur la nourriture ont ainsi grandi au fur et à mesure que le contenu et la provenance de nos aliments nous sont apparus comme essentiels à notre bonne santé.

L'achat d'un aliment en grande surface s'accompagne de multiples informations sur la date de péremption, le contenu calorifique ou encore les composants du produit. Ces informations sont nécessaires pour rassurer le consommateur sur la qualité de ce qu'il mange. Même s'il ne lit pas les informations, leur présence rassure, car elle garantit que le fabricant n'a rien à cacher.

Le succès d'initiatives comme Nutriscore ou Yuka qui permettent d'évaluer en un clic la qualité nutritive des aliments que l'on achète dans un magasin montre l'appétit des Français pour une information claire, simple et quantifiée sur la qualité des produits achetés.

Dans le domaine de la santé, la loi bioéthique en cours d'examen impose d'accompagner les algorithmes réalisant un diagnostic ou un traitement médical d'une notice explicative. Elle ne fait ici que reprendre un principe de bon sens : plus l'objet a un impact fort, plus je veux avoir des garanties sur son fonctionnement.

Nous pensons nécessaire de multiplier ce type d'initiatives pour les biais à fort impact, en concentrant les efforts de transparence sur la nature et la qualité des données utilisées et sur les objectifs des algorithmes. Les biais viennent avant tout (cf. partie I) des données et (cf. partie II) du choix des objectifs que l'algorithme doit optimiser (cf. partie II). C'est donc sur ces éléments qu'il faut se concentrer.

En reprenant l'exemple de Yuka, notre proposition revient à proposer la publication de deux types d'information. Tout d'abord, une liste du type de données utilisées pour entraîner l'algorithme d'évaluation des qualités nutritives des aliments. Cette liste pourra être complétée d'un indicateur de la qualité de ces données (données homogènes, représentatives, échantillon assez vaste, etc.), ainsi que d'informations sur les conditions d'acquisition et de gestion. En effet, Il faut s'attaquer à la qualité des données sous-jacentes (les ingrédients) plutôt qu'au code de l'algorithme (la recette), plus visible mais souvent moins générateur de biais.

Ensuite, une publication des objectifs que l'algorithme doit atteindre est nécessaire. Celui-ci est-il entraîné pour minimiser la valeur calorifique des aliments recommandés, mettre en avant certain types de régimes alimentaires, par exemple sans sel, ou valoriser certain types d'aliments, par exemple bio? La nature des objectifs est un

élément essentiel pour comprendre où l'algorithme veut nous emmener et identifier des biais éventuels.

Dans cet exemple, la transparence est vis-à-vis du consommateur ou de la consommatrice. Néanmoins, une transparence vis-à-vis d'un tiers de confiance, organisme public, laboratoire de recherche ou autre est tout à fait envisageable et permettrait de protéger le secret professionnel qui entoure ces informations stratégiques.

La deuxième mesure de ce cadre spécifique concerne le droit de recours. Qui dit décision importante, prise ou influencée par l'algorithme, dit droit de contester celle-ci. Ce principe, sanctuarisé dans le droit français et européen pour les décisions automatisées, publiques ou à fort impact, est un élément absolument essentiel pour le développement d'algorithmes équitables. Une critique légitime de Parcoursup est notamment le fait de n'avoir pas ouvert clairement de voie de recours interne au processus. Cette seconde mesure est de nature à renforcer la confiance dans de tels algorithmes.

Le droit de recours face aux algorithmes pourrait s'appuyer largement sur le droit de recours tel qu'il existe aujourd'hui, et notamment sur l'institution du Défenseur des Droits. Lorsqu'un salarié soupçonne par exemple qu'une décision discriminante a été prise à son encontre, il peut saisir le Défenseur des Droits. Ce n'est alors plus à lui de prouver qu'il y a eu discrimination, mais bien à l'entreprise attaquée de prouver qu'il n'y en a pas eu. Un droit de recours où la charge de la preuve repose sur le développeur de l'algorithme doit nécessairement aller de pair avec l'autorisation de collecter certaines variables protégées à des fins de test (cf. proposition 4 sur l'équité active).

Un tel droit n'est pas toujours possible. Si l'on évoque les algorithmes de conduite autonome, une procédure de recours interne ne sera que de peu d'utilités en cas d'accident. En revanche, une redondance d'algorithmes, c'est à dire la présence d'un second algorithme qui viendrait analyser la première décision et éventuellement la contredire est une solution qu'il faut promouvoir, notamment lorsque la sécurité des personnes est en jeu. Ainsi, deux algorithmes de conduite autonome pourraient, lorsqu'ils donnent des ordres différents, alerter le conducteur sur une limite de l'algorithme.

Ce cadre spécifique ne nécessite pas une grande loi sur les biais algorithmiques commune à tous les secteurs d'activités. Nous pensons que l'adoption de bonnes pratiques par les entreprises et administrations concernées, l'usage de dispositifs existants comme le Défenseur des droits et l'ajout au cas par cas de dispositions dans des législations sectorielles sont suffisants.

b. Proposition 7 : soutenir l'émergence de labels pour renforcer la confiance du citoyen dans les usages critiques et accélérer la diffusion des algorithmes bénéfiques

L'engagement de transparence vis-à-vis de consommateurs ou vis-à-vis d'entreprises clientes peut être difficile à mettre en œuvre. En effet, des informations telles que les objectifs des algorithmes, les variables utilisées en entrée, ou les poids entre les différents variables, ont un fort caractère stratégique. Cela est particulièrement vrai dans des domaines comme la banque, l'assurance, ou la conduite autonome où les performances des algorithmes vont durablement différentier les concurrents.

Par ailleurs, la transparence est fortement chronophage pour la personne qui doit analyser les éléments fournis et se forger une opinion. Le consommateur lisant les notices d'achats électroménagers et la PME analysant les spécifications d'une machine industrielle achetée devront investir un temps considérable pour se rassurer sur la qualité et le fonctionnement de leur acquisition.

C'est pourquoi l'industrie a depuis longtemps développé des labels de qualité et de sécurité qui garantissent aux acheteurs particuliers ou entreprises une conformité avec un certain niveau d'exigence. Pour les usages les plus critiques, il s'agit de certifications obligatoires, reposant sur la conformité de l'objet avec des cas de sécurité, ou *safety cases*[55]. C'est notamment la norme dans l'automobile, l'aviation ou encore la santé. Pour des usages moins critiques, il peut s'agir simplement de labels (facultatifs), qui donnent une indication à l'acheteur sur les caractéristiques du produit (label agriculture biologique AB par exemple).

Nous recommandons de transposer cette logique de qualité industrielle aux algorithmes en favorisant l'émergence de labels et certifications spécifiques. Le processus de définition d'une certification étant lent et contraignant par rapport au rythme d'évolution de la technologie, c'est l'émergence de labels qu'il faudra prioriser à court terme.

Face à la difficulté de définir des normes pour les biais, il semble illusoire de définir un label garantissant un algorithme sans biais. Néanmoins, des labels pourraient se concentrer sur l'auditabilité des algorithmes, sur la qualité des données ou encore sur la présence d'un processus d'évaluation des risques de biais au sein de l'entreprise.

55 Définition d'une situation face à laquelle l'objet doit répondre à un certain nombre de critères de sécurité. Par exemple, les *crash tests* dans l'automobile ou les études cliniques pour les médicaments.

Le label Fair data use[56] est un exemple du type d'initiatives qu'il faut encourager dans chaque secteur, particulièrement pour les algorithmes à fort impact. Ce label s'obtient après un audit des algorithmes pour garantir l'absence de discriminations, le respect du RGPD et des règles RSE de l'entreprise. Ce label, délivré pour une durée d'un an, consiste en une évaluation de l'algorithme cible par un algorithme auditeur qui analyse des critères précis comme la présence de variables sensibles, la transparence ou encore la loyauté du traitement.

Le développement de labels permettrait également de développer au sein de l'écosystème français des compétences et des méthodes d'audit des algorithmes. À titre d'exemple, la méthode d'augmentation de données permet de tester de façon très robuste l'indépendance de l'algorithme vis-à-vis de certaines variables. Elle consiste à générer des données artificielles, et par exemple de générer des individus identiques en tout point à ceux réellement présents dans la base de données, sauf en ce qui concerne leur genre. On peut alors vérifier qu'à profil égal, une femme et un homme sont bien traités de la même façon par l'algorithme. D'autres méthodes permettent d'inférer des variables exclues explicitement de la base de données, mais qui peuvent pourtant être capturées par l'algorithme. Des spécialistes de ces méthodes formeraient les premières briques d'une industrie du test de l'algorithme.

Une difficulté que rencontreront les labels et certifications lors de leur extension au domaine des biais algorithmiques est la fréquence d'évolution des codes sources et des données d'entraînement. Des labels portant sur l'équipe développant l'algorithme (ses processus internes, ses pratiques, sa composition) plutôt que sur les algorithmes eux-mêmes pourront permettre de pallier cette difficulté.

c. Proposition 8 : développer une capacité d'audit des algorithmes à fort impact

Dans le secteur de la justice, les algorithmes qui cherchent à prédire la récidive, et donc la libération sous caution, ne sont pas en eux-mêmes moins performants que les juges. Ces derniers font aussi des choix qui peuvent être biaisés, et sont de même soumis à des injonctions contradictoires, sur l'égalité entre des groupes (il faut se tromper autant de fois pour les jeunes que pour les plus âgés), et l'égalité entre des individus (il faut que je traite deux profils identiques indépendamment de leur couleur de peau). Mais l'algorithme génère un problème nouveau quand il est développé par une société privée, dont l'algorithme est protégé par le secret des

56 Site internet du label *"Fair data use"* par Maathics.

affaires, les données d'entraînement ne sont pas disponibles, et qui ne peut être audité. Sans aucun recours sur la décision qui a été prise, la justice perd de son poids.

Nous sommes convaincus que nos recommandations de formation et de bonnes pratiques favoriseront le développement d'algorithmes moins biaisés. L'émergence de labels permettra de diffuser ces bonnes pratiques, et de valoriser les développeurs vertueux d'algorithmes à fort impact. Dans les cas où des biais apparaîtraient néanmoins, et avec un fort impact, des possibilités de recours donneront la possibilité aux individus ou à la société civile de ne pas subir le biais sans capacité de réagir.

Il demeure que dans certains cas extrêmement problématiques ou sensibles, il faudra pouvoir auditer des algorithmes, leurs données et leur fonctionnement. Ces audits d'algorithmes à fort impact pourraient être effectués par des tierces parties, sur l'exemple des auditeurs des comptes financiers des entreprises, ou par l'État.

Les cas de Parcoursup, de la reconnaissance faciale par la police, des algorithmes de prédiction dans la justice américaine sont de multiples preuves que l'État concentre beaucoup des utilisations à fort impact. Or sa capacité d'expertise et d'audit, en tant que régulateur et acheteur, est aujourd'hui à la fois limitée et éclatée. Sept autorités et services de l'État sont potentiellement compétents pour l'audiovisuel et la communication[57]. La création d'un pôle d'expertise capable d'auditer les algorithmes, initiée en 2019, est donc une bonne nouvelle.

57 Autorité de la concurrence, CSA, CNIL, ARCEP, DGE, DGCCRF, Direction générale des médias et des industries culturelles.

L'État se dote d'un pôle d'expertise algorithmique

L'idée d'une fusion de plusieurs de ces régulateurs pour créer un régulateur du numérique est souvent avancée : CSA et CNIL, CNIL et ARCEP, CSA et ARCEP. Sans vouloir nous prononcer à ce sujet, nous notons que si l'audiovisuel et la communication sont en première ligne, ils ne sont pas les seuls à être transformés par les algorithmes.

La transversalité des acteurs du numérique, et l'omniprésence à venir des algorithmes, requiert de créer une capacité d'expertise numérique au sein de l'État, mobilisée sur différentes missions. La diversité et l'importance des projets permettraient de faciliter le recrutement de profils techniques aujourd'hui très recherchés.

Le projet de loi relatif à la communication audiovisuelle et à la souveraineté culturelle à l'ère du numérique, déposé en décembre 2019, prévoit la création d'un « pôle d'expertise numérique » de vingt personnes à la Direction générale des entreprises (ministère de l'Economie). Ce projet répond à un besoin qui dépasse largement le seul domaine audiovisuel.

Les algorithmes évoluent constamment, soit parce qu'ils sont améliorés, soit parce qu'ils apprennent au fur et à mesure de leur utilisation, grâce au feedback. Un audit ponctuel d'un algorithme risque malheureusement d'être rapidement obsolète. Une étude d'AlgoTransparency sur le biais de l'algorithme de YouTube en faveur de contenus radicaux peut être démentie 6 mois plus tard parce que YouTube prend régulièrement en compte les critiques.

Dans les cas où l'impact d'un algorithme est fort, il est souhaitable d'avoir une assurance continue sur l'absence de biais, Un tiers de confiance comme les sociétés d'audit ou l'État pourrait choisir un contrôle proprement numérique[58]. Plutôt que d'établir un audit reporté sur papier une fois par an, l'entité tierce pourrait récupérer via des API sécurisées soit des résultats de test, soit des données permettant d'auditer l'algorithme. Elle pourrait vérifier ainsi de façon continue que ces algorithmes conviennent à un certain nombre de critères, notamment en matière de biais.

58 Grossman N., *Regulation, the internet way*, Data-smart city solutions, ASH Center, Harvard Kennedy School, 8 Avril 2015.

CONCLUSION

Les biais des algorithmes ont fait leur apparition dans le débat américain sur les discriminations à l'ère du numérique. Des cas particulièrement frappants dans le domaine de la justice, du recrutement ou de l'accès aux services financiers ont marqué les opinions et sensibilisé à la fois le grand public, les chercheurs, les entreprises et les pouvoirs publics à ce sujet.

Penser que cette prise de conscience outre-Atlantique nous protège parfaitement des risques associés à ces biais est pourtant insuffisant. La définition de l'équité, du comportement que l'algorithme devrait adopter pour garantir des décisions juste, n'est pas un concept universel. Par ses modes de vie et son histoire politique et juridique, la France peut et doit avoir sa propre doctrine en ce qui concerne les biais algorithmiques.

Cette approche française devra concilier d'une part un certain retard pris par l'Europe en matière de numérique qu'une régulation contre les biais algorithmiques ne doit pas aggraver davantage, et d'autre part un risque bien réel de déstabilisation sociale.

Elle devra également prendre en compte la flexibilité nécessaire dont ont besoin les acteurs de terrain, pour adapter à chaque contexte les objectifs des algorithmes et les exigences d'équité et de performances appliquées aux algorithmes.

Enfin, cette approche devra prendre en compte le formidable potentiel de réduction des discriminations que représentent les algorithmes. L'enjeu est certes de savoir si l'algorithme est biaisé, mais surtout s'il l'est davantage que la personne qu'il remplace ou assiste. Sans nier les risques propres des algorithmes comme le manque de transparence ou la capacité à démultiplier une décision biaisée, les progrès que nous pouvons apercevoir en matière de discriminations dans notre société grâce aux algorithmes sont réels.

Nous sommes convaincus qu'il est aujourd'hui trop tôt pour proposer une loi sur les biais algorithmiques ou un contrôle *ex ante* par l'État des algorithmes. Le cadre juridique en matière de discrimination et de numérique offre déjà de nombreuses solutions et son implémentation mériterait d'être affermie. L'État, ayant dès aujourd'hui de grandes difficultés à faire appliquer dans son entièreté le règlement européen sur la protection des données, serait par ailleurs difficilement en mesure de contrôler des algorithmes si nombreux et dont la complexité est croissante.

Nous pensons au contraire que cette doctrine française en matière de biais algorithmiques devrait, compte tenu de l'ampleur encore modérée du phénomène en

Europe, reposer sur trois piliers. Tout d'abord, un effort essentiel de formation, afin de garantir que l'ensemble des acteurs de la chaîne de valeur des algorithmes prennent conscience des risques liés au déploiement d'algorithmes.

Ensuite, la création d'une capacité, pour les acteurs publics et privés de tester les algorithmes à la recherche d'éventuels biais.

Enfin, il est essentiel de dissocier dès aujourd'hui les algorithmes les plus sensibles des autres et de leur appliquer un cadre spécifique. Ce sont ceux portant atteinte aux droits fondamentaux, mettant en jeu la sécurité physique ou psychologique des personnes, et restreignant l'accès à des services essentiels. Pour ces algorithmes, droits de recours, transparence, labellisation et audit par des tiers sont des étapes qui devront nécessairement être incluses dans leur développement.

L'initiative législative et réglementaire sur les algorithmes et l'intelligence artificielle envisagée par la nouvelle Commission Européenne sera à ce titre, une étape essentielle dans le développement d'une doctrine française et européenne face aux risques des biais des algorithmes.

REMERCIEMENTS

L'Institut Montaigne remercie particulièrement les personnes suivantes pour leur contribution à ce travail.

Présidents du groupe de travail

- **Anne Bouverot,** présidente du conseil d'administration de Technicolor et présidente, Fondation Abeona (co-présidente)
- **Thierry Delaporte,** directeur général adjoint, Capgemini (co-président)

Rapporteurs

- **Arno Amabile,** ingénieur, Corps des Mines (rapporteur)
- **Théophile Lenoir,** responsable du programme Numérique, Institut Montaigne
- **Tanya Perelmuter,** directrice de stratégie et partenariats, Fondation Abeona (rapporteure générale)
- **Basile Thodoroff,** ingénieur, Corps des Mines (rapporteur)

Membres du groupe de travail

- **Gilles Babinet,** conseiller numérique, Institut Montaigne
- **Ingrid Bianchi,** fondatrice / directrice, Diversity Source Manager
- **David Bounie,** directeur du département sciences économiques et sociales, Télécom Paris
- **Dominique Cardon,** directeur, Médialab de Sciences Po
- **Anna Choury,** *Advanced Data Analytics Manager,* Airbus
- **Stephan Clémençon,** enseignant-chercheur, Télécom Paris
- **Dominique Latourelle,** *Head of RTB,* iProspect
- **Sébastien Massart,** directeur de la stratégie, Dassault Systèmes
- **Bernard Ourghanlian,** *Chief Technology Officer and Chief Security Officer,* Microsoft France
- **Guillemette Picard,** *Chief Health Officer,* Nabla
- **Christian de Sainte Marie,** directeur du centre des études avancées, IBM France
- **François Sillion,** *Director,* Advanced Technologies Center Paris, Uber
- **Serge Uzan,** vice-président, Conseil national de l'ordre des médecins

Ainsi que :

- **Joan Elbaz,** assistante chargée d'études, Institut Montaigne
- **Margaux Tellier,** assistante chargée d'études, Institut Montaigne
- **Julie Van Muylders,** assistante chargée d'études, Institut Montaigne

Les personnes auditionnées ou rencontrées dans l'élaboration de ce travail

- **Éric Adrian,** directeur général, UiPath France
- **Prabhat Agarwal,** *Deputy Head of Unit E-Commerce and Platforms,* DG Connect, European Commission
- **Sacha Alanoca,** *Senior AI Policy Researcher & Head of Community Development,* The Future Society
- **Christine Bargain,** directrice RSE de 2011 à 2018, Groupe La Poste
- **Marie Beaurepaire,** cheffe de projets, Afmd
- **Bertrand Braunschweig,** directeur de coordination du programme national de recherche en intelligence artificielle
- **Alexandre Briot,** *Artificial Intelligence Team Leader,* Valeo
- **Clément Calauzènes,** *Senior Staff Research Lead,* Criteo AI Lab
- **Laurent Cervoni,** directeur intelligence artificielle, Talan
- **Florence Chafiol,** avocate associée, August Debouzy
- **Guillaume Chaslot,** *Mozilla Fellow* et fondateur, Algotransparency
- **Raja Chatila,** professeur d'intelligence artificielle, de robotique et d'éthique et membre du groupe d'experts de haut niveau sur l'intelligence artificielle, Commission européenne
- **Bertrand Cocagne,** directeur innovation & technologies Lending & Leasing, Linedata Services
- **Guillaume De Saint Marc,** *Senior Director, Chief Technology and Architecture Office,* Cisco
- **Marie-Laure Denis,** présidente, CNIL
- **Christel Fiorina,** coordonnatrice du volet économique de la stratégie nationale IA, Service de l'économie numérique de la Direction générale des Entreprises, Ministère de l'Économie et des Finances
- **Marie-Anne Frison-Roche,** professeure, Sciences Po
- **Vincent Grari,** *Research Data Scientist,* AXA
- **Arthur Guillon,** *Senior Machine Learning Engineer,* easyRECrue
- **Nicolas Kanhonou,** directeur, promotion de l'égalité et de l'accès aux droits, Défenseur des droits
- **Djamil Kemal,** *co-CEO,* Goshaba
- **Yann Le Biannic,** *Data Science Chief Expert*, SAP
- **Agnès Malgouyres,** responsable intelligence artificielle, Siemens Healthineers France
- **Stéphane Mallat,** professeur, Collège de France
- **Sébastien Mamessier,** *Senior Research Engineer,* Uber
- **Claire Mathieu,** directrice de recherche, CNRS
- **Marc Mézard,** directeur, ENS
- **Nicolas Miailhe,** co-fondateur et président, The Future Society

- **Christophe Montagnon,** directeur de l'organisation, des systèmes d'information et de la qualité, Randstad France
- **Christelle Moreux,** responsable juridique, Siemens Healthcare
- **François Nédey,** responsable de l'unité Technique et Produits et membre du comité exécutif, Allianz
- **Bertrand Pailhès,** coordonnateur de la stratégie française en intelligence artificielle jusqu'en novembre 2019 et directeur des technologies et de l'innovation, CNIL
- **Cédric Puel,** *Head of Data and Analytics,* BNP Paribas Retail Banking and Services
- **Pete Rai,** *Principal Engineer in the Chief Technology and Architecture Office,* Cisco
- **Boris Ruf,** *Research Data Scientist,* AXA
- **Bruno Sportisse,** président-directeur général, Inria
- **Pierre Vaysse,** directeur technique particuliers, pricing, data & pilotage, Allianz
- **Renaud Vedel,** coordonnateur ministériel en matière d'intelligence artificielle, Ministère de l'Intérieur
- **Fernanda Viégas,** *Co-lead,* PAIR Initiative, Google

Les opinions exprimées dans cette note n'engagent ni les personnes précédemment citées ni les institutions qu'elles représentent.

LES PUBLICATIONS DE L'INSTITUT MONTAIGNE

- Retraites : pour un régime équilibré (mars 2020)
- Espace : le réveil de l'Europe ? (février 2020)
- Données personnelles : comment gagner la bataille ? (décembre 2019)
- Transition énergétique : faisons jouer nos réseaux (décembre 2019)
- Religion au travail : croire au dialogue - Baromètre du Fait Religieux Entreprise 2019 (novembre 2019)
- Taxes de production : préservons les entreprises dans les territoires (octobre 2019)
- Médicaments innovants : prévenir pour mieux guérir (septembre 2019)
- Rénovation énergétique : chantier accessible à tous (juillet 2019)
- Agir pour la parité : performance à la clé (juillet 2019)
- Pour réussir la transition énergétique (juin 2019)
- Europe-Afrique : partenaires particuliers (juin 2019)
- Media polarization « à la française » ? Comparing the French and American ecosystems (mai 2019)
- L'Europe et la 5G : le cas Huawei (partie 2, mai 2019)
- L'Europe et la 5G : passons la cinquième ! (partie 1, mai 2019)
- Système de santé : soyez consultés ! (avril 2019)
- Travailleurs des plateformes : liberté oui, protection aussi (avril 2019)
- Action publique : pourquoi faire compliqué quand on peut faire simple (mars 2019)
- La France en morceaux : baromètre des Territoires 2019 (février 2019)
- Énergie solaire en Afrique : un avenir rayonnant ? (février 2019)
- IA et emploi en santé : quoi de neuf docteur ? (janvier 2019)
- Cybermenace : avis de tempête (novembre 2018)
- Partenariat franco-britannique de défense et de sécurité : améliorer notre coopération (novembre 2018)
- Sauver le droit d'asile (octobre 2018)
- Industrie du futur, prêts, partez ! (septembre 2018)
- La fabrique de l'islamisme (septembre 2018)
- Protection sociale : une mise à jour vitale (mars 2018)
- Innovation en santé : soignons nos talents (mars 2018)
- Travail en prison : préparer (vraiment) l'après (février 2018)
- ETI : taille intermédiaire, gros potentiel (janvier 2018)
- Réforme de la formation professionnelle : allons jusqu'au bout ! (janvier 2018)
- Espace : l'Europe contre-attaque ? (décembre 2017)
- Justice : faites entrer le numérique (novembre 2017)

- Apprentissage : les trois clés d'une véritable transformation (octobre 2017)
- Prêts pour l'Afrique d'aujourd'hui ? (septembre 2017)
- Nouveau monde arabe, nouvelle « politique arabe » pour la France (août 2017)
- Enseignement supérieur et numérique : connectez-vous ! (juin 2017)
- Syrie : en finir avec une guerre sans fin (juin 2017)
- Énergie : priorité au climat ! (juin 2017)
- Quelle place pour la voiture demain ? (mai 2017)
- Sécurité nationale : quels moyens pour quelles priorités ? (avril 2017)
- Tourisme en France : cliquez ici pour rafraîchir (mars 2017)
- L'Europe dont nous avons besoin (mars 2017)
- Dernière chance pour le paritarisme de gestion (mars 2017)
- L'impossible État actionnaire ? (janvier 2017)
- Un capital emploi formation pour tous (janvier 2017)
- Économie circulaire, réconcilier croissance et environnement (novembre 2016)
- Traité transatlantique : pourquoi persévérer (octobre 2016)
- Un islam français est possible (septembre 2016)
- Refonder la sécurité nationale (septembre 2016)
- Bremain ou Brexit : Europe, prépare ton avenir ! (juin 2016)
- Réanimer le système de santé - Propositions pour 2017 (juin 2016)
- Nucléaire : l'heure des choix (juin 2016)
- Un autre droit du travail est possible (mai 2016)
- Les primaires pour les Nuls (avril 2016)
- Le numérique pour réussir dès l'école primaire (mars 2016)
- Retraites : pour une réforme durable (février 2016)
- Décentralisation : sortons de la confusion / Repenser l'action publique dans les territoires (janvier 2016)
- Terreur dans l'Hexagone (décembre 2015)
- Climat et entreprises : de la mobilisation à l'action / Sept propositions pour préparer l'après-COP21 (novembre 2015)
- Discriminations religieuses à l'embauche : une réalité (octobre 2015)
- Pour en finir avec le chômage (septembre 2015)
- Sauver le dialogue social (septembre 2015)
- Politique du logement : faire sauter les verrous (juillet 2015)
- Faire du bien vieillir un projet de société (juin 2015)
- Dépense publique : le temps de l'action (mai 2015)
- Apprentissage : un vaccin contre le chômage des jeunes (mai 2015)
- Big Data et objets connectés. Faire de la France un champion de la révolution numérique (avril 2015)
- Université : pour une nouvelle ambition (avril 2015)
- Rallumer la télévision : 10 propositions pour faire rayonner l'audiovisuel français

(février 2015)
- Marché du travail : la grande fracture (février 2015)
- Concilier efficacité économique et démocratie : l'exemple mutualiste (décembre 2014)
- Résidences Seniors : une alternative à développer (décembre 2014)
- Business schools : rester des champions dans la compétition internationale (novembre 2014)
- Prévention des maladies psychiatriques : pour en finir avec le retard français (octobre 2014)
- Temps de travail : mettre fin aux blocages (octobre 2014)
- Réforme de la formation professionnelle : entre avancées, occasions manquées et pari financier (septembre 2014)
- Dix ans de politiques de diversité : quel bilan ? (septembre 2014)
- Et la confiance, bordel ? (août 2014)
- Gaz de schiste : comment avancer (juillet 2014)
- Pour une véritable politique publique du renseignement (juillet 2014)
- Rester le leader mondial du tourisme, un enjeu vital pour la France (juin 2014)
- 1 151 milliards d'euros de dépenses publiques : quels résultats ? (février 2014)
- Comment renforcer l'Europe politique (janvier 2014)
- Améliorer l'équité et l'efficacité de l'assurance-chômage (décembre 2013)
- Santé : faire le pari de l'innovation (décembre 2013)
- Afrique-France : mettre en oeuvre le co-développement Contribution au XXVI^e^ sommet Afrique-France (décembre 2013)
- Chômage : inverser la courbe (octobre 2013)
- Mettre la fiscalité au service de la croissance (septembre 2013)
- Vive le long terme ! Les entreprises familiales au service de la croissance et de l'emploi (septembre 2013)
- Habitat : pour une transition énergétique ambitieuse (septembre 2013)
- Commerce extérieur : refuser le déclin
  Propositions pour renforcer notre présence dans les échanges internationaux (juillet 2013)
- Pour des logements sobres en consommation d'énergie (juillet 2013)
- 10 propositions pour refonder le patronat (juin 2013)
- Accès aux soins : en finir avec la fracture territoriale (mai 2013)
- Nouvelle réglementation européenne des agences de notation : quels bénéfices attendre ? (avril 2013)
- Remettre la formation professionnelle au service de l'emploi et de la compétitivité (mars 2013)
- Faire vivre la promesse laïque (mars 2013)
- Pour un « New Deal » numérique (février 2013)

- Intérêt général : que peut l'entreprise ? (janvier 2013)
- Redonner sens et efficacité à la dépense publique 15 propositions pour 60 milliards d'économies (décembre 2012)
- Les juges et l'économie : une défiance française ? (décembre 2012)
- Restaurer la compétitivité de l'économie française (novembre 2012)
- Faire de la transition énergétique un levier de compétitivité (novembre 2012)
- Réformer la mise en examen Un impératif pour renforcer l'État de droit (novembre 2012)
- Transport de voyageurs : comment réformer un modèle à bout de souffle ? (novembre 2012)
- Comment concilier régulation financière et croissance : 20 propositions (novembre 2012)
- Taxe professionnelle et finances locales : premier pas vers une réforme globale ? (septembre 2012)
- Remettre la notation financière à sa juste place (juillet 2012)
- Réformer par temps de crise (mai 2012)
- Insatisfaction au travail : sortir de l'exception française (avril 2012)
- Vademecum 2007 – 2012 : Objectif Croissance (mars 2012)
- Financement des entreprises : propositions pour la présidentielle (mars 2012)
- Une fiscalité au service de la « social compétitivité » (mars 2012)
- La France au miroir de l'Italie (février 2012)
- Pour des réseaux électriques intelligents (février 2012)
- Un CDI pour tous (novembre 2011)
- Repenser la politique familiale (octobre 2011)
- Formation professionnelle : pour en finir avec les réformes inabouties (octobre 2011)
- Banlieue de la République (septembre 2011)
- De la naissance à la croissance : comment développer nos PME (juin 2011)
- Reconstruire le dialogue social (juin 2011)
- Adapter la formation des ingénieurs à la mondialisation (février 2011)
- « Vous avez le droit de garder le silence… » Comment réformer la garde à vue (décembre 2010)
- Gone for Good ? Partis pour de bon ? Les expatriés de l'enseignement supérieur français aux États-Unis (novembre 2010)
- 15 propositions pour l'emploi des jeunes et des seniors (septembre 2010)
- Afrique - France. Réinventer le co-développement (juin 2010)
- Vaincre l'échec à l'école primaire (avril 2010)
- Pour un Eurobond. Une stratégie coordonnée pour sortir de la crise (février 2010)
- Réforme des retraites : vers un big-bang ? (mai 2009)

- Mesurer la qualité des soins (février 2009)
- Ouvrir la politique à la diversité (janvier 2009)
- Engager le citoyen dans la vie associative (novembre 2008)
- Comment rendre la prison (enfin) utile (septembre 2008)
- Infrastructures de transport : lesquelles bâtir, comment les choisir ? (juillet 2008)
- HLM, parc privé
  Deux pistes pour que tous aient un toit (juin 2008)
- Comment communiquer la réforme (mai 2008)
- Après le Japon, la France…
  Faire du vieillissement un moteur de croissance (décembre 2007)
- Au nom de l'Islam… Quel dialogue avec les minorités musulmanes en Europe ?
  (septembre 2007)
- L'exemple inattendu des Vets
  Comment ressusciter un système public de santé (juin 2007)
- Vademecum 2007-2012
  Moderniser la France (mai 2007)
- Après Erasmus, Amicus. Pour un service civique universel européen (avril 2007)
- Quelle politique de l'énergie pour l'Union européenne ? (mars 2007)
- Sortir de l'immobilité sociale à la française (novembre 2006)
- Avoir des leaders dans la compétition universitaire mondiale (octobre 2006)
- Comment sauver la presse quotidienne d'information (août 2006)
- Pourquoi nos PME ne grandissent pas (juillet 2006)
- Mondialisation : réconcilier la France avec la compétitivité (juin 2006)
- TVA, CSG, IR, cotisations…
  Comment financer la protection sociale (mai 2006)
- Pauvreté, exclusion : ce que peut faire l'entreprise (février 2006)
- Ouvrir les grandes écoles à la diversité (janvier 2006)
- Immobilier de l'État : quoi vendre, pourquoi, comment (décembre 2005)
- 15 pistes (parmi d'autres…) pour moderniser la sphère publique (novembre 2005)
- Ambition pour l'agriculture, libertés pour les agriculteurs (juillet 2005)
- Hôpital : le modèle invisible (juin 2005)
- Un Contrôleur général pour les Finances publiques (février 2005)
- Les oubliés de l'égalité des chances (janvier 2004 - Réédition septembre 2005)

Pour les publications antérieures se référer à notre site internet :
www.institutmontaigne.org

INSTITUT
MONTAIGNE

ABB FRANCE
ABBVIE
ACCURACY
ACTIVEO
ADIT
ADVANCY
AIR FRANCE - KLM
AIR LIQUIDE
AIRBUS
ALLEN & OVERY
ALLIANZ
ALVAREZ & MARSAL FRANCE
AMAZON WEB SERVICES
AMBER CAPITAL
AMUNDI
ARCHERY STRATEGY CONSULTING
ARCHIMED
ARDIAN
ASTORG
ASTRAZENECA
AUGUST DEBOUZY
AVRIL
AXA
BAKER & MCKENZIE
BANK OF AMERICA MERRILL LYNCH
BEARINGPOINT
BESSÉ
BNP PARIBAS
BOLLORÉ
BOUGARTCHEV MOYNE ASSOCIÉS
BOUYGUES
BROUSSE VERGEZ
BRUNSWICK
CAISSE DES DÉPÔTS
CANDRIAM
CAPGEMINI
CAPITAL GROUP
CAREIT
CARREFOUR
CASINO
CHAÎNE THERMALE DU SOLEIL
CHUBB
CIS
CISCO SYSTEMS FRANCE
CMA CGM
CNP ASSURANCES
COHEN AMIR-ASLANI
COMPAGNIE PLASTIC OMNIUM
CONSEIL SUPÉRIEUR DU NOTARIAT
CORREZE & ZAMBEZE
CRÉDIT AGRICOLE
CRÉDIT FONCIER DE FRANCE
D'ANGELIN &CO.LTD
DASSAULT SYSTÈMES
DE PARDIEU BROCAS MAFFEI
DENTSU AEGIS NETWORK
DRIVE INNOVATION INSIGHT - DII
EDF
EDHEC BUSINESS SCHOOL
EDWARDS LIFESCIENCES
ELSAN
ENEDIS
ENGIE
EQUANCY
ESL & NETWORK
ETHIQUE & DÉVELOPPEMENT
EURAZEO
EUROGROUP CONSULTING
EUROSTAR
FIVES
FONCIA GROUPE
FONCIÈRE INEA
GALILEO GLOBAL EDUCATION
GETLINK
GIC PRIVATE LIMITED
GIDE LOYRETTE NOUEL
GOOGLE
GRAS SAVOYE
GROUPAMA
GROUPE EDMOND DE ROTHSCHILD
GROUPE M6
HAMEUR ET CIE
HENNER
HSBC FRANCE
IBM FRANCE
IFPASS
ING BANK FRANCE
INSEEC
INTERNATIONAL SOS
INTERPARFUMS
IONIS EDUCATION GROUP
ISRP
JEANTET ASSOCIÉS
KANTAR
KATALYSE
KEARNEY
KPMG S.A.
LA BANQUE POSTALE

SOUTIENNENT L'INSTITUT MONTAIGNE

INSTITUT
MONTAIGNE

LA PARISIENNE ASSURANCES
LAZARD FRÈRES
LINEDATA SERVICES
LIR
LIVANOVA
L'ORÉAL
LOXAM
LVMH
M.CHARRAIRE
MACSF
MALAKOFF MÉDÉRIC
MAREMMA
MAZARS
MCKINSEY & COMPANY FRANCE
MÉDIA-PARTICIPATIONS
MEDIOBANCA
MERCER
MERIDIAM
MICHELIN
MICROSOFT FRANCE
MITSUBISHI FRANCE S.A.S
NATIXIS
NEHS
NESTLÉ
NEXITY
OBEA
ODDO BHF
ONEPOINT
ONDRA PARTNERS
ONET
OPTIGESTION
ORANGE
ORANO
ORTEC GROUPE
PAI PARTNERS
PRICEWATERHOUSECOOPERS
PRUDENTIA CAPITAL
RADIALL
RAISE
RAMSAY GÉNÉRALE DE SANTÉ
RANDSTAD
RATP
RELX GROUP
RENAULT
REXEL
RICOL LASTEYRIE CORPORATE FINANCE
RIVOLIER
ROCHE
ROLAND BERGER
ROTHSCHILD MARTIN MAUREL
SAFRAN
SANOFI
SAP FRANCE
SCHNEIDER ELECTRIC
SERVIER
SGS
SIA PARTNERS
SIACI SAINT HONORÉ
SIEMENS FRANCE
SIER CONSTRUCTEUR
SNCF
SNCF RÉSEAU
SODEXO
SOFINORD - ARMONIA
SOLVAY
SPRINKLR
SPVIE
STAN
SUEZ
TALAN
TECNET PARTICIPATIONS SARL
TEREGA
TETHYS
THE BOSTON CONSULTING GROUP
TILDER
TOTAL
TRANSDEV
UBER
UBS FRANCE
UIPATH
VEOLIA
VINCI
VIVENDI
VOYAGEURS DU MONDE
WAVESTONE
WAZE
WENDEL
WORDAPPEAL
WILLIS TOWERS WATSON

SOUTIENNENT L'INSTITUT MONTAIGNE

INSTITUT
MONTAIGNE

Imprimé en France
Dépôt légal : mars 2020
ISSN : 1771-6764
Achevé d'imprimer en mars 2020

INSTITUT
MONTAIGNE

IL N'EST DÉSIR PLUS NATUREL QUE LE DÉSIR DE CONNAISSANCE

Algorithmes : contrôle des biais S.V.P.

Constituer automatiquement une *playlist* avec nos chansons préférées ou trouver le résultat le plus pertinent *via* un moteur de recherche : les algorithmes nous assistent tout le long de la journée. S'ils sont formidablement efficaces, ils posent aussi des questions. Que se passerait-il si un algorithme de recrutement laissait systématiquement de côté les femmes ou des minorités ethniques ? Comment s'assurer que ces erreurs soient mises en lumière et corrigées ?

Ce rapport tente de donner une perspective française à cette problématique aujourd'hui essentiellement traitée sous un prisme américain. Il poursuit l'étude de Télécom Paris et de la Fondation Abeona, *Algorithmes : biais, discrimination et équité*, publiée en 2019. Sur la base de ce constat technique, nous avons voulu, à travers la quarantaine d'entretiens réalisés, apporter des solutions concrètes pour limiter les dérives potentielles et redonner confiance dans les algorithmes.

Rejoignez-nous sur :

Suivez chaque semaine notre actualité en vous abonnant à notre newsletter sur : www.institutmontaigne.org

Institut Montaigne
59, rue La Boétie - 75008 Paris
Tél. +33 (0)1 53 89 05 60
www.institutmontaigne.org

10 €
ISSN 1771-6764
MARS 2020