Les algorithmes sont-ils racistes ? Éléments d'analyse sociologique des discriminations en contexte numérique

Mariame Tighanimine

Socio

Socio
La nouvelle revue des sciences sociales

18 | 2023
Racisme, antisémitisme, discriminations

Les algorithmes sont-ils racistes ?

Éléments d'analyse sociologique des discriminations en contexte numérique

Are algorithms racist? Elements of sociological analysis of discrimination in the digital context

Mariame Tighanimine

OpenEdition Journals

Édition électronique
URL : https://journals.openedition.org/socio/14648
DOI : 10.4000/socio.14648
ISSN : 2425-2158

Éditeur
Les éditions de la Maison des sciences de l'Homme

Édition imprimée
Date de publication : 14 novembre 2023
Pagination : 29-57
ISBN : 978-2-7351-2953-9
ISSN : 2266-3134

Référence électronique
Mariame Tighanimine, « Les algorithmes sont-ils racistes ? », *Socio* [En ligne], 18 | 2023, mis en ligne le 29 septembre 2023, consulté le 07 janvier 2025. URL : http://journals.openedition.org/socio/14648 ;
DOI : https://doi.org/10.4000/socio.14648

Les algorithmes sont-ils racistes ?

Éléments d'analyse sociologique des discriminations en contexte numérique

Mariame TIGHANIMINE

En décembre 2020, la chercheuse en informatique Timnit Gebru, co-directrice de l'Ethical Artificial Intelligence Team chez Google, annonce avoir été renvoyée à la suite de la publication d'un de ses articles de recherche (Bender *et al.*, 2021) questionnant l'éthique des algorithmes de traitement automatique du langage[1] et la course aux intelligences artificielles plus puissantes[2] à laquelle son employeur – en tant que géant de l'industrie

1. Le traitement automatique du langage ou *natural language processing* est une branche pluridisciplinaire de l'informatique permettant aux machines de comprendre et de traduire le langage humain, et de générer des outils de traitement de la langue naturelle pour diverses applications comme les assistants vocaux ou les agents conversationnels.

2. Plus précisément, cet article alertait sur le risque d'un déploiement précipité d'algorithmes puissants, s'entraînant sur de grandes quantités de données téléchargées sur Internet.

technologique – participe activement. Cette chercheuse américano-éthiopienne-érythréenne prolifique est également connue pour ses travaux sur les biais algorithmiques et l'exploration des données. Ils font d'ailleurs autorité au point d'influencer la législation américaine sur le sujet de la reconnaissance faciale par les algorithmes, grâce notamment à un article (voir Buolamwini et Gebru, 2018) qui montre comment les systèmes de reconnaissance faciale d'IBM et de Microsoft incorporent bien plus de préjugés raciaux et sexistes que les autres logiciels de reconnaissance faciale, particulièrement lorsqu'il s'agit d'identifier les personnes à la peau foncée.

Après son renvoi, Timnit Gebru a été la cible d'une campagne médiatique de dénigrement à laquelle son employeur a participé, la présentant comme une militante et non comme la scientifique remarquée qu'elle est dans son domaine. Théoricienne, elle est aussi une praticienne saluée qui a, par exemple, été membre de l'équipe d'ingénieur·es ayant développé le premier iPad. Elle est aussi réputée pour ses actions en faveur de la diversité ethnique, sociale ou encore de genre dans les métiers de la technologie à travers des initiatives comme Black in AI[3]. Passée par Stanford, Apple, Microsoft et Alphabet (Google), elle est l'une des chercheuses en intelligence artificielle (IA) les plus remarquées de sa génération.

À elle seule, Timnit Gebru incarne une part des nombreux problèmes qui se posent au sein de la recherche en IA, et également dans les entreprises des *Big Tech*[4] (ces dernières étant elles-mêmes productrices de nombreux

3. Créé en 2017, le collectif Black in AI se donne l'objectif de résoudre le problème du manque de diversité dans la recherche en IA (et ses conséquences sur les biais algorithmiques, l'éthique ou encore l'accès aux carrières scientifiques). Son lancement a eu lieu lors d'un *workshop* de l'une des plus deux importantes conférences du domaine, la Conference on Neural Information Processing Systems (NeurIPS). Voir l'article éditorial du journal *Nature,* n° 564, « How one conference embraced diversity. Improvements to an AI event accused of sexism are long overdue », publié le 12 décembre 2018 sur nature.com : <https://www.nature.com/articles/d41586-018-07718-x>.

4. Parmi les entreprises concernées, on trouve Alphabet (Google), Amazon, Facebook (devenu Meta la semaine des révélations de l'affaire des *Facebook files*), Apple et Microsoft, appelées *Big Five* ou GAFAM. Côté chinois, l'équivalent appelé *Big Four* ou BATX compte les entreprises Baidu, Alibaba, Tencent et Xiaomi, elles aussi productrices de recherche en sciences de l'information et particulièrement en IA. Début 2023, il semblerait même que la Chine devance les États-Unis en production d'articles de recherche liés au sujet de l'intelligence artificielle.

travaux académiques sur les IA). Son cas renvoie à la fois aux sujets liés aux travaux académiques sur la responsabilité, la transparence et l'éthique[5] des systèmes informatiques, et à la question des producteur·rices de la recherche en IA (et aux effets potentiels de leurs trajectoires et caractéristiques socio-démographiques et économiques sur les recherches produites).

Au sein des arènes académiques, industrielles et médiatiques liées aux milieux des nouvelles technologies de l'information, sa qualité de femme noire (voir Johnson, 2020) a été pointée du doigt principalement pour la disqualifier (voir Schiffer, 2021). Ce fait malmène un certain mythe techniciste de la neutralité[6] des technologies et de celles et ceux qui les conçoivent, auquel adhèrent de nombreux·ses chercheur·euses. Selon cette même croyance, les données massives et les outils d'intelligence artificielle qui les travaillent auraient la capacité de transcrire le monde social tel qu'il est sous forme de données. Ces dernières seraient neutres, et leur éditorialisation est bien peu questionnée. Ce mythe de la neutralité doit nous pousser à interroger la place des représentations sociales (Durkheim, 1898 ; Moscovici, 1961) dans le monde des algorithmes pouvant être définies comme « une forme de connaissance socialement élaborée et partagée ayant une visée pratique et concourant à la construction d'une réalité commune à un ensemble social » (Jodelet, 2019 : 53), et qui constituent pour les individus un premier point de contact au monde. Ces représentations sociales produisent des données qui portent en elles des biais pouvant être identifiés et objectivés. Elles sont le fruit de rapports de

5. Les préoccupations éthiques au sujet des algorithmes vont au-delà de l'espace académique universitaire (ou des laboratoires intégrés au *Big Tech*) pour concerner toutes les couches de travailleur·euses de l'industrie de la tech. Voir Berrebi-Hoffmann et Chapus (2022).

6. Dans son livre *Race After Technology. Abolitionist Tools for the New Jim Code* (2019), la sociologue Ruha Benjamin remet en cause la neutralité supposée des technologies. Elle développe ainsi le concept de « *New Jim Code* », forgé à partir des lois Jim Crow issues des *Blacks Codes* instaurant la ségrégation raciale aux États-Unis dès la fin du XIX^{e} siècle, qu'elle définit comme « *The employment of new technologies that reflect and reproduce existing inequities but that are promoted and perceived as more objective or progressive than the discriminatory systems of a previous era* » (l'utilisation de nouvelles technologies qui reflètent et reproduisent les inégalités existantes, mais qui sont promues et perçues comme plus objectives ou progressistes que les systèmes discriminatoires d'époque antérieure), Benjamin (2019 : 6).

force et de domination, et l'on ne peut faire l'économie de leurs origines, de leurs contextes et de leurs conditions de production.

Ce questionnement est d'autant plus urgent au regard de l'irruption d'une série de technologies numériques avancées depuis le milieu des années 2010, qui a soulevé à nouveaux frais le problème des discriminations et du racisme dans le quotidien des individus et au sein d'un certain nombre d'institutions étatiques, particulièrement dans leur instrumentation numérique. L'apparition de publications académiques sur le sujet, d'ouvrages et de débats qui fleurissent dans plusieurs arènes cache divers enjeux, et amène à regarder différemment le racisme, l'antisémitisme et les discriminations aujourd'hui.

Qu'est-ce que ces controverses techniques nous disent du racisme lui-même ? Comment l'introduction des algorithmes auprès des populations ou de certains publics, et au sein des institutions conduit à repenser le racisme et les discriminations ? Comment les traiter différemment dans nos sociétés numérisées ? C'est à tenter de proposer des réponses, à l'aune des débats récents et de recherches actuelles, que cet article est consacré.

Nous commencerons par une courte histoire de l'émergence d'une prise de conscience des effets du numérique sur la résurgence et l'extension du racisme dans nos sociétés. Nous présenterons ensuite le cas précis des algorithmes de recommandation pour mieux saisir les effets d'invisibilisation et de classification des algorithmes apprenants ou *machine learning* dans la structuration de nos sociétés numériques. Il faudra ici distinguer trois niveaux de discrimination et de définition du racisme : le niveau individuel et les discriminations directes, les discriminations indirectes, les discriminations systémiques et/ou le racisme institutionnel. Nous emprunterons exemples et analyses à notre recherche et à nos enquêtes de terrain en cours[7].

7. À partir de quatre enquêtes de terrain sur les journalistes, les syndicats, les chercheurs en sécurité et éthique des IA et des groupes organisés en ligne opérant un travail de désinformation, ma thèse décrit et analyse les différentes façons dont le numérique et les algorithmes transforment des professions, des institutions, des corps intermédiaires et plus largement nos démocraties.

Les possibilités du numérique : un racisme augmenté ?

A priori, rien ne laissait penser qu'un simple calcul mathématique pourrait un jour se retrouver au cœur des transformations et de la diffusion du racisme et de l'antisémitisme dans nos sociétés. Pourtant, le renforcement des préjugés raciaux dans l'opinion n'est aujourd'hui plus dissociable de leur circulation sur Internet et les réseaux sociaux (Noble et Tynes, 2016 ; Matamoros-Fernández, 2017). Plus encore, un débat s'est développé depuis quelques années autour du rôle de l'intelligence artificielle dans l'extension de discriminations fondées sur l'origine, la race ou le sexe. Il est vrai que, depuis peu, police, justice, banques, assurances, services publics et administrations d'État mettent en place des usages automatiques d'attribution de droits et de peines, dont les effets se sont avérés particulièrement biaisés (Eubanks, 2018 ; Noble, 2018)[8].

Dès lors, il s'agit d'analyser les modes d'expression du racisme contemporain, comme ses formes institutionnelles ou systémiques, à la lumière de ce que permet, diffuse, construit ou amplifie l'omniprésence du numérique dans nos sociétés. Toutefois, ce caractère omniprésent n'aide guère à véritablement saisir le phénomène. Le recours constant aux outils numériques dans nos modes d'accès à l'information, dans nos formes de travail et de consommation, dans nos relations aux institutions – tout comme le caractère technique de leurs fonctionnements – semble d'abord faire obstacle à la réflexion, ou du moins entretenir une certaine confusion. Faut-il réguler la liberté d'expression sur les réseaux sociaux et comment ? En quoi

8. Dans son livre *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor,* Eubanks (2018) analyse trois programmes sociaux mis en place aux États-Unis pour traiter de l'automatisation des inégalités, c'est-à-dire de la manière dont les technologies perfectionnées perpétuent et accroissent les discriminations envers les couches les plus défavorisées. Elle montre notamment que l'utilisation de ces outils s'inscrit dans la continuité d'une gestion punitive et disciplinaire de la pauvreté remontant au XIX^e^ siècle (elle forge d'ailleurs le terme de « *digital poorhouse* », ou hospices numériques). Les outils numériques ne seraient que des amplificateurs de cette tendance ancienne, offrant une échelle et une vitesse de déploiement jusque-là jamais égalées. Voir également Noble et son *Algorithms of oppression. How Search Engines Reinforce Racism* (2018). La chercheuse y montre l'étendue des biais racistes et sexistes dans les résultats de recherche d'images de moteurs de recherche comme Google Search.

un algorithme, défini comme « une suite finie d'instructions simples permettant (ou non) de résoudre un problème[9] » (Boullier et El-Mhamdi, 2020), peut-il à lui seul véhiculer ou renforcer des préjugés humains et des discriminations qui relèvent de rapports sociaux et institutionnels au sein de nos sociétés ? Un algorithme n'est-il pas forcément objectif ? Tout comme les données qu'il traite ? N'est-ce pas la société elle-même qui est raciste et qui opère des discriminations ? Les algorithmes font-ils autre chose que de refléter les biais des données dans lesquelles ils puisent ?

On le voit, pour saisir la question des intrications actuelles du racisme et du numérique dans nos sociétés, on ne peut s'arrêter au simple constat de l'extension et de l'expression de la haine et des préjugés sur les réseaux sociaux. Ce sont les algorithmes eux-mêmes qu'il nous faut comprendre. L'organisation actuelle de l'information numérique et l'automatisation de décisions institutionnelles et publiques jouent-elles un rôle dans la redéfinition du racisme et, au-delà, des discriminations dans nos démocraties sociales ? Et dans la visibilité ou l'invisibilité de groupes sociaux et de leurs caractéristiques présumées ?

Les biais des données et le racisme direct automatisé

Une première discussion sur les liens entre racisme et numérique se centre sur la qualité des données statistiques et bases de données qui équipent nos États de droit, d'une part, et sur leur traitement algorithmique qui en amplifierait les biais, d'autre part. L'usage du mode conditionnel peut étonner et pourtant, il ne devrait pas. Les algorithmes de *machine learning* et leurs modèles prédictifs sont souvent assimilés à des boîtes noires (Pasquale, 2015), dont l'opacité rend impossible la compréhension, y compris pour leurs concepteur·rices. Cela ne peut être attribué à une volonté individuelle ou collective, mais plutôt à la complexité inhérente aux objets manipulés. En outre, au fil des années, l'écart s'est creusé entre le nombre de lignes de code que les développeur·euses humain·es écrivent pour

9. « Un algorithme est une suite finie d'instructions simples permettant (ou non) de résoudre un problème. Plus important encore, un algorithme est une suite d'instructions tellement simples et non ambiguës qu'elle doit permettre à un agent, humain ou machine, de résoudre un problème sans fournir d'autres efforts intellectuels que l'application rigoureuse de ces instructions » (Boullier et El-Mhamdi, 2020).

établir la première version d'un algorithme d'apprentissage, et le nombre de lignes que peut écrire l'un des algorithmes d'apprentissage les plus avancés aujourd'hui capable de contenir mille milliards de paramètres et donc, d'être impossible à auditer par un être humain (contrairement aux algorithmes classiques de quelques milliers de lignes de code). Étant donné que ces paramètres résultent des données d'entraînement, l'illisibilité des modèles de très grande taille (*very large models*) qui en résultent pousse les experts interrogés sur notre terrain de recherche à penser qu'aujourd'hui, les données manipulent les algorithmes bien plus que les concepteur·rices de ces derniers (voir El-Mhamdi *et al.*, 2022). On peut entendre ou lire à propos des algorithmes qu'ils ne sont pas racistes, mais que les données le sont. La réalité est un peu plus complexe ; toujours est-il qu'une fois qu'un algorithme s'est entraîné, a appris et s'est déployé, il devient ce qu'il a appris. Si par exemple les données étaient sexistes, l'algorithme devient sexiste.

En 2016, Microsoft et son moteur de recherche Bing conduisent une expérience qui démontre remarquablement et rapidement ce premier point. L'entreprise déploie sur les réseaux sociaux un *chatbot*, c'est-à-dire un automate virtuel capable de converser et d'apprendre à s'exprimer en circulant, lisant et apprenant des autres utilisateur·rices. Elle le prénomme Tay et son compte est supposé être celui d'une adolescente. Moins de vingt-quatre heures après son lancement sur le réseau social Twitter, Tay se met à poster des insultes racistes et sexistes, ainsi que des messages haineux et complotistes. Microsoft doit alors retirer son *chatbot* du réseau social. En 2016, après le déploiement de Tay, de nombreux médias ont écrit sur la nature extrémiste voire « (néo)nazie » des messages postés par Tay. Une étude menée en interne en 2021 par des chercheur·euses de Twitter (Huszár *et al.*, 2022) révèle que les contenus « de droite » sont plus mis en avant par les algorithmes de la plateforme, sans que cette dernière ni ses chercheurs puissent vraiment l'expliquer. Un élément de réponse peut être avancé grâce aux travaux menés sur l'activisme en ligne, pointant la suractivité des militants de droite (Schradie, 2019), voire d'extrême droite par rapport aux autres tendances politiques. Cette suractivité, Pierre Haski, président du conseil d'administration de Reporters sans frontières, journaliste passé par de nombreux médias dont *Libération* (où il a été rédacteur en chef) et le *pure player* Rue89 (dont il est cofondateur), nous la confirme lors d'un entretien mené en 2019 :

> *La crise démocratique a trouvé un exutoire dans l'existence des réseaux sociaux qui permettent de parler sans filtres, et dans lesquels se sont engouffrés, et ça, je pense qu'il faut en être conscient, des groupes minoritaires qui ont découvert une manière d'exister, là où un univers médiatique* mainstream *ne leur permettait pas de s'exprimer. Et aujourd'hui, ces groupes ont un coup d'avance, parce qu'ils ont des années d'expérience. Je vois la fachosphère qui est la plus perfectionnée dans le domaine. Ça fait sept, huit ans, dix ans qu'elle sait faire sur Internet, qu'elle sait… Y compris sur un plan technique, ce qu'on cherche tous à faire sur nos médias : valoriser leur référencement, jouer avec l'algorithme de Facebook et de Google, etc. Ils sont très forts parce que ça fait plus de dix ans qu'ils font ça. Comme ils n'avaient pas accès aux grands médias, ils se sont engouffrés dans la brèche d'Internet. Et donc, on se retrouve dans cette situation où on a, d'un côté, des médias et, de l'autre, un univers social ouvert dans lequel des minorités agissantes ont un poids surreprésenté…*

Si l'expression raciste et extrémiste « directe » est facilement identifiable par les individus et les plateformes et par leurs systèmes de modération, des formes de détournement et d'amoindrissement des idées les plus nauséabondes ont vu le jour. L'utilisation de périphrases ou de codes pour contourner la modération (automatique) des algorithmes sur les plateformes de réseaux sociaux afin d'éviter la suppression, la suspension ou le *shadow banning* (une technique de modération consistant à rendre un·e utilisateur·rice ou ses productions invisibles ou moins visibles pour les autres membres du réseau social, tout en les gardant généralement visibles pour lui·elle) renvoie à l'*algospeak* (voir Lorenz, 2022), c'est-à-dire la création d'un nouveau langage ou vocabulaire (par exemple, « *unalive* » pour « *death* »), la modification de la syntaxe d'un mot (« vi0l » pour « viol »), voire l'utilisation d'émojis ou le recours aux métaphores, euphémismes et autres paraphrases[10].

10. Autre exemple, pour exprimer des idées antisémites, des internautes ont trouvé des moyens détournés et des expressions d'apparence neutre comme celle des « dragons célestes » dans l'Internet francophone pour désigner les Juifs. Cela évite les suppressions de contenus et de comptes, ou les bannissements pour messages haineux. Cette expression particulière renvoie à la culture manga et plus précisément à *One Piece* dans lequel les dragons célestes sont des nobles issus d'une lignée de rois au fonde-

Toutefois, le développement de ces expressions et formes de résistance face aux plateformes est à double tranchant. Le contournement des règles de modération *via* l'*algospeak* peut autant profiter à des groupes extrémistes et racistes qu'à des catégories très signalées ou harcelées (minorités ethniques, sexuelles, etc.) par ces mêmes groupes extrémistes.

On peut qualifier ce premier effet de racisme automatisé « direct », à l'instar de la littérature sur les discriminations qui fait la différence entre discriminations directes, indirectes et systémiques (Dhume, 2016). Si le jeu de données d'entraînement contient des biais sexistes, racistes et/ou validistes, les algorithmes entraînés avec ces données seront racistes, sexistes et/ou validistes. Cet effet simple, au-delà de son caractère marquant, ne permet pourtant pas de cerner l'ensemble des questions que posent les données. Une seconde interrogation plus fondamentale est celle de la construction sociale des données et de l'information quantitative en général, de leurs catégories de classement et de leur caractère négocié et performatif dont, par exemple, Desrosières (1993) pour les catégories socio-professionnelles, Salais *et al.* (1986) pour les chiffres du chômage, Schor (2009) dans le cas des stratifications ethniques et raciales aux États-Unis ou encore Didier (2009) et la statistique agricole ont montré l'importance.

Le racisme embarqué ou racisme « indirect »

La sociologie et l'histoire réfléchissent depuis longtemps au problème du racisme embarqué dans la construction des données. Dans *Compter et classer. Histoire des recensements américains* (2009), Paul Schor s'intéresse à une période allant de la fin du XVIII[e] siècle jusqu'à 1940 et montre comment le recensement se révèle être avant tout un instrument pour orienter la répartition des sièges à la Chambre des représentants et donc, une opération hautement politique. Il explique comment la mise en place du mécanisme de l'*apportionment* (répartition) est un épisode fondateur et constitutif de l'histoire des recensements américains, désignant la population comme mesure pour la répartition du pouvoir et de la fiscalité entre les nouveaux États fédérés. À travers la pratique régulière du recensement, dont les règles et les popula-

ment d'une organisation appelée « gouvernement mondial ». Ils bénéficient d'un statut d'intouchables, autant de caractéristiques qui font échos aux pires clichés antisémites.

tions à recenser ont été inscrites dans la Constitution, la définition d'une population a émergé selon le statut des habitants – indiens, esclaves ou libres. Si les premiers ont été exclus du recensement car non soumis à l'impôt, les deuxièmes, considérés comme hommes et biens ont été dénombrés à partir du principe des 3/5 (d'un homme libre) dans le calcul de l'assiette des impôts et de la représentation. Ainsi, même si la Constitution faisait une distinction non selon la couleur ou la race, mais selon le statut, compter à part les esclaves dès le premier recensement de 1790 a généré une distinction raciale au sein de la population en posant les catégories « Blancs » et « Noirs » comme équivalents respectivement au statut de « libre » et d'« esclave ». La conséquence directe de ce *three-fifths compromise* (compromis des trois cinquièmes) est son utilisation pour déterminer le nombre de sièges que chaque État aurait à la Chambre des représentants des États-Unis. Ces opérations ont fini par y donner une représentation disproportionnée des États esclavagistes par rapport aux électeurs des États libres jusqu'à la guerre de Sécession. Pour Schor, cette articulation entre le partage politique et l'assignation identitaire est devenue le fondement des classifications ultérieures, au détriment du projet de libéralisation de la société américaine. Le principe racial finit par se fixer au cours des recensements du XIX^e^ siècle, avec l'apparition de la catégorie des mulâtres où est classé tout individu ayant un seul ancêtre d'ascendance africaine sub-saharienne, selon le principe social et juridique de la *one-drop rule* (règle de l'unique goutte de sang). Il influence par la suite le destin des multiples vagues d'immigration que connaissent les États-Unis au cours des siècles suivants. Si les populations asiatiques ont été recensées à partir du critère de la race, celles issues de pays d'Europe ont été saisies selon leurs origines nationales.

À travers le cas étasunien, on note l'incidence des traditions singulières, autonomes et des éléments ancrés dans les cultures nationales. Les débats sur l'introduction de la race, de l'ethnie ou du religieux dans les politiques de comptage et de classification des populations ont concerné différents pays et aires géographiques et politiques, ainsi que leurs opérateurs comme les statisticien·nes ; les institutions de ces pays leur ont permis d'entretenir de fortes relations à partir du milieu du XIX^e^ siècle, moment d'intensification des échanges scientifiques internationaux (Brian, 2002).

Les algorithmes s'entraînent sur des données d'actions, de décisions et de comportements passés, souvent le résultat de choix et surtout

de discriminations anciennes. D'une certaine manière, il s'agit de partir de ces données pour interpréter le présent et le futur. Dès lors, au-delà des biais de données, ce sont l'histoire et la culture d'un pays et, avec elles, les politiques passées et les comportements d'organisations et d'institutions telles que la police et la justice qui sont embarqués dans les jeux de données à partir desquels les algorithmes apprennent, au risque de renforcer les préjugés, les discriminations et le racisme. L'exemple des logiciels de prévention des risques de récidive est tout à fait éloquent, particulièrement celui de « COMPAS » (voir Corbett-Davies *et al.*, 2016) développé aux États-Unis par la société Northpointe afin de fournir aux juges, lors des audiences, une estimation des risques de récidive des prévenu·es. La controverse autour de ce logiciel peut sembler datée et pourtant, elle reste importante à rappeler car elle est devenue une étude de cas emblématique dans la littérature sur l'équité[11] en apprentissage automatique, notamment pour sa mise en lumière de l'existence de métriques d'équité (Castelnovo *et al.*, 2022) différentes et non compatibles, ayant des conséquences concrètes sur les individus. Ce système d'apprentissage automatisé a été entraîné à partir d'une base de données sur la délinquance répertoriant les réponses d'accusé·es à 137 questions (par exemple « *How many prior juvenile felony offense arrests*[12] ? »), et produisant un score de risque sur une échelle de 1 à 10, ainsi que des labels pouvant qualifier le risque de récidive de « faible », « moyen » ou « élevé ». COMPAS fonctionne comme une technologie de

11. Il existe toute une littérature en informatique dédiée à la *fairness* (équité), portant particulièrement sur la correction des biais algorithmiques dans les processus de décision automatisés. Les chercheurs·euses qui en sont à l'origine savent que les données d'entraînement des modèles d'apprentissage automatique sont collectées et compilées par des individus ayant des objectifs, empreints de préjugés et inscrits dans des contextes sociaux et idéologiques divers. La recherche en *fairness*, volontiers ouverte à l'interdisciplinarité et discutant ainsi avec la philosophie morale et politique, le droit ou encore les sciences sociales, produit des travaux sur l'intelligibilité et la transparence des systèmes algorithmiques, l'élaboration de métriques d'équité ou encore sur les effets politiques et sociaux de certaines catégories d'algorithmes. Il existe une conférence scientifique annuelle consacrée à ces travaux, nommée « ACM Conference on Fairness, Accountability, and Transparency » (ACM FAccT). Pour en savoir plus, voir Laufer *et al.* (2022).

12. Voir le document comportant les questions, mis à disposition par le site ProRepublica : « Sample-COMPAS-Risk-Assessment-COMPAS-"CORE" » : <https://www.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE.html>.

suivi dans la mesure où elle contrôle, agrège et enregistre toutes les données concernant les activités et les comportements des accusé·es afin de les « gérer ». Avec ses labels de tri, elle opère un classement des prévenu·es, et les peines sont distribuées selon la qualification attribuée. Par exemple, les prévenus à risque « élevé » sont plus susceptibles de recevoir une peine d'incarcération tandis que les personnes à risque « faible » peuvent bénéficier d'une probation (bien que la décision finale revienne aux juges). COMPAS, comme d'autres technologies de tri, s'est finalement révélée orientée en défaveur des accusé·es afro-américain·es. ProPublica, organisation à but non lucratif à l'origine d'enquêtes journalistiques considérées d'intérêt public, publie en 2016 (voir Angwin *et al.*, 2016) une enquête portant sur les scores COMPAS obtenus par 7 000 personnes arrêtées en 2013 et 2014 dans le comté de Broward en Floride. Les journalistes de ProPublica ont aussi vérifié combien, parmi ces personnes, ont été accusées de nouveaux crimes au cours des deux années suivantes, en utilisant la même référence que les créateurs de l'algorithme[13]. La comparaison entre les prédictions de récidive formulées par le logiciel et ce qui s'est réellement passé les années suivantes a mis en évidence que le taux de récidive réel des Noirs a été largement en deçà des estimations du logiciel, alors que celui des Blancs a été nettement minoré. Ici, la discrimination apparaît indirecte.

Les jugements rendus par cette technologie semblent être une approximation imparfaite de ceux des humains. Si les algorithmes d'intelligence artificielle ne sont pas forcément programmés pour formuler des jugements racistes, ils opèrent une généralisation à partir des données d'entraînement qui les rend racistes.

13. Voir le document intitulé « Practitioners-Guide-to-COMPAS-Core », qui indique comment la notion de récidive est définie. « *For most of our analysis of COMPAS risk scores, we defined recidivism as a new arrest within two years. We based this decision on Northpointe's practitioners guide, which says that its recidivismscore is meant to predict "a new misdemeanor or felony offense within two years of the COMPAS administration date* » (Pour la majeure partie de notre analyse des scores de risque du COMPAS, nous avons défini la récidive par une nouvelle arrestation dans les deux ans. Nous avons fondé cette décision sur le guide du praticien de Northpointe, qui indique que son score de récidive est censé prédire « un nouveau délit ou crime dans les deux ans suivant la date d'administration du COMPAS) » : <https://www.documentcloud.org/documents/2840784-Practitioner-s-Guide-to-COMPAS-Core.html#document/p30/a296482>.

L'amplification des biais : les algorithmes fonctionnent en discriminant

De manière générale, le travail des algorithmes consiste à analyser des données et à les classer en leur attribuant des scores et ainsi, à conférer à certaines d'entre elles plus d'importance qu'à d'autres. Pourtant, la nature et la source de ces données sont variées et ne découlent pas toujours des mêmes composantes du monde social que le travail des algorithmes prétend parfaitement numériser. Que se passe-t-il lorsque la décision, par exemple d'attribution de droits ou de recrutement, est confiée aux algorithmes ? Plusieurs exemples récents montrent que s'automatisent également les inégalités et les discriminations structurelles. En 2018, l'entreprise Amazon développe un outil de ressources humaines automatisé qui s'est avéré plus favorable aux hommes qu'aux femmes (Dastin, 2018). En effet, l'algorithme a été entraîné à partir des curriculums vitae de candidat·es qui ont par la suite été embauché·es au sein de l'entreprise entre 2004 et 2014 (principalement des hommes). L'historique de CV dont disposait l'entreprise était donc biaisé avec une sous-représentation des femmes parmi les salarié·es occupant les postes les mieux payés. En outre, l'outil semblait accorder une importance à l'université d'origine des nouveaux·elles candidat·es. Ainsi, un·e postulant·e diplômé·e d'une université X, qui comptait un certain nombre de ses alumnis parmi les salarié·es d'Amazon, avait plus de chance d'être recruté·e qu'un·e candidat·e issu·e d'une autre université moins bien représentée dans la base de données. Or, les États-Unis ont la particularité de posséder des *women's* colleges (Harwarth *et al.*, 1997), ces établissements de l'enseignement supérieur de premier cycle dont les populations étudiantes sont composées exclusivement ou presque de femmes. Ainsi, l'intelligence artificielle a pénalisé les candidates formées dans ces universités non mixtes, et est même allée jusqu'à pénaliser les *curriculum vitae* contenant le mot « femme ». Ici, la technologie a reproduit une inégalité historique. Le fonctionnement même des systèmes numériques intelligents consiste à discriminer et donc à amplifier également les biais des données. C'est en ce sens que l'on pourrait aller jusqu'à parler – en ayant en tête les promesses et l'utopie numérique d'un humain augmenté – d'un racisme augmenté.

Classification, survisibilisation et invisibilisation : le cas des algorithmes de recommandation

Peut-on se contenter de montrer qu'il y a racisme et amplification de racisme embarqué dans les jeux de données ? Ne peut-on aller au-delà et tenter de comprendre comment ces technologies transforment aujourd'hui notre instrumentation sociale qui se retrouve *de facto* à jouer un rôle actif dans l'élaboration et la diffusion des biais ? Pour ce faire, intéressons-nous un moment à la manière dont sont conçus certains algorithmes, puis attachons-nous à saisir des algorithmes particuliers à l'origine du fonctionnement de Facebook (Meta) ou de YouTube (Google) : les algorithmes de recommandation de contenus.

Des choix sociaux au cœur de la fabrique des algorithmes

Dans les disciplines des statistiques et de l'apprentissage automatique, les chercheur·euses nomment « *ground truth* » ou « vérité de terrain » la connaissance exhaustive d'une information que l'on peut qualifier de vraie ou de réelle, mais à laquelle nous n'avons pas facilement accès dans son entièreté. Les données empiriques issues de l'observation ou de mesures recueillies directement sur le terrain permettent de s'approcher de cette *ground truth*. Ce concept désigne donc la réalité qui existe en dehors des modèles, celle que ceux-ci doivent espérer modéliser en collectant, traitant et étiquetant les données issues d'un échantillonnage aussi proche et aussi représentatif que possible de cette *ground truth*. Dans cet environnement, les biais sont considérés comme des écarts à la réalité ou vérité, et leur neutralisation devient essentielle dans le travail des chercheur·euses en apprentissage automatique. Seulement, les réalités de nos mondes sociaux sont elles-mêmes biaisées et produisent des données qui le sont. Les nombreux apports de la sociologie vont même au-delà de la notion de biais – dans son acception aussi bien statistique que psychologique – en documentant comment les trajectoires, les normes, les inscriptions dans des groupes sociaux façonnent les individus et leurs pratiques, et plus généralement toutes les structures de nos sociétés. Si ces dernières sont organisées de façon inégalitaire, injuste ou encore raciste, les technologies qui y seront développées ont des chances d'en être, au mieux, de simples miroirs, au pire, des miroirs déformés et exagérés. Des *verbatim* tirés de notre

travail ethnographique portant sur la structuration et le développement d'un collectif de chercheurs en informatique et mathématiques agissant au sein et à la marge de l'arène professionnelle permettent de comprendre l'importance et la spécificité des algorithmes de recommandation dans les sujets qui nous occupent.

> *Afin de réaliser les tâches d'apprentissage de manière à « passer à l'échelle », il n'est pas possible de créer des instructions individualisées, à la main, pour chaque donnée. Une fonction « objectif*[14] *» permet de traduire l'intention du concepteur·rice, à travers ce qu'il/elle entend par « écart à la normale ». Cette fonction estime l'écart entre la prédiction du modèle et l'observation empirique réelle dans l'échantillon, et donne une évaluation au modèle. L'apprentissage consiste en l'ajustement des paramètres du modèle jusqu'à ce que la valeur moyenne de l'écart observé sur tous les échantillons, estimée par cette fonction objectif, soit la plus petite possible. Ce paradigme d'une extrême simplicité permet de faire face, avec quelques lignes de code, à des données de plus en plus diverses et complexes. Mais cette simplicité a un prix : la perte en fiabilité. Ou, pour reprendre Alan Turing,* « If a machine is expected to be infallible, it cannot also be intelligent[15] » (El-Mahdi El-Mhamdi, chercheur en intelligence artificielle à l'École polytechnique et ancien *senior scientist* chez Google, juillet 2021).

Du fait du réductionnisme qui est présent dans leur conception, les algorithmes d'apprentissage peuvent fortement diverger des mondes sociaux qu'ils sont supposés modéliser. Si les acteurs sociaux agissent et s'expriment selon des points de vue socialement, économiquement ou politiquement situés, il est une tendance consistant à croire qu'il en est autrement pour les algorithmes. Ces derniers parcourent des données massives à la recherche de régularités statistiques, opèrent des corrélations avec le risque de produire de fausses interprétations et d'envisager

14. Appelée aussi fonction de perte ou *loss function*. Les sociologues ayant déjà employé en méthodes quantitatives la régression linéaire sont familiers d'un exemple de fonction de perte : la somme des carrés de l'erreur utilisée dans la méthode dite des moindres carrés.

15. « Si l'on attend d'une machine qu'elle soit infaillible, elle ne peut pas non plus être intelligente. »

des causalités entre des données qui ne sont en réalité que du bruit, soit des phénomènes aléatoires et hasardeux.

Autre problème, celui de la fausseté des données. Différentes catégories d'acteur·rices politiques, sociaux·ales et économiques tentent de manipuler les algorithmes d'apprentissage et de les reprogrammer en attaquant la base de données d'entraînement. Les attaques par empoisonnement cherchent à « modifier le comportement du système d'IA en introduisant des données corrompues en phase d'entraînement (ou d'apprentissage) » (CNIL) – et donc, au moment où le système d'intelligence artificielle construit un modèle à partir de données –; elles sont nombreuses et dangereuses. Injecter de fausses données sous la forme de likes, clics, temps de visionnage, campagnes de désinformation, création de contenus et autres activités factices pour donner l'impression que ces contenus sont populaires et appréciés par le plus grand nombre constitue aujourd'hui, pour tout système politique démocratique, la plus grande menace représentée par une certaine catégorie d'algorithmes : les algorithmes de recommandation. Ces derniers touchent aujourd'hui le plus d'humains à travers le monde.

> *Il y a une expression qui n'est pas du tout anodine dans ce que j'ai pu dire, celle de « valeur moyenne ». On aurait pu la remplacer par « valeur médiane ». L'esprit de l'opération serait* a priori *le même, mais les conséquences sociales ne seront pas du tout les mêmes. Un algorithme construit en apprenant sur des données utilisant la valeur moyenne va par exemple être plus « inclusif » vis-à-vis des valeurs rarement observées dans un échantillon. En contrepartie, il va être plus facilement manipulable par un acteur malveillant ou* via *l'introduction par inadvertance de données erronées*[16]*. À l'inverse, un algorithme construit en apprenant sur les mêmes données, mais en utilisant la valeur médiane va être résilient face aux erreurs ou à la manipulation par des extrêmes. Cette résilience va être au prix de moins d'inclusivité vis-à-vis des points de données sous-représentés dans l'échantillon. Le simple choix de médiane ou moyenne – et il y a plein d'autres choix possibles – n'a rien d'objectif et donnera des résultats très différents. Ce choix est un choix social. C'est même le nom de toute une*

16. La thèse de doctorat de El-Mhamdi porte sur ce type de considérations. Voir El-Mhamdi (2020).

> *discipline, la « théorie du choix social », que les concepteurs de systèmes de vote et les économistes connaissent bien. Mais cette théorie donne rarement des réponses définitives. Elle permet juste d'informer les choix qui doivent,* in fine, *revenir à la société* (El-Mahdi El-Mhamdi, chercheur en intelligence artificielle à l'École polytechnique et ancien *senior scientis*t chez Google, juillet 2021).

Spécificité et importance des algorithmes de recommandation

Revenons un peu en arrière. La numérisation de l'information est apparue avant l'apparition du Web accompagnant celle de l'ordinateur. Les livres des bibliothèques ont été parmi les premiers supports informationnels numérisés, et ont nécessité la création d'outils, de logiciels pour faciliter et automatiser leur recherche. C'est dans ce but que les moteurs de recherche ont été créés, afin de faciliter la recherche de ressources à partir de requêtes composées de termes et adressées à un ordinateur. Dès 1990 et la création du Web, les données ont commencé à être nombreuses, pour devenir au fil du temps massives. Les moteurs de recherche proposaient une quantité de résultats difficilement consultables dans leur intégralité. C'est pour répondre à ce problème que les systèmes de recommandation comme outil de filtrage de l'information (Pazzani et Billsus, 2007 ; Ekstrand *et al.*, 2011) ont été créés. Face à la multitude de données informationnelles accessibles (livres, musiques, images, pages Web…), le système opère une sélection des éléments d'information pouvant intéresser l'utilisateur en comparant son profil et ses préférences à celles exprimées par d'autres. Un système de recommandation est donc un système de filtrage de l'information dit collaboratif car il repose sur l'action collaborative d'utilisateur·rices qui recommandent à des pairs des éléments d'informations en votant et attribuant des notes. On peut citer le cas de Usenet (Hauben et Hauben, 1997), considéré comme l'un des premiers systèmes de recommandation permettant de noter des articles. Les notes attribuées par les utilisateur·rices ayant lu les articles servaient de données afin de faire des prédictions pour les autres ne les ayant pas lus mais pouvant être intéressés par ces contenus. C'est tout simplement l'automatisation d'une évaluation par les pairs dans la vraie vie, avec néanmoins la possibilité de multiplier le nombre de pairs en ligne. Avec l'arrivée des réseaux sociaux, le système

de vote s'est généralisé et complexifié. Liker, disliker, retweeter, partager, signaler du contenu comme inapproprié sont des manières d'exprimer un avis, mais, surtout, des manières de voter (Boullier, 2020). D'autres outils peuvent mesurer le plébiscite d'un contenu. Le temps passé à faire défiler son fil d'actualité Twitter, celui à lire un post Facebook, le temps de visionnage d'une vidéo sur Youtube sont aussi des manières de voter pour du contenu. Ce sont des comportements à partir desquels les algorithmes de recommandation des plateformes de réseaux sociaux accordent un score, procèdent à la (sur)visibilisation du contenu et finalement à sa recommandation à d'autres pairs aux caractéristiques similaires, ou, au contraire, à son invisibilisation et à sa non-recommandation.

L'extrait d'entretien qui suit est issu d'un échange avec l'un des chercheurs du collectif que nous étudions. Précisons que ces derniers sont à l'origine de réflexions et de productions scientifiques aussi bien théoriques qu'appliquées sur les questions de sécurité et d'éthique des intelligences artificielles. Ils sont porteurs d'un projet ambitieux de refonte épistémologique de l'informatique considérée comme une « philosophie morale calculable ». Répondre (techniquement) aux questions philosophiques posées par les outils informatiques et de calcul relève selon eux de l'urgence, dans la mesure où chaque requête soumise à un algorithme est en réalité un dilemme moral que l'algorithme doit résoudre dans un temps contraint. Notre échange a porté sur l'un de leurs projets consacré au développement d'un algorithme collaboratif adossé à une plateforme nommée Tournesol[17], pensée comme un outil d'audit des algorithmes de recommandation des réseaux sociaux. Une base de données appelée MNIST pour Modified ou Mixed National Institute of Standards and Technology a alors été évoquée. Il s'agit d'une base de données de chiffres écrits à la main, très utilisée en apprentissage automatique, permettant de tester

17. Cet algorithme sur <https://tournesol.app/> permet à une communauté de contributeurs certifiés d'identifier et d'amplifier des contenus préalablement jugés « d'utilité publique ». Selon plusieurs critères (liés à la pertinence ou à la qualité de l'information transmise), différents contenus sous la forme de vidéos sont ainsi comparés. Des algorithmes spécifiques sont ensuite utilisés pour traduire de manière équitable et sécurisée les critères en scores pour chacun des contenus. Supposés exprimer l'avis de la communauté des contributeurs de la plateforme, ces scores sont utilisés pour faire des recommandations de contenus.

les algorithmes (la reconnaissance de l'écriture manuscrite étant vue comme un problème difficile[18]).

> *Le* ***machine learning*** *considère de plus en plus le problème de l'apprentissage dit hétérogène. Hétérogène dans le problème qui nous concerne signifie que différentes personnes ont différents avis. Reconnaître un chiffre dans des images homogènes, c'est facile, nous serons presque tous d'accord. S'il y a un 1 dans une image, nous serons tous d'accord ou presque. Certains verront peut-être un 7, mais sinon, on sera quasiment tous d'accord. Mais les algorithmes importants ne sont pas les algos de reconnaissance de chiffres, ce sont surtout les algorithmes de recommandation* (Lê-Nguyên Hoang, chercheur en intelligence artificielle et président de l'association Tournesol, juillet 2021).

L'algorithme de recommandation de Tournesol a justement été pensé comme une solution pour ne pas subir ce qu'une poignée de minoritaires extrémistes (comme des régimes autoritaires ou des groupes organisés complotistes, racistes ou sexistes) a décidé de faire voir aux internautes sur un média comme YouTube où, rappelons-le, trois vidéos sur quatre consommées sur la plateforme le sont à la suite d'une recommandation algorithmique (voir Solsman, 2018). Produire des algorithmes de recommandation éthiques, sécurisés, promouvant la qualité nécessite de disposer de beaucoup de données. L'hypothèse sous-jacente dans le travail de ces chercheurs est que les experts (Berrebi-Hoffmann et Lallement, 2009) sont minoritaires. Si un algorithme élimine les opinions minoritaires, les opinions expertes disparaîtront. Les algorithmes de recommandation auxquels nous sommes confrontés au quotidien *via* nos modes de consommation de produits et de services divers, et par lesquels les entreprises et gouvernements tentent d'établir des profils d'utilisateurs ou de citoyens (célibataires, fans de films d'action, malades, au chômage, récidivistes…), recherchent, à travers le traitement de jeux de données brutes, des modèles et des corrélations qui font glisser les individus de catégories éprouvées socialement – mais dont on remet en cause les fondements inégalitaires et

18. Devenue un test standard, MNIST regroupe 60 000 images d'apprentissage et 10 000 images de test, issues d'une base de données antérieure.

dimensions déterministes – vers une individualisation allant de pair avec une personnalisation, les deux étant présentées comme plus égalitaires et neutres. Cette individualisation est surtout génératrice d'une indifférence au fait que malgré une inscription dans des contextes collectifs, chaque individu est singulier et n'est pas une statistique[19].

Algorithmes, racisme institutionnel et discriminations systémiques

Par nature, les algorithmes d'intelligence artificielle utilisés dans les systèmes de recommandation de contenus se modifient et s'ajustent en changeant leurs paramètres[20]. Frances Haugen, ancienne salariée américaine de la plateforme Facebook au sein de la *Civic integrity team* et lanceuse d'alerte à l'origine de l'affaire dite des *Facebook Files*[21], indique qu'un rapport interne à l'entreprise datant de 2019 fait état de plaintes formulées par de nombreux partis politiques européens concernant la modification de l'algorithme de recommandation. Ils ont déclaré constater que seuls les

19. Utile pour formuler une prévision à l'échelle d'un groupe, une statistique ne l'est plus quand il s'agit de le faire à l'échelle d'un individu. C'est d'ailleurs ce que nous avons tenté d'illustrer dans notre livre *Notre histoire France* (Tighanimine, 2022) qui raconte le parcours de deux ouvriers issus de l'immigration postcoloniale et met en avant des moments dans leurs trajectoires – encore trop peu étudiées – où leur agentivité s'est exprimée, leur permettant de connaître une destinée différente de celle des membres de leur groupe d'origine.

20. Voir la page « Vocabulaire de l'intelligence artificielle » éditée par le ministère de l'Enseignement supérieur et de la Recherche, où est proposée une définition qui permet de comprendre la notion de paramètre dans l'apprentissage automatique : « Un algorithme d'apprentissage automatique comporte un modèle dont il modifie les paramètres, de valeur initiale en général aléatoire, en fonction du résultat constaté » (<https://www.enseignementsup-recherche.gouv.fr/fr/bo/19/Hebdo6/CTNR1832601K.htm>).

21. Il s'agit de la révélation de nombreux documents de l'entreprise Facebook survenue en 2021, et comprenant des rapports de la recherche menée en interne par la plateforme. Avant ces révélations, une autre lanceuse d'alerte, Sophie Zhang, ancienne salariée de Facebook, *data scientist* au sein de l'équipe *Fake Engagement* a dénoncé en 2020 l'inaction de la plateforme face à des preuves d'opérations de manipulation politique. Zhang a précédé Haugen dans les révélations mais n'a pas bénéficié de la même préparation, ni des mêmes soutiens et tribunes. Si l'ampleur des données de Haugen y est pour quelque chose, il n'est pas inintéressant de souligner le fait que Zhang soit une femme trans d'origine asiatique. Voir Hao (2021).

messages polarisants, de nature à susciter des réactions radicales ou de colère, semblaient fonctionner, alors que, par le passé, ce sont les contenus liés à leurs idées qui étaient plébiscités. Haugen a également déclaré lors de son audition au Sénat français[22] que « l'algorithme de recommandation de la fonctionnalité "Reels" [qui permet de partager de courtes vidéos sur Instagram] qui n'était pas programmé pour être raciste [...] "a compris" que les contenus concernant les personnes de couleur étaient dans l'ensemble moins visionnés que les autres, et a donc réduit leur visibilité, puisqu'ils ne permettaient pas de maximiser le temps passé par les utilisateurs sur l'application[23]. »

Les cas de polarisation et de manipulation politiques avaient déjà été évoqués lors de l'affaire Cambridge Analytica en 2016. Dans *Le fabuleux chantier. Rendre l'intelligence artificielle robustesse bénéfique* (Hoang et El-Mhamdi, 2019), les chercheurs montrent dans une section consacrée à « La démocratisation de la cyberguerre » comment il est possible de manipuler les algorithmes de recommandation de contenus des réseaux sociaux avec peu de moyens, en prenant l'exemple des publicités achetées par l'Internet Research Agency (IRA)[24].

> *À la suite du Brexit et des élections américaines de 2016, le nom d'une entreprise allait faire le tour du monde : Cambridge Analytica (dissoute en mai 2018). Mais Cambridge Analytica (CA) allait en fait n'être que le début d'une prise de conscience mondiale d'une nouvelle vulnérabilité de la démocratie moderne : la possibilité de manipuler l'électorat de manière à la fois personnalisée et massive* [...] *46 000 dollars seulement ! L'IRA n'a investi que 0,05 % de ce qu'ont investi*

22. Voir le communiqué de presse du Sénat français : « Audition le 10 novembre au Sénat de Frances Haugen, la "lanceuse d'alerte" de Facebook » : <https://www.senat.fr/presse/cp20211028a.html>.

23. Rapport d'information du Sénat français sur la « Proposition de résolution au nom de la commission des affaires européennes, en application de l'article 73 quater du Règlement, sur la proposition de règlement du Parlement européen et du Conseil relatif à un marché intérieur des services numériques : amplifier la législation européenne sur les services numériques (DSA) pour sécuriser l'environnement en ligne » : <https://www.senat.fr/rap/r21-274/r21-2748.html>.

24. Un fait révélé par l'enquête du Congrès américain faisant suite aux accusations d'ingérences russes dans l'élection présidentielle américaine de 2016.

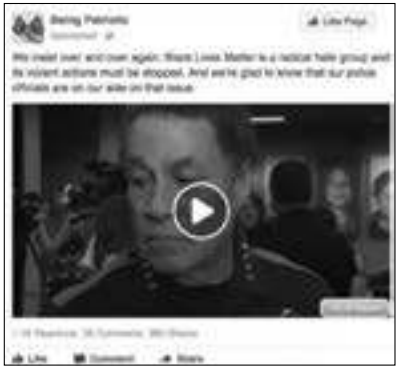

Fig. 1. Publicité affichée dans la communauté hostile au mouvement Black Lives Matter

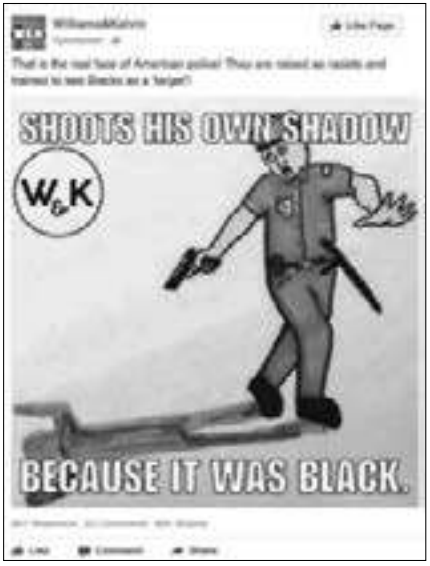

Fig. 2. Publicité affichée dans la communauté ayant des affinités avec Black Lives Matter

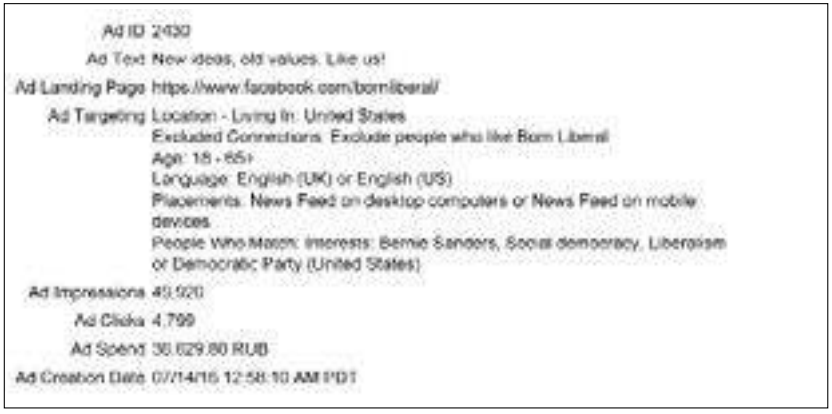

Fig. 3. Facture pour une publicité ayant ciblé l'audience favorable à Bernie Sanders

Cette facture en roubles russes (comme pour le règlement des 3500 autres publicités dévoilées par le Congrès étasunien) pour une publicité ayant ciblé l'audience favorable à Bernie Sanders lors des élections de 2016, dans le cadre d'une campagne appelant à l'abstention.

> *les équipes de campagne des candidats officiels. Plus précisément, l'IRA n'aurait acheté que 46 000 dollars de publicité, quand les candidats officiels ont dépensé 81 millions, soit 1 760 fois plus* (Hoang et El-Mhamdi, 2019 : 50).

Sur les publicités ci-contre, nous disent les auteurs, le mouvement Black Lives Matter est « utilisé par l'acheteur [...] afin de faire camper chaque camp encore plus loin de l'autre [...] Une page "patriote" dénonce ce mouvement comme radical et haineux, tout en plébiscitant les forces de l'ordre américaines [...]. Un post qui se moque de policiers américains qui seraient entraînés à voir le "Noir" comme cible, au point de tirer sur son ombre noire. Chacun des deux posts publicitaires a reçu une quantité importante de partages par des comptes authentiques. » (Hoang et El-Mhamdi, 2019 : 51)

Un troisième document présenté est une facture en rouble russe attestant de l'achat d'une publicité à destination du public de l'ancien candidat démocrate Bernie Sanders, dans le but de l'encourager à l'abstention.

On le voit donc, les effets de catégorisation, de classification et d'invisibilisation sont liés à la fois aux systèmes de vote des algorithmes et de classification hérités, dans les jeux de données, d'une culture/de traditions nationales. Ces effets réinterrogent les dimensions systémiques ou institutionnelles du racisme et des discriminations. Ils vont même au-delà dans la mesure où les questions de classification renvoient à deux autres sujets centraux en sociologie depuis Durkheim, ceux de la hiérarchie et de l'ontologie, et finalement à des questions éminemment morales. Cela n'a d'ailleurs pas échappé au collectif de chercheurs que nous étudions, caractérisé par deux exigences : l'une de rationalité, s'exprimant par exemple dans la production scientifique ; l'autre de justice, considérant que leur milieu professionnel doit prendre conscience des effets sociaux et politiques des technologies qu'ils développent ou contribuent à développer, dont le racisme et les discriminations. Ce collectif de chercheurs tend, selon nous, vers l'idéal durkheimien du groupe professionnel cohésif (Durkheim, 2020 [1902]) qui doit réguler son activité car cette dernière peut être déstabilisatrice de l'ordre social. C'est aussi au sein du groupe professionnel que doivent se constituer la morale et le droit professionnels.

Si les catégories du racisme sont constamment questionnées des deux côtés de l'Atlantique, ces dernières années les notions de racisme systémique et celle de racisme institutionnel ont fait un retour remarqué

dans les travaux académiques et les mobilisations politiques antiracistes. Daniel Sabbagh (2022) a récemment questionné l'« extension du domaine du racisme » car le racisme concerne des entités et des objets de différente nature mais rassemblés sous la même étiquette. Est-ce une confusion ou une extension liée au caractère multiforme de ce phénomène ? C'est particulièrement le racisme systémique que Sabbagh pointe du doigt comme une catégorie « attrape-tout » dont les effets les plus ambigus sont « la réification et l'homogénéisation tendancielles des groupes raciaux ». La notion de racisme institutionnel, quant à elle, forgée aux États-Unis par des militants des droits civiques appartenant au mouvement politique afro-américain Black Power, est très discutée ou même décriée dans les mondes politiques et académiques français. Ses détracteur·rices avancent qu'à la différence des États-Unis de la ségrégation raciale avec les lois Jim Crow, ou de la France des lois sur le statut des Juifs du régime de Vichy, la France d'aujourd'hui n'institutionnalise pas les discriminations et l'exclusion fondée sur la race. Certaines relations et interactions sociales, ainsi que certains mécanismes et comportements sociaux, font pourtant système, à en croire la surreprésentation de populations ayant des origines ethniques bien précises (souvent issues des immigrations africaines) dans des quartiers dits défavorisés, cumulant un certain nombre de difficultés sociales et économiques et ne connaissant qu'une faible mobilité sociale (Simon, 2017). Si des corrélations statistiques peuvent être établies, cela ne nous dit rien du poids de la variable raciale dans la reproduction des inégalités, ni sur le caractère racial des inégalités constatées. Voilà pourquoi les partisans de la notion de racisme institutionnel mobilisent les survivances d'un « legs colonial » (Bayart et Bertrand, 2006). Certain·es vont jusqu'à dessiner une chaîne causale partant de la colonisation jusqu'aux immigrations postcoloniales pour expliquer des inégalités et des rapports sociaux en France désavantageant les populations issues des anciens pays colonisés[25].

25. Nous discutons dans l'un de nos écrits les conséquences d'une telle lecture dans un contexte militant, et ses effets sur les possibilités de penser l'émancipation, pour les descendant·es d'immigré·es d'anciennes colonies françaises (Tighanimine, 2021).

Ce qui est sûr, c'est que ces legs culturels, historiques et politiques se retrouvent dans la sous-représentation de populations « subalternes » dans les jeux de données des pays qui conçoivent, produisent et fabriquent l'intelligence artificielle[26]. Ils conduisent ainsi nos systèmes numériques à surreprésenter les catégories dominantes et invisibiliser ou discriminer, par les processus que nous venons de décrire (vote, classification…), les populations (colonisées, racisées, féminines) que nos histoires statistiques depuis le XIX^e^ siècle ont systématiquement moins représentées.

Remerciements : Je tiens à remercier Isabelle Berrebi-Hoffmann, co-directrice de thèse, pour nos nombreux échanges lors de la construction des arguments de cet article.

26. Pays qui peuvent tout de même exploiter des travailleurs·euses (du clic) pauvres, « subalternes » du Sud global pour modérer les contenus des réseaux sociaux ou entraîner des IA. Voir par exemple l'enquête récente menée par *Time* sur les travailleurs·euses Kényan·es payé·es deux dollars de l'heure pour rendre ChatGPT « moins toxique » (Perrigo, 2023).

Références bibliographiques

Angwin, Julia, Larson, Jeff, Mattu, Surya et Kirchner, Lauren, 2016, « Machine bias. There's software used across the country to predict future criminals. And it's biased", https://www.propublica.org, 23 mai : <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.

Arrow, Kenneth J., 2015, « The theory of discrimination », dans Albert Rees et Orley Ashenfelter, *Discrimination in Labor Markets*, Princeton, Princeton University Press, p. 1-33.

Bayart, Jean-François et Bertrand, Romain, 2006, « De quel "legs colonial" parle-t-on ? », *Esprit*, décembre, p. 134-160.

Bender, Emily M., Gebru, Timnit, McMillan-Major, Angelina et Mitchell, Margaret, 2021, « On the dangers of stochastic parrots: Can language models be too big? », dans *Proceedings of the 2021 ACM conference on fairness, accountability, and transparence*, New York, Association for Computing Machinery, p. 610-623.

Benjamin, Ruha, 2019, *Race After Technology : Abolitionist Tools for the New Jim Code*, Medford, Polity Press.

Berrebi-Hoffmann, Isabelle et Chapus, Quentin, 2022, « Des luttes éthiques aux luttes sociales », *Réseaux,* nº 231, p. 71-107.

Berrebi-Hoffmann, Isabelle et Lallement, Michel, 2009, « À quoi servent les experts ? », *Cahiers internationaux de sociologie*, nº 126,p. 5-12 : <https://doi.org/10.3917/cis.126.0005>

Boullier, Dominique, 2020, *Comment sortir de l'emprise des réseaux sociaux,* Paris, Le Passeur.

Boullier, Dominique et El-Mhamdi, El-Mahdi, 2020, « Le *machine learning* et les sciences sociales à l'épreuve des échelles de complexité algorithmique », *Revue d'anthropologie des connaissances*, nº 14 : <https://doi.org/10.4000/rac.4260>

Brian, Éric, 2002, « Transactions statistiques au XIXe siècle : mouvements internationaux de capitaux symboliques », *Actes de la recherche en sciences sociales,* nº 145, p. 34-46.

Brun, Solène et Cosquer, Claire, 2022, *Sociologie de la race*, Paris, Armand Colin.

Buolamwini, Joy et Gebru, Timnit, 2018, « Gender shades: Intersectional accuracy disparities in commercial gender classification », *Conference on fairness, accountability and transparency. Proceedings of Machine Learning Research,* vol. 81, p. 1-15 : <https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf>.

Castelnovo, Alessandro *et al.*, 2022, « A clarification of the nuances in the fairness metrics landscape », *Scientific Reports,* nº 12, art. 4209 : <https://www.nature.com/articles/s41598-022-07939-1>.

Corbett-Davis, Sam, Pierson, Emma, Feller, Avi et Goel, Sharad, 2016, « A computer program used for bail and sentencing decisions was labeled biased against blacks:It's actually not that clear », *washingtonpost.com,* 17 octobre : <https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-

an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/>.

D'Ignazio, Catherine et Klein, Lauren F., 2020, *Data feminism,* Cambridge, MIT press.

Dastin, Jeffrey, 2018, « Amazon scraps secret AI recruiting tool that showed bias against women », *reuters.com,* 11 octobre : <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G>.

Desrosières, Alain et Thévenot, Laurent, 1988, *Les catégories socioprofessionnelles,* Paris, La Découverte.

Dhume, Fabrice, 2016, « Du racisme institutionnel à la discrimination systémique ? Reformuler l'approche critique », *Migrations Société,* n° 163, p. 33-46.

Didier, Emmanuel, 2009, *En quoi consiste l'Amérique ? Les statistiques, le New Deal et la démocratie,* Paris, La Découverte.

Durkheim, Émile, 1898, « Représentations individuelles et représentations collectives », *Revue de Métaphysique et de Morale,* t. VI, mai, p. 273-302.

–, 2020 [1902], « Quelques remarques sur les groupements professionnels », dans id., *Sociologie politique. Une anthologie*, éditée par Florence Hulak, Paris, Presses Universitaires de France, p. 55-90.

Ekstrand, Michael D., Riedl, John, T. et Konstan, Joseph, A., 2011, *Collaborative Filtering Recommender Systems*, Boston et Delft, Now Publishers.

El-Mhamdi, El-Mahdi, 2020, *Robust Distributed Learning*, thèse n° 7218, Lausanne, EPFL : <https://doi.org/10.5075/epfl-thesis-7218>.

El-Mhamdi, El-Mahdi, Farhadkhani, Sadegh, Guerraoui, Rachid, Gupta, Nirupam, Hoang, Lê-Nguyên, Pinot, Rafael, Rouault, Sébastien et Stephan, John, 2022, « On the impossible security of large AI models", *arXiv* 2209.15259 : <https://doi.org/10.48550/arXiv.2209.15259>.

Eubanks, Virginia, 2018, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*, New-York, St. Martin's Press.

Hao, Karen, 2021, « She risked everything to expose Facebook. Now she's telling her story », *technologyreview.com,* 29 juillet : <https://www.technologyreview.com/2021/07/29/1030260/facebook-whistleblower-sophie-zhang-global-politicalmanipulation/>.

Harwarth, Irene, DeBra, Elizabeth et Maline, Mindi, 1997, *Women's Colleges in the United States: History, Issues, and Challenges,* Washington, National Institute on Postsecondary Education, Libraries, and Lifelong Learning, U.S. Dept. of Education.

Hauben, Michael et Hauben, Ronda, 1997, *Netizens: On the History and Impact of Usenet and the Internet*, Los Alamitos, IEEE Computer Society Press.

Hoang, Lê-Nguyen et El-Mhamdi, El-Mahdi, 2019, *Le fabuleux chantier. Rendre l'intelligence artificielle robustement bénéfique,* Les Ulis, EDP Sciences.

Hoang, Lê-Nguyen, Faucon, Louis et El-Mhamdi, El-Mahdi, 2021, « Recommendation algorithms, a neglected opportunity for public health », *Revue Médecine et Philosophie*, nº 4, p. 16-24 : <https://philarchive.org/archive/HOARAA>.

Huszár, Ferenc *et al.*, 2022, « Algorithmic amplification of politics on Twitter », *Proceedings of the National Academy of Sciences* vol. 119, nº 1 : <https://doi.org/10.1073/pnas.2025334119>.

Jodelet, Denise, 1989, *Les représentations sociales,* Paris, Presses universitaires de France.

Johnson, Khari, 2020, « Timnit Gebru: Google's "dehumanizing" memo paints me as an angry Black woman », *venturebeat.com*, 10 décembre : <https://venturebeat.com/business/timnit-gebru-googles-dehumanizing-memo-paints-me-as-an-angry-black-woman/>.

Laufer, Benjamin et al., 2022, « Four years of FAccT: A reflexive, mixed-methods analysis of research contributions, shortcomings, and future prospects », *FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency,* juin, p. 401-426 : <https://doi.org/10.1145/3531146.3533107>.

Lorenz, Taylor, 2023, « Internet "algospeak" is changing our language in real time, from "nip nops" to "le dollar bean" », *washigtonpost.com,* 8 avril : <https://www.washingtonpost.com/technology/2022/04/08/algospeak-tiktok-le-dollar-bean>.

Matamoros-Fernández, Ariadna, 2017, « Platformed racism: The mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube », *Information, Communication & Society*, vol. 20, nº 6, p. 930-946.

McGuffie, Kris et Newhouse, Alex, 2020, « The radicalization risks of GPT-3 and advanced neural language models », *arXiv* : 2009.06807 : <https://doi.org/10.48550/arXiv.2009.06807>.

Moscovici, Serge, 1961, *La psychanalyse. Son image et son public,* Paris, Presses universitaires de France.

Noble, Safiya Umoja, 2018, *Algorithms of oppression. How Search Engines Reinforce Racism,* New York, New York University Press.

Noble, Safiya Umoja et Tynes, Brendesha M., 2016, *The Intersectional Internet: Race, Sex, Class, and Culture Online,* New York, Peter Lang International Academic Publishers.

Pasquale, Frank, 2015, *The Black Box Society: The Secret Algorithms That Control Money and Information,* Cambridge, Harvard University Press.

Pazzani, Michal J. et Billsus, Daniel, 2007, « Content-based recommendation systems », dans Peter Brusilovsky, Alfred Kobsa, Wolfgang Nejdl (dir.), *The Adaptive Web*, Berlin et Heidelberg, Springer, p. 325-341.

Perrigo, Billy, 2023, « Exclusive: OpenAI used Kenyan workers on less than $2 per hour to make ChatGPT less toxic », *time.com,* 18 janvier : <https://time.com/6247678/openai-chatgpt-kenya-workers/>.

Sabbagh, Daniel, 2022, « Le "racisme systémique" : un conglomérat problématique », *Mouvements,* vol. 2, numéro hors série, p. 56-74.

Salais, Robert, Baverez, Nicolas et Reynaud, Bénédicte, 1986, *L'invention du chômage,* Paris, Presses universitaires de France.

Schiffer, Zoe, 2021, « Timnit Gebru was fired from Google – then the harassers arrived", *theverge.com,* 5 mars : <https://www.theverge.com/22309962/timnit-gebru-google-harassment-campaign-jeff-dean>.

Schor, Paul, 2009, *Compter et classer. Histoire des recensements américains,* Paris, Éditions de l'EHESS.

Schradie, Jen, 2019, *The Revolution That Wasn't: How Digital Activism Favors Conservatives,* Cambridge, Harvard University Press.

Simon, Patrick, 2017, « Les descendants d'immigrés et la question de l'intégration », *Regards croisés sur l'économie,* nº 20, p. 81-92.

Solsman, Joan E., 2018, « YouTube's AI is the puppet master over most of what you watch », *cnet.com,* 10 janvier : <https://www.cbsnews.com/news/ces-youtubes-ai-is-the-puppetmaster-over-what-you-watch/>.

Stavo-Debauge, Joan, 2008, « Faut-il s'en remettre aux pouvoirs de la statistique pour agir contre les discriminations et réaliser le droit ? La catégorisation ethnique et raciale en question au Royaume-Uni et en France », dans Antoine Lyon-Caen et Adalberto Perulli (dir.), *Efficacia e diritto del lavoro,* Padoue, Cedam, p. 163-194.

–, 2011, « En quête d'une introuvable action antidiscriminatoire. Une sociologie de ce qui fait défaut », *Politix,* vol. 94, nº 2, p. 81-105.

Stokely, Carmichael et Hamilton, Charles V., 2009 [1967], *Le Black Power. Pour une politique de libération aux États-Unis,* Paris, Payot.

Tighanimine, Mariame, 2021, *Dévoilons-nous. Manifeste antiraciste et féministe*, Paris, Éditions de l'Olivier.

–, 2022, *Notre histoire de France*, Paris, Stock.