Algorithmes : prévenir l'automatisation des discriminations

En partenariat avec la CNIL

CNIL.
COMMISSION NATIONALE
INFORMATIQUE & LIBERTÉS

Face au droit, nous sommes tous égaux

Défenseurdesdroits
RÉPUBLIQUE FRANÇAISE

Propos liminaires

Dans la crise sanitaire inédite que le monde traverse, l'usage des outils numériques a connu une accélération et une diversification sans précédent en suscitant des débats majeurs. Ces outils numériques reposent souvent sur des algorithmes sans que l'on en soit toujours conscient ou informé.

Le recours à des algorithmes comme support à la décision privée ou publique n'est pas nouveau : le calcul automatisé de l'appréciation du risque financier par les banques (« *scoring* ») qui consiste à combiner divers critères tirés des renseignements fournis par les demandeurs de prêt s'est ainsi généralisé depuis des décennies. Mais l'utilisation intensive des algorithmes, grâce à la nouvelle puissance de calcul des ordinateurs et à l'exploitation massive de données désormais très nombreuses, constitue, comme le relève le Conseil d'Etat, un « tournant inédit »[1].

En quelques années, les usages des algorithmes se sont diversifiés dans le secteur privé comme au sein des administrations[2]. On retrouve aujourd'hui de tels procédés dans des domaines aussi essentiels pour les individus que l'accès aux prestations sociales[3], la police et la justice[4], le fonctionnement des organisations telles que les hôpitaux, l'accès aux services publics ou encore les procédures d'embauche[5].

Depuis 2006, les technologies de l'apprentissage automatique (« *machine learning* ») connaissent un véritable essor. Une fois déployés, ces systèmes apprenants évoluent sans cesse pour se perfectionner.

Ces évolutions technologiques, qui se poursuivent, constituent des sources indéniables de progrès pour les individus et la société en permettant de produire des résultats rapides, plus fiables et individualisés et des analyses inédites sur de nombreux terrains.

Toutefois, **la CNIL et le Défenseur des droits ont, chacun dans leur domaine de compétences, déjà fait part de leurs préoccupations quant à l'impact de certains systèmes algorithmiques sur les droits fondamentaux**[6].

C'est dans cette perspective que s'inscrit aujourd'hui le Défenseur des droits, en souhaitant, en partenariat avec la CNIL, mettre en lumière **les risques considérables de discrimination que peuvent faire peser sur chacun et chacune d'entre nous l'usage exponentiel des algorithmes dans toutes les sphères de notre vie**.

Ce sujet est longtemps resté un angle mort du débat public. Il ne doit plus l'être.

[1] Conseil d'Etat, *Puissance publique et plateformes numériques : accompagner « l'ubérisation »*, La documentation française, 2017, p. 59.

[2] Voir par exemple, Direction interministérielle du numérique et du système d'information et de communication de l'Etat, DINSIC, Guide des algorithmes publics 2019.

[3] Délégation Nationale à la Lutte contre la Fraude, *Le « data mining », une démarche pour améliorer le ciblage des contrôles*, Paris, le 14 janvier 2014.

[4] Soraya Amrani Mekki, « Justice prédictive et accès au juge », *La Justice Prédictive*, Actes du Colloque du 12 février 2018 organisé par l'ordre de avocats au Conseil d'Etat et à la Cour de cassation à l'occasion de son Bicentenaire en partenariat avec l'Université Paris-Dauphine PSL, Paris, Dalloz, 2018.

[5] Christine Bargain, Marie Beaurepaire, Dorothée Prud'homme, *Recruter avec des algorithmes ? Usages, opportunités et risques*, AFMD, 2019.

[6] CNIL, Travaux sur le système APB (décision n°2017-053 du 30 août 2017) ; *Comment permettre à l'Homme de garder la main ?* Rapport sur les enjeux éthiques des algorithmes et de l'intelligence artificielle, 15 décembre 2017. Défenseur des droits, Guide - Recruter avec des outils numériques sans discriminer publié en 2015, Avis n°15-25 du 1er décembre 2015 sur la sécurité dans les gares ; *Rapport Lutte contre la fraude aux prestations sociales : à quel prix pour les droits des usagers ?*, septembre 2017, Décisions Parcoursup (2018-323 du 21 décembre 2018 et 2019-21 du 18 janvier 2019), Avis 18-26 du 31 octobre 2018 sur le projet de loi de programmation et de réforme pour la justice, avis 19-11 du 5 septembre 2019 sur le projet de loi bioéthique.

Pourquoi les algorithmes peuvent être discriminatoires ?

À première vue, les algorithmes permettent de trier, classer ou d'ordonner des informations en se débarrassant des préjugés et biais propres aux affects des humains. Ils seraient ainsi plus à même de réaliser l'égalité de traitement attendue en appliquant les mêmes critères et pondérations quelle que soit, par exemple, l'origine ou l'orientation sexuelle du demandeur.

En réalité, il n'y a ni magie technologique ni neutralité mathématique : les algorithmes sont conçus par des humains et à partir de données reflétant des pratiques humaines. Ce faisant, **des biais peuvent être ainsi intégrés à toutes les étapes de l'élaboration et du déploiement des systèmes : dès l'intention qui préside à l'élaboration de l'algorithme en amont, pendant la réalisation du code informatique, celle du code exécutable, celle de l'exécution, celle du contexte d'exécution et celle de la maintenance**[7].

Certains biais, tout à fait intentionnels, peuvent être issus de l'intégration dans un algorithme d'un critère de discrimination interdit. Certains motifs peuvent être pris en compte parmi les critères de l'algorithme dans certains contextes spécifiques tels que l'état de santé pour l'assurance, l'âge pour le prêt bancaire ou encore le lieu de résidence pour moduler des primes, si leur utilisation est jugée proportionnée à un finalité légitime[8]. En revanche, les critères tels le sexe ou l'origine ne sauraient, quel que soit le contexte, constituer des critères légaux.

Les effets discriminatoires des algorithmes reposent toutefois le plus souvent sur des mécanismes moins visibles que l'intégration d'un critère prohibé bien identifiable dans l'algorithme.

Des données biaisées

La mécanique discriminatoire repose, fréquemment, sur les biais des données sélectionnées et utilisées par l'algorithme fermé ou qui nourrissent l'algorithme apprenant dans sa phase d'apprentissage puis ultérieurement.

L'un des biais fréquents repose sur le **manque de représentativité des données mobilisées**. Par exemple, une étude a permis en 2018 d'expliquer pourquoi certains systèmes de reconnaissance faciale, qui reposent sur des techniques d'apprentissage[9], rencontraient de plus grandes difficultés pour identifier les femmes, les personnes non blanches et davantage encore les femmes de couleur, en produisant un taux d'erreur significatif pour ces populations : le stock de données sur lesquelles le modèle s'appuyait était marqué par une très forte prédominance des visages masculins et blancs[10]. Le problème peut être voisin pour les technologies d'identification vocale : faute d'avoir pensé aux femmes, à leur voix et d'avoir été construit (et donc alimenté par des données « féminines ») et testé en ce sens, le système fonctionne moins bien pour le personnel féminin[11].

Les données intégrées aux systèmes algorithmiques ou mobilisées pour entraîner un système d'apprentissage automatique peuvent être par ailleurs biaisées lorsqu'elles sont **la traduction mathématique de pratiques et comportements passés souvent discriminatoires et des discriminations systémiques opérant au sein de la société**.

[7] Barocas S., Selbst et Andrew D. « Big data's disparate impact », California Law Review, June 2016 Vol. 104, n°3, pp.671-732.

[8] C.E., 30 octobre 2001, n° 204909, *association française des Stés financières*. Voir l'article «Testing, scoring, ranking... », Revue trimestrielle de droit civil, juillet-septembre 2002, n° 3, p. 498.

[9] CNIL, Reconnaissance faciale. Pour un débat à la hauteur des enjeux, 15 novembre 2019.

[10] Selon l'étude de Joy Buolamwini, chercheuse du MIT, le taux d'erreur du Logiciel Rekognition d'Amazone est de 1% pour les hommes de peaux claire, 7% pour les femmes de peau claire, 12% pour les hommes de couleur, 35% pour les femmes de couleur. Voir Hardesty, Larry. "Study Finds Gender and Skin-Type Bias in Commercial Artificial-Intelligence Systems." *MIT News*, February 11, 2018.

[11] Dans le cadre d'un stage centré sur la reconnaissance vocale des pilotes d'hélicoptère, lorsqu'une femme a été mise aux commandes, le système a moins bien fonctionné, ce qui constitue un problème crucial pour la sécurité (TUAL M., « La diversité humaine est un enjeu central pour le développement de l'intelligence artificielle », *Le Monde*, 30/07/2018).

Dans les données d'emploi disponibles, les femmes sont moins représentées et tendent à occuper certaines filières de métiers et des postes et rémunérations moindres. Sur la base de telles données, un algorithme pourrait déduire que les femmes ne sont pas aussi productives que les hommes et n'accèdent pas autant à des postes à responsabilité. En conséquence, un algorithme utilisé pour le recrutement utilisant des données biaisées reproduira ces biais, voir les accentuera[12].

Fausse neutralité des algorithmes et vrais effets discriminatoires

La mobilisation de critères neutres en apparence c'est-à-dire ne relevant pas des motifs prohibés de discriminations, peut avoir des effets discriminatoires comme l'a souligné le Défenseur des droits dans sa décision Parcoursup[13]. En l'occurrence, la prise en compte du critère apparemment neutre de l'établissement d'origine par les algorithmes des universités pourrait conduire, indirectement, à discriminer les jeunes d'origine immigrée, compte tenu de la forte ségrégation résidentielle et scolaire notamment observée en Ile-de-France.

Le plus souvent, c'est la combinaison de plusieurs critères neutres qui peut emporter ces effets discriminatoires. Les critères et données en question peuvent même nous sembler très lointains des motifs prohibés, mais leur corrélation permet des résultats similaires à ceux qu'on aurait obtenu en utilisant directement la caractéristique protégée. Les algorithmes apprenants, et les multiples corrélations de données massives qu'ils permettent, peuvent facilement générer de tels effets. Dans ce cas, l'appartenance à une catégorie protégée se retrouve encodée dans des données « neutres ».

Le programme, élaboré de sorte qu'il maximise ses capacités à trouver des caractéristiques similaires dans une masse de données, recréé un ensemble correspondant à la catégorie protégée, et lui applique un traitement spécifique.

Afin de cibler sa publicité, la société de supermarchés américaine Target a ainsi mis au point un modèle prédictif qui repère les clientes enceintes à partir de leurs habitudes d'achat sur 25 produits[14]. De tels modèles pourraient être utilisés à des fins discriminatoires, ou avoir des effets discriminatoires.

Les algorithmes peuvent combiner plusieurs sources de biais et venir ruiner les meilleures intentions. Plusieurs études américaines ont démontré récemment le caractère discriminatoire des principaux modèles de systèmes dits « intelligents » chargés de détecter automatiquement les propos haineux pour les modérer : la probabilité de voir son message signalé comme offensant ou haineux par le système était 1,5 fois plus élevée pour les internautes afro-américains. Ces biais sont issus des données d'apprentissage : le panel de données a été conçu par des humains qui ont, en amont, classé comme offensant ou haineux les messages marqués par les grossièretés. Ils sont aussi renforcés par les limites techniques du système qui peine à identifier les nuances d'une langue et à contextualiser les propos relevant de l'argot ou de l'ironie par exemple[15]. De tels risques ne peuvent pas être ignorés à l'heure où, en France, les plateformes disposant dorénavant de seulement 24h pour retirer des propos haineux signalés, vont donc utiliser pleinement des procédés algorithmiques[16].

[12] EEOC, Conference - Big Data in the Workplace: Examining Implications for Equal Employment Opportunity Law, 13 octobre 2016.

[13] Décision 2019-021 du 18 janvier 2019 relative au fonctionnement de la plateforme nationale de préinscription en première année de l'enseignement supérieur (Parcoursup).

[14] Voir KIM PT, « Data-driven discrimination at work », 58 Wm. and Mary Law Review 857, 2016.

[15] "The algorithms that detect hate speech online are biased against black people", 15 août 2019, vox.com.

[16] Proposition de loi visant à lutter contre les contenus haineux sur internet adopté définitivement par l'Assemblée nationale le 13 mai 2020 et destinée à retirer certains contenus haineux sous 24 h des réseaux sociaux, des plates-formes collaboratives et des moteurs de recherche.

Des discriminations invisibles et potentiellement massives

Les effets discriminatoires des algorithmes ne sont bien souvent mesurables par les chercheurs qu'à l'échelle des groupes. Ils risquent de rester tout à fait invisibles pour les victimes.

De plus, alors que les biais cognitifs d'un être humain varient en fonction des circonstances et se transcrivent de manière contingente en pratiques discriminatoires, **les biais discriminatoires intégrés par un algorithme s'appliquent de manière automatique et pourraient systématiser les discriminations**. Il existe un risque majeur de renforcer « essentialisation » et « stéréotypes » car le caractère prédictif de l'algorithme est basé sur le comportement ou les caractéristiques homogénéisées de groupes. Ces systèmes risquent ainsi « de renforcer les discriminations en leur donnant une apparence d'objectivité »[17].

Si les effets discriminatoires de l'algorithme ne seront pas toujours repérables à l'échelle individuelle, le système algorithmique, neutre en apparence, peut produire des discriminations pour les groupes sociaux protégés, qui se traduiront par exemple par un moindre accès aux biens recherchés ou à un taux d'erreur plus important du système les concernant. Ce risque de discrimination est d'autant plus grand pour les groupes sociaux qui font déjà l'objet de discriminations systémiques majeures au sein de la société, par exemple, les femmes, les personnes en situation de handicap ou les personnes issues de l'immigration.

Intégrant des pratiques discriminatoires antérieures dans le cadre du jeu de données mobilisées pour son apprentissage, **les biais des systèmes dits « intelligents » tendent à se renforcer au fil de leur déploiement**.

Le logiciel Predpol permet à de nombreuses forces de police d'orienter leurs interventions et de « rationaliser » l'activité policière en déterminant des "points chauds" où le risque d'infraction est le plus élevé, afin de renforcer les patrouilles. Ce modèle prend aussi en compte des facteurs d'influence comme la densité de population, la présence à proximité de bars, ou encore l'existence de moyens de transport. Toutefois, la prédominance des informations relatives aux lieux d'occurrence des délits et crimes passés est problématique. Aux Etats-Unis comme dans d'autres pays, les contrôles, les arrestations et les lieux où la police décide de patrouiller visent bien davantage les minorités et certains territoires. Sur suggestion de Predpol, les forces de police se rendront en majorité dans ces quartiers et y constateront de nouvelles infractions, venant ainsi approvisionner la base d'apprentissage de nouvelles données biaisées. Les algorithmes peuvent ainsi former des boucles de rétroaction par lesquelles stéréotypes, discriminations et inégalités se confortent mutuellement, contribuant ainsi à cristalliser durablement des situations d'inégalité[18]. **Seul un contrôle précis et régulier des résultats de l'algorithme apprenant permettra de s'assurer que l'algorithme ne devient pas discriminatoire au fil des encodages successifs**.

Enfin, il faut ajouter que **ces systèmes tendent à davantage cibler et contrôler, et ce faisant stigmatiser, les membres des groupes sociaux déjà défavorisés et dominés**[19]. Plusieurs associations ont ainsi lancé en 2019 une action en justice contre l'Etat néerlandais afin de faire établir l'illégalité d'un algorithme mis au point par le Ministère des affaires sociales et de l'emploi pour prévoir la probabilité qu'un individu s'engage dans la fraude aux prestations et à l'impôt.

[17] Dunja Mijatovic, Commissaire aux droits de l'Homme, « Protéger les droits de l'Homme à l'ère de l'intelligence artificielle », Carnet des droits de l'Homme du Commissaire, Strasbourg, 3 juillet 2018.

[18] Hiring by Algorithm: predicting and Preventing disparate impact - Ifeoma Ajunwa, Sorelle Freidler, Carlos Scheidegger, Suresh Venkatasubramanian ; Draft de janvier 2016.

[19] Virginia Eubanks, *Automating inequalities. How High-tech tools profiles, police, and punish the Poor* ; St. Martin's Press, janvier 2018.

Au cours de l'audience, le gouvernement a reconnu que cet algorithme visait les quartiers comptant un plus grand nombre de bénéficiaires de l'aide sociale, malgré l'absence d'éléments démontrant que ces quartiers sont responsables de taux plus élevés de fraude aux prestations[20].

Recommandations

Le droit de la non-discrimination doit être effectivement respecté en toutes circonstances, y compris quand une décision implique le recours à un algorithme.

La mobilisation extensive des algorithmes constitue, pour reprendre l'expression de Cathy O'Neil, une « bombe à retardement » au regard des enjeux d'égalité[21]. Néanmoins, malgré les premières alertes du Rapport Villani[22] et quelques initiatives[23], la prise de conscience tarde à émerger en France : **les concepteurs d'algorithmes, comme les organisations achetant et utilisant ce type de systèmes, n'affichent pas la vigilance nécessaire pour éviter une forme d'automatisation invisible des discriminations**.

Pourtant, le principe de loyauté, qui place la notion d'« intérêt des utilisateurs » comme obligation du responsable de l'algorithme, tout comme le principe de vigilance et de réflexivité qui organise un questionnement régulier, méthodique et délibératif à l'égard des objets apprenants, devraient guider la réflexion et l'action[24].

Faut-il le rappeler, **la non-discrimination n'est pas une option mais renvoie à un cadre juridique** qui prévoit une grille d'analyse permettant l'identification des inégalités de traitement, afin de mettre en œuvre un droit fondamental, celui de ne pas être discriminé.

Les organisations qui utilisent des algorithmes ne sauraient échapper à leurs responsabilités sous couvert d'ignorance, d'incompétence technologique ou d'opacité des systèmes. **Les biais algorithmiques doivent pouvoir être identifiés puis corrigés et les auteurs de décisions discriminatoires issues de traitement algorithmiques doivent pouvoir être sanctionnés.**

Comme le souligne la littérature existante, le manque de transparence des systèmes mis en œuvre et les corrélations de données permises par les algorithmes, souvent de manière totalement invisible, rendent les protections offertes par le droit incertaines voire inefficaces.

Ainsi, comment exercer son droit au recours quand on ne sait même pas que l'on a été victime d'une discrimination liée à un algorithme, que l'organisation utilisatrice de l'algorithme elle-même n'en a pas conscience, que le concepteur de l'algorithme ne veut ou ne peut expliquer comment fonctionne cet outil ? Comment savoir qu'un algorithme discrimine tel ou tel groupe social ? Et dès lors, comment sanctionner ces atteintes aux droits ? Les travaux menés avec nos homologues européens dans le cadre du réseau Equinet[25] comme le séminaire pluridisciplinaire « Algorithmes, biais et lutte contre les discriminations organisé les 28

[20] Open Democracy, « Welfare surveillance on trial in the Netherlands », 8 November 2019. La Cour de la Haye a rendu un jugement le 5 février 2020 reconnaissant que le gouvernement avait violé le droit à la vie privée et familiale prévu par l'article 8 de la CEDH et a ordonné d'arrêter d'utiliser cet algorithme. Les juges ont appuyé leur jugement sur le fait que l'algorithme Syri manquait de transparence. La Cour n'a pas répondu sur le terrain de la possible violation de l'art. 22 du RGPD qui prévoit l'interdiction de certaines décisions automatisées.

[21] Cathy O'Neil, *Algorithmes. La bombe à retardement*, Les Arènes, 2018 (USA, 2016).

[22] Cédric Villani, *Donner un sens à l'intelligence artificielle : pour une stratégie nationale et européenne*, Rapport au Gouvernement, 8 mars 2018.

[23] Telecom Paris Tech, *Algorithmes : biais, discrimination et équité*, février 2019 ; Aude Bernheim, Flora Vincent, *L'intelligence artificielle, pas sans elles !*, Laboratoire de l'égalité, éditions Belin, 2019 ; Institut Montaigne, *Rapport Algorithmes : contrôle des biais SVP*, mars 2020 ; Rapport collectif sur commande de la mission Etalab, *Ethique et responsabilité des algorithmes publics*, ENA, Promotion 2018-2019 « Molière », Juin 2019.

[24] 40e Conférence mondiale des autorités de protection des données, Déclaration sur l'éthique et la protection des données, 23 octobre 2018.

[25] Equinet, *Regulating for an equal AI : A New Role for Equality Bodies. Meeting the new challenges to equality and non-discrimination from increased digitisation and the use of Artificial Intelligence*, juin 2020.

Défenseur des droits

TSA 90716 - 75334 Paris Cedex 07

Tél. : 09 69 39 00 00

www.defenseurdesdroits.fr

Toutes nos actualités :

www.defenseurdesdroits.fr

Défenseurdesdroits

RÉPUBLIQUE FRANÇAISE