

QUELLES MESURES POUR UNE UTILISATION NON-DISCRIMINATOIRE DE L'INTELLIGENCE ARTIFICIELLE & DES ALGORITHMES ?

Nelly CHATUE-DIOP

Chief Data Officer | **Betclik Group**

Soumia MALINBAUM

Vice-Présidente Business Development | **Keyrus Group**

La démocratisation des algorithmes d'intelligence artificielle (Machine Learning, Deep Learning, NLP, etc.) permet désormais à toutes les entreprises d'accéder aux bénéfices de l'intelligence artificielle. Attention toutefois à l'effet « boîte noire » et aux biais potentiels de ces outils qui, en reproduisant des stéréotypes, peuvent non seulement aller à l'encontre des lois en vigueur¹, mais aussi des intérêts économiques de l'entreprise, de ses engagements éthiques et de sa politique en faveur de la diversité.

Il y a encore une dizaine d'années, l'intelligence artificielle était une discipline scientifique, cantonnée aux laboratoires de recherche, et ses applications paraissaient ne pas devoir concerner de sitôt le grand public et le commun des entreprises. L'explosion actuelle des cas d'usage et la multiplication des solutions basées sur des algorithmes d'intelligence artificielle ouvrent indiscutablement aux entreprises de nouvelles perspectives de gain d'efficacité, de productivité et de création de valeur. Mais l'engouement actuel pour ces technologies et les opportunités qu'elles représentent doivent être tempérés par la mise en évidence, dans certaines applications, de biais ayant pour conséquence des discriminations et des atteintes aux droits fondamentaux des personnes.

UNE INDISPENSABLE PRISE DE CONSCIENCE

Tout le monde a entendu parler de ce logiciel de prédiction de récidive utilisé par les juges américains qui pénalisait les populations afro-américaines et, plus récemment, de l'algorithme d'Apple Pay Card² accordant un plafond de crédit plus élevé aux hommes qu'aux femmes, malgré des revenus équivalents. Ces discriminations racistes et sexistes n'étaient pas volontaires. Leur mise au jour n'en a pas moins jeté méfiance et discrédit sur des solutions fondamentalement conçues pour accélérer des processus et optimiser des prises de décision – notamment, et là est tout le paradoxe – en réduisant la part de subjectivité entrant dans tout arbitrage réalisé par des humains.

Fortement relayées par les médias, de telles révélations ont un impact négatif sur la réputation des organisations concernées.

1 | Loi du 27 mai 2008 portant diverses dispositions d'adaptation au droit communautaire dans le domaine de la lutte contre les discriminations. Au niveau européen, outre le RGPD, 5 directives fixent des garde-fous en matière de discrimination, que ce soit dans le monde du travail ou dans l'accès aux services.

2 | <https://www.bostonglobe.com/business/2019/11/11/apple-founder-steve-wozniak-says-goldman-apple-card-algorithm-discriminates-against-women/IKSjZDC5nuFhbAWSRD6s9H/story.html>

Les lois anti-discrimination s'appliquent aussi aux algorithmes

En France, la loi du 27 mai 2008 recense **25 critères de discrimination** : âge, sexe, origine, appartenance réelle ou supposée à une ethnie, appartenance ou non à une prétendue race, appartenance ou non à une nation, état de santé, grossesse, handicap, caractéristiques génétiques, orientation sexuelle, identité de genre, opinions politiques, activités syndicales, opinions philosophiques, religion, situation de famille, apparence physique, nom, mœurs, lieu de résidence, perte d'autonomie, vulnérabilité économique, domiciliation bancaire, capacité à s'exprimer en français.

L'utilisation de ces critères est prohibée dans les 7 situations suivantes : accès à l'emploi, rémunération, accès aux biens et services publics et privés, accès à un lieu accueillant du public, accès à la protection sociale, éducation et formation.

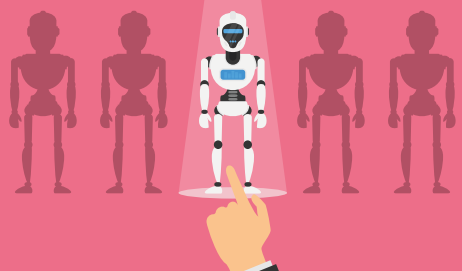


Elles ont cependant un triple mérite :

- Faire reculer l'idée trop bien installée d'une IA neutre et infaillible, intrinsèquement juste et équitable parce que reposant sur la science la plus objective qui soit – les mathématiques.
- Pousser les gouvernements et les instances de régulation à mettre l'accent, en regard des bénéfices attendus de l'IA, sur les implications sociétales et éthiques de ces technologies – notamment les « **risques potentiels, tels que l'opacité de la prise de décisions, la discrimination fondée sur le sexe ou sur d'autres motifs**, l'intrusion dans nos vies privées ou encore l'utilisation à des fins criminelles ». ³
- Alerter les entreprises sur la nécessité de s'assurer que les outils algorithmiques qu'elles développent ou utilisent sont exempts de biais pouvant induire des discriminations. Cette précaution et cette vigilance s'imposent **en particulier pour les applications à base de Machine Learning/Deep Learning utilisées dans le cadre de processus de recrutement, de gestion des carrières, de profilage des clients et, plus généralement, pour caractériser des personnes en vue de prendre une décision.**



POURQUOI L'IA N'EST NI MAGIQUE NI NEUTRE ?



Pour le comprendre, il faut revenir sur les spécificités des algorithmes de Machine Learning/Deep Learning/NLP, cœur des « intelligences artificielles » actuelles. La particularité de ces algorithmes est leur capacité à apprendre. C'est ce qui leur confère la capacité de se spécialiser afin de traiter des problématiques et des cas pour lesquels ils n'ont pas été explicitement programmés. **Le procédé n'est pas magique ! Il est statistique** : pour se spécialiser, l'algorithme apprend à partir d'un grand nombre d'exemples, souvent qualifiés ou « étiquetés » par des petites mains. Il identifie dans ces données d'apprentissage des invariants, des corrélations ou des récurrences qui lui permettent d'élaborer un modèle capable, dans un deuxième temps, de traiter correctement – c'est-à-dire avec une probabilité de réussite élevée – des données qu'il n'a jamais vues.

Quant à la neutralité que l'on attribue à ces outils, elle ne peut être que relative. En tant que succession d'instructions et de formules mathématiques, les algorithmes n'ont certes pas de volonté intrinsèque. En revanche, ils sont **toujours orientés par l'intention et la finalité de ceux qui les ont conçus, puis de ceux qui les utilisent**. L'IA – conçue par des humains, entraînée par des humains, évaluée par des humains et utilisée par des humains – embarque inévitablement **des biais, des partis pris, des préjugés humains que l'automatisation rend invisibles, mais qu'elle reproduit et amplifie**.

DES BIAIS OMNIPRÉSENTS MAIS INVISIBILISÉS

Ces systèmes algorithmiques que l'on voudrait « objectifs » comportent en réalité trois points de faiblesse – trois sources de biais ou d'opacité – qui peuvent conduire un modèle jugé performant à véhiculer des stéréotypes de genre ou d'origine débouchant sur des discriminations, positives ou négatives, à l'insu des développeurs et des utilisateurs.

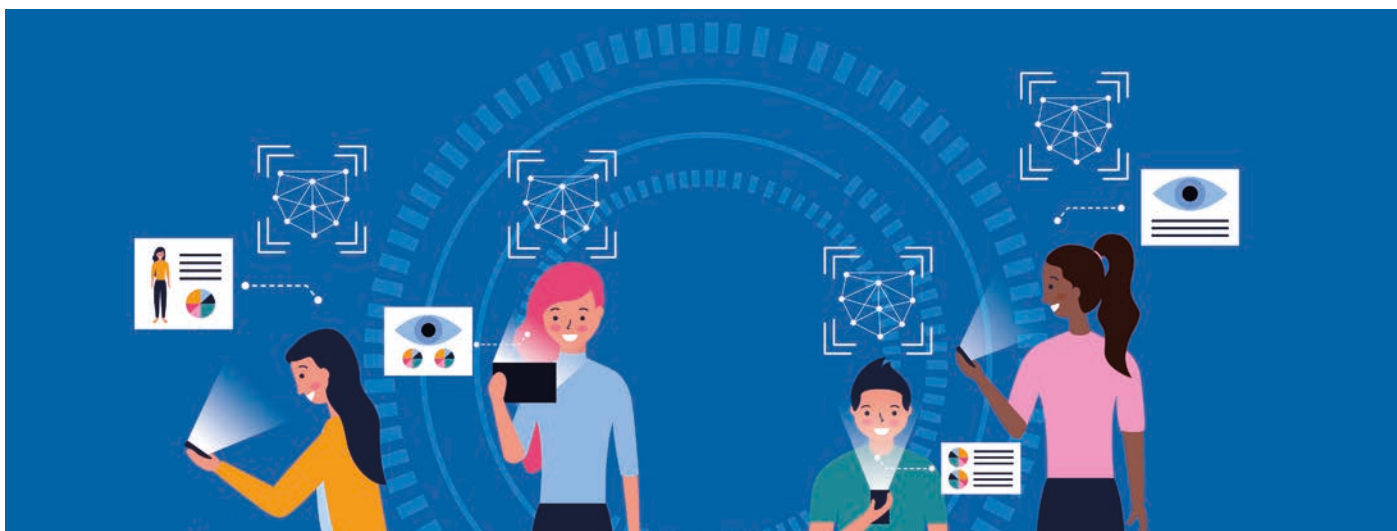
LES ALGORITHMES EUX-MÊMES

Les algorithmes d'apprentissage utilisés aujourd'hui par les développeurs d'applications ne sont pas, pour la plupart, créés sur mesure par ces derniers. Ils sont basés sur des « produits sur étagère ».

En libre accès, ces algorithmes génériques ont en grande majorité été développés par des scientifiques (universitaires ou travaillant dans les GAFAM) ayant pour priorité de valider la précision de leur modèle mathématique et d'éviter le biais de sur-apprentissage, et non de s'assurer de la généralisation en toute équité du modèle en question.

Non seulement **aucun de ces algorithmes n'a été conçu avec un objectif explicite de non-discrimination** mais, de plus, ils ont été **développés par une population singulièrement homogène** – en l'occurrence, des Européens et des Nord-Américains blancs de sexe masculin, véhiculant une culture et une vision du monde qui inévitablement imprègnent leur production.

³ Commission européenne, Livre Blanc « Intelligence artificielle, une approche européenne axée sur l'excellence et la confiance », 19 février 2020.
https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_fr.pdf



Ces biais socioculturels profonds sont a priori sans incidence pour des algorithmes visant à reconnaître des chiens ou des chats sur des images. Ils sont en revanche problématiques s'il s'agit de repérer des candidats dans des banques de CV.

LES DONNÉES D'APPRENTISSAGE

Le modèle étant construit à partir des données, on comprend la nécessité cruciale de disposer, pour l'apprentissage et la validation du modèle lui-même, de jeux de données non seulement de grande taille, mais aussi **représentatifs de la diversité des situations/cas à traiter**. Un algorithme de reconnaissance faciale entraîné sur une banque d'images comportant plus d'hommes que de femmes reconnaîtra « automatiquement » mieux les hommes que les femmes. Qui plus est, si la banque d'images ne reflète pas la diversité ethnique de la société, le modèle se révélera incapable de reconnaître les personnes appartenant aux groupes sous-représentés, a fortiori les femmes.

Si, pour des raisons statistiques, la qualité de l'apprentissage dépend beaucoup de l'ampleur des jeux de données, **un entraînement réalisé sur des données internes, même nombreuses, conduit assez logiquement à reproduire des biais antérieurs**. L'algorithme mis en place en 2015 par Amazon pour faciliter le recrutement de nouveaux talents a été entraîné sur les centaines de milliers de CV reçus par la firme pendant 10 ans. Son utilisation a été suspendue lorsqu'il est apparu qu'il attribuait de mauvaises notes à des profils de femmes qualifiées et proposait systématiquement des candidats hommes sous-qualifiés. Les hommes constituant l'écrasante majorité des cadres recrutés dans le passé, l'algorithme en avait déduit qu'il fallait sous-estimer les femmes !

Enfin, **expurger les données d'apprentissage de toute référence au genre, à l'origine, à la religion et autre critère potentiellement discriminatoire ne garantit pas non plus la neutralité** des futurs traitements. De telles informations peuvent en effet être déduites d'autres champs pris en compte par le modèle et a priori considérés comme neutres. Ainsi, le type de cosmétiques acheté par une personne permet de savoir s'il s'agit d'une femme ou d'un homme. La nuance de fond de teint achetée par une femme révèle de facto sa carnation et donc son origine ethno- raciale...

LES CRITÈRES DE PERFORMANCE

Un modèle est jugé performant lorsque, après la phase d'apprentissage, il traite correctement un pourcentage élevé de cas qu'il n'a jamais vus. Les tests pré-opérationnels consistent à évaluer la précision du modèle et à juger si le taux d'erreur est acceptable au regard de l'objectif poursuivi. Ces tests sont généralement réalisés sur des échantillons globaux **qui ne permettent pas de savoir si le modèle discrimine ou pas sur des critères de genre, d'origine ou autres. Ces critères ne sont tout simplement pas pris en compte**. Deux exemples :

- Un algorithme de speech recognition retranscrit sans erreur 98,7 % des messages vocaux qui lui sont soumis. Cette performance sera jugée satisfaisante pour un voicebot de service client, mais on peut découvrir, dans un deuxième temps, que la performance est bien moindre avec des locuteurs ayant certains accents⁴.

4 | <https://hbr.org/2019/05/voice-recognition-still-has-significant-race-and-gender-biases>



– Un algorithme visant à identifier les clients prêts à résilier leur abonnement repère 86,2 % sur un échantillon de test global. Pouvoir engager une action spécifique pour retenir ces clients est un gain réel pour l'entreprise. Mais est-elle sûre que les résultats sont aussi bons sur toutes les catégories – d'âge, de sexe, de lieu de résidence... – composant l'échantillon ? Faute de tests par classe, on ne le sait pas et cela peut conduire l'entreprise à engager des actions de rétention inutiles sur de nombreux « faux positifs »⁵ catégoriels. Imaginons maintenant qu'il s'agisse d'une application destinée non pas à prévenir le churn, mais à détecter la fraude dans une compagnie d'assurance ? Quelles seraient les conséquences pour l'entreprise ?



Comme le souligne un récent rapport de l'Institut Montaigne, « les cas avérés de biais algorithmiques en Europe et en France sont encore très limités »⁶. Cette situation s'explique par la rareté des déploiements à grande échelle de ce type d'outils dans la sphère la plus sensible : celle des politiques publiques – de santé, d'éducation, de logement, d'aide sociale... – où toute infraction avérée ou suspicion de discrimination ne manquerait pas d'être dénoncée. C'est ce qui s'est produit en 2018 avec l'algorithme de Parcoursup, accusé par ses détracteurs de comporter des biais sociologiques, voire idéologiques, favorisant la reproduction des inégalités sociales⁷.

C'est précisément parce que le recours à l'IA et aux algorithmes d'apprentissage est en train de se généraliser, tant dans le secteur public que dans les entreprises, que lutter durablement contre les discriminations algorithmiques devient une priorité. Cela demande de s'attaquer aux causes mises en évidence précédemment et pour cela – plutôt que de légiférer a priori – d'en appeler à la responsabilité et à la déontologie des différents acteurs. À ce titre, trois axes d'action nous paraissent devoir être travaillés en parallèle :

5 | En l'occurrence, des personnes n'ayant pas l'intention de résilier leur abonnement.

6 | Institut Montaigne, « Algorithmes : contrôle des biais S.V.P. », mars 2020. <https://www.institutmontaigne.org/ressources/pdfs/publications/algorithmes-controle-des-biais-svp.pdf>

7 | Parcoursup est l'application destinée à recueillir et gérer les vœux d'affectation des futurs étudiants de l'enseignement supérieur public français.

- ACCROÎTRE LA DIVERSITÉ DES ÉQUIPES DE CONCEPTION** – Les équipes qui développent aujourd'hui des algorithmes fondamentaux ou qui construisent des solutions à partir d'algorithmes existants sont **notoirement peu diversifiées** en termes de formation, et, en outre, **trop peu féminisées**. Cet état de fait favorise l'homogénéisation des approches et des points de vue, ainsi que la reproduction inconsciente de pratiques porteuses de stéréotypes. La **diversification des recrutements**, en termes de formation initiale, de profil et de parcours, de même que les **initiatives promouvant les métiers du numérique, de la data et des mathématiques auprès des femmes** doivent systématiquement être encouragées – tant dans les écoles d'ingénieurs et les centres de recherche que chez les éditeurs de solutions et dans les ESN.
- RENFORCER LES PROTOCOLES DE VALIDATION** – Une organisation faisant de la diversité et de la non discrimination un engagement prioritaire ne peut pas se contenter de juger la performance d'un algorithme d'apprentissage uniquement sur la précision globale de ses résultats. Qu'il s'agisse d'outils développés en interne ou de solutions externes, elle doit **formaliser ses critères d'évaluation** pour y intégrer la non-discrimination et mettre en place des protocoles lui permettant de **s'assurer de l'absence de biais rédhibitoires au niveau de l'algorithme et des données d'apprentissage et de validation du modèle**.
Tester les algorithmes avant déploiement « en s'inspirant des études cliniques de médicaments » comme le propose l'Institut Montaigne est une piste intéressante. Dans ce cadre, une bonne pratique serait, **outre l'audit systématique et approfondi des données**, de tester le modèle sur des sous-ensembles catégoriels caractérisés pour vérifier l'absence d'écart de performance par rapport aux résultats obtenus sur des échantillons de données non segmentés. La mise en place de ce type de méthodologie suppose de pouvoir **s'appuyer sur les compétences de Data Scientists dûment sensibilisés aux questions de discrimination algorithmique** et, plus largement aux questions éthiques.
- AMÉLIORER L'INFORMATION DES DÉCIDEURS ET DES MANAGERS** – La décision de recourir à une solution d'IA obéit en premier lieu à des objectifs de gain de performance et d'efficacité opérationnelle. Elle est souvent prise par **des managers ne soupçonnant même pas que les outils et solutions qu'ils choisissent sur des critères métiers sont susceptibles de comporter des biais discriminatoires**. L'Association Française des Managers de la Diversité souligne à juste titre que les éditeurs et « promoteurs de solutions ne mentionnent presque jamais les questions de diversité et/ou de lutte contre les discriminations parmi les bénéfices attendus de ces outils »⁸. Ils ne mentionnent pas plus les risques de biais discriminatoires et d'autant moins que leurs solutions ne sont pas testées sous cet angle. Outre l'indispensable acculturation des décideurs et managers aux problématiques algorithmiques, **la labellisation des produits/solutions/services présentant le moins de risques de cette nature** – par les associations professionnelles et/ou les organismes de certification – leur fournirait un précieux point de repère et limiterait la diffusion des solutions les moins vertueuses ou les plus opaques.

Ces pistes de travail ne sont pas des incantations. Nombre d'acteurs travaillent déjà dans ce sens et démontrent qu'il est tout à fait possible de corriger les biais discriminatoires, **dès lors qu'il y a prise de conscience et volonté de les débusquer**. Par exemple, constatant des différences stigmatisantes dans les traductions de Google Translate dans certaines langues, Google a entièrement revu le processus et les données d'apprentissage de façon à générer des traductions genrées pour toutes les phrases où le modèle d'origine proposait uniquement une version masculine. **La mise en place d'indicateurs spécifiques a permis de mesurer les progrès et de réduire ce biais jusqu'à plus de 90 %** dans les traductions du hongrois, du finnois, du persan et du turc vers l'anglais. Le nouveau modèle n'en est que plus pertinent et comprend enfin que dans ces langues « doctor » et « engineer » peuvent être féminins !⁹

Ce qui est en jeu aujourd'hui, à travers ce type de démarches, c'est l'acceptabilité sociale de ces technologies qui sont vouées à jouer un rôle croissant et structurant dans nos vies. Si l'on veut que ces technologies contribuent à la construction d'une société inclusive et juste, cette acceptabilité est conditionnée, en Europe et de manière peut-être encore plus aiguë en France, **par le respect de principes éthiques, au premier rang desquels figurent la non-discrimination et l'égalité de traitement des personnes**. C'est dans cet esprit que la Commission européenne a publié, en avril 2019, des *Lignes directrices en matière d'éthique pour une IA digne de confiance*¹⁰. Non contraignante, **cette proposition de cadre éthique place « la diversité, la non-discrimination et l'équité » parmi les 7 exigences essentielles d'une IA éthique**¹¹ et s'adresse à tous ceux « qui conçoivent, mettent au point, déploient, mettent en œuvre, utilisent l'IA ou sont soumises à ses incidences ».

Si la formalisation d'un cadre éthique commun nous paraît aller dans le bon sens, une question fondamentale doit néanmoins être posée : à date, les instances européennes réfléchissant à ces questions d'éthique dans l'IA sont-elles elles-mêmes représentatives de la diversité européenne ? En termes d'équilibre femmes/hommes probablement, mais en termes d'origines sociales, culturelles et ethniques ? La question reste ouverte...

8 | AFMD, « Recruter avec des Algorithmes ? Usages, opportunités et risques », mai 2019.

9 | La démarche est détaillée dans l'article « A Scalable Approach to Reducing Gender Bias in Google Translate », avril 2020, <https://ai.googleblog.com/2020/04/a-scalable-approach-to-reducing-gender.html>

10 | <https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines>

11 | Les 6 autres sont : facteur humain et contrôle humain ; robustesse technique et sécurité ; respect de la vie privée et gouvernance des données ; transparence ; bien-être sociétal et environnemental ; responsabilisation.

À PROPOS DES AUTEURS



Nelly CHATUE-DIOP

Nelly est Chief Data Officer au sein du Groupe Betcllic et siège en tant que Présidente du Conseil d'administration de L2F, start-up spécialisée dans l'IA. Avant de rejoindre Betcllic, elle a passé 6 ans dans la grande distribution en tant que Directrice Pricing. Nelly a débuté sa carrière en tant que Consultante Informatique et Data.

Par ailleurs, très sensible aux enjeux du numérique pour la jeunesse africaine, elle a co-fondé Nzingha Lab au Cameroun, studio

lab pour des solutions locales utilisant la Blockchain et/ou l'IA.

Nelly a obtenu son MBA à HEC Paris et à la London Business School et son diplôme d'ingénieur en informatique à l'ESCE Lyon.



Soumia MALINBAUM

Après avoir créé et dirigé pendant 15 ans Spécimen, une ESN créée en 2006 au Groupe Keyrus, Soumia intègre ce dernier au poste de DRH. Aujourd'hui, elle est Vice-Présidente Business Development, en charge des grands comptes sur l'ensemble des offres du Groupe et de l'accompagnement de l'engagement RSE. En parallèle, elle siège au Conseil d'administration du Syntec Numérique, dont elle préside la commission « Formation, e-éducation ». Elle est aussi engagée dans

la lutte contre les discriminations et la promotion de toutes les diversités.

Soumia est co-fondatrice de l'AFMD, siège au conseil de surveillance du groupe Lagardère et au conseil d'administration de Nexity. Elle est chevalier de l'Ordre national de la Légion d'Honneur.

KEYRUS

Keyrus, créateur de valeur à l'ère de la Data et du Digital

Acteur international du conseil et des technologies, spécialiste de la Data et du Digital, Keyrus a pour mission d'aider les entreprises à tirer profit du paradigme de la Donnée et du Numérique pour accroître leur performance, faciliter et accélérer leur transformation et générer de nouveaux leviers de croissance, et de compétitivité.

Plaçant l'innovation au cœur de sa stratégie, Keyrus développe une proposition de valeur unique sur le marché autour d'une offre novatrice qui s'appuie sur la combinaison de trois expertises majeures et convergentes :

Data Intelligence

Data Science - Intelligence Artificielle - Big Data & Cloud Analytics - Business Intelligence - EIM - CPM/EPM

Digital Experience

Innovation & Stratégie Digitale - Marketing Digital - DMP & CRM - Commerce Digital - Performance Digitale - User Experience

Conseil en Management & Transformation

Stratégie & Innovation - Transformation Digitale - Pilotage de la Performance - Accompagnement des Projets

Présent dans 19 pays et sur 4 continents, le Groupe Keyrus emploie 3 200 collaborateurs.

Plus d'informations sur : www.keyrus.fr