L'INTELLIGENCE ARTIFICIELLE AU SERVICE DE LA LUTTE CONTRE LES DISCRIMINATIONS DANS LE RECRUTEMENT : NOUVELLES PROMESSES ET NOUVEAUX RISQUES

Alain Lacroux, Christelle Martin-Lacroux

Management Prospective Ed. | « Management & Avenir »

2021/2 N° 122 | pages 121 à 142

ISSN 1768-5958

Article disponible en ligne à l'adresse :

https://www.cairn.info/revue-management-et-avenir-2021-2-page-121.htm

L'Intelligence artificielle au service de la lutte contre les discriminations dans le recrutement : nouvelles promesses et nouveaux risques

Alain LACROUX[1]
Christelle MARTIN-LACROUX[2]

Résumé

Le marché des outils d'aide au recrutement intégrant des modules d'intelligence artificielle est en plein essor. Parmi les arguments utilisés pour promouvoir ces dispositifs, figure la promesse que ceux-ci permettraient de favoriser un recrutement non discriminatoire, en raison de leur capacité supposée à éliminer les biais de jugement humains. L'objectif de cette revue des recherches est de montrer que ces promesses sont difficilement tenables, car la correction de certains biais de jugement est contrecarrée par l'émergence de nouveaux biais induits par l'usage de l'IA[3].

1. **Alain LACROUX** : Professeur de sciences de gestion, Université Polytechnique des Hauts de France (UPHF), IAE de Valenciennes, Campus du Mont Houy – CRISS (Centre de recherche interdisciplinaire en sciences de la société) – alain.lacroux@uphf.fr

2. **Christelle MARTIN-LACROUX** : Maître de Conférences en sciences de gestion, Université Grenoble Alpes IUT2 – CERAG (Centre d'études et de recherche appliquées à la gestion, EA 7521) – christelle.martin-lacroux@univ-grenoble-alpes.fr

3. La présente proposition s'appuie sur une communication présentée lors d'un symposium consacré à l'intelligence artificielle organisé lors du congrès 2019 de l'AGRH (Bordeaux, 13 au 15 novembre 2019).

Abstract

The recruitment tools market integrating artificial intelligence modules is rapidly expanding. Among the arguments used to promote these devices is the promise that they would promote non-discriminatory recruitment because of their supposed ability to eliminate human judgment bias. The purpose of this review is to show that these promises are difficult to keep, as the correction of some judgement biases is hampered by the emergence of new biases induced by AI.

Introduction

Wendy Hall, auteur d'un rapport sur l'intelligence artificielle a remis récemment au goût du jour un célèbre adage des spécialistes de l'analyse de données dans un article du *Financial Times* en notant : « *We used to talk about garbage in, garbage out ; now, with AI, we talk about bias in, bias out*[4] ». Cette phrase illustre le paradoxe engendré par les outils de recrutement fondés sur l'intelligence artificielle qui sont présentés par leurs auteurs comme des armes au service d'un recrutement « objectif », mais qui peinent à honorer leurs promesses, et sont même susceptibles de faire naître de nouveaux biais de décision chez les recruteurs. L'Intelligence artificielle (IA dans la suite de cet article) est une notion relativement large et parfois floue : on utilise ce terme dès lors qu'un système informatique est capable d'appliquer une règle de décision de manière autonome en s'appuyant sur l'information disponible dans une base d'apprentissage. Les algorithmes de *machine learning* (apprentissage supervisé adaptatif) sont capables de faire des prédictions sur des données, à partir de règles implémentées par le créateur de l'outil, et sont capables de les adapter en fonction de la quantité de données fournies. Dans les algorithmes de *deep learning* (apprentissage profond) qui se développent à partir de 2010, la logique est poussée plus loin. Les programmeurs tentent de mimer le fonctionnement du cerveau humain, en créant des algorithmes basés sur des réseaux de neurones artificiels (programmes informatiques capables d'appliquer une règle simple). Dans ce cas, les règles de décision peuvent être créées à partir d'analyses de classification menées sur une grande masse de données. Le *deep learning* se distingue par une logique *data driven*, très inductive.

L'un des secteurs dans lesquels l'IA gagne le plus rapidement en influence est l'instrumentation RH. Les arguments principaux avancés par les promoteurs des outils de recrutement intégrant des modules d'IA sont qu'ils permettent un *process* plus rapide, plus efficace et plus inclusif (exempt de biais discriminatoires). Toutes les phases du recrutement seraient concernées. Parmi les avantages clé mis en lumière par les professionnels pour promouvoir leurs produits, figure la promesse que ceux-ci permettraient de parvenir à un recru-

4. Alya Ram, *Financial Times*, 31 mai 2018.

tement non discriminatoire, en raison de leur « objectivité » et de leur capacité à éliminer les biais de jugement humains interpersonnels liés par exemple à l'activation de stéréotypes. Cette capacité à favoriser la diversité dans le recrutement est un argument à la fois technique, légal et stratégique. Sur le plan technique, la promesse est d'obtenir une diversité de manière quasi automatique, avec un gros gain de temps sur le processus de recrutement. Sur le plan légal, la promesse commerciale est d'éviter de faire l'objet de poursuites pour pratiques de recrutement discriminatoires. Sur le plan stratégique enfin, la diversité de main d'œuvre obtenue grâce à des procédures exemptes de biais de sélection « humains » est vue comme source d'avantage compétitif (Bear *et al*., 2010 ; Hoogendoorn & Van Praag, 2012). Le terme de « recrutement prédictif », largement promu par les créateurs des outils vient donner corps à cette promesse d'efficacité et d'objectivité.

L'objectif de cet article est de montrer que cette promesse est encore partiellement illusoire pour plusieurs raisons, dont la plus préoccupante tient au fonctionnement même des algorithmes au cœur des « boîtes noires » que sont les outils du recrutement prédictif. Ces algorithmes sont par essence non transparents, car ils constituent un avantage concurrentiel permettant la différenciation sur un marché qui est très vite devenu extrêmement disputé. Ils sont souvent juridiquement protégés, et les mécanismes permettant l'évaluation et le classement des postulants en évitant les critères discriminatoires demeurent opaques… parfois même aux yeux des développeurs (cf. *infra*).

Notre objectif est de montrer que les outils fondés sur l'IA ne contribuent pas toujours à réduire la discrimination à l'embauche, et qu'ils peuvent même engendrer de nouveaux biais de jugement. Notre démonstration prendra appui sur une revue de littérature ciblée et transdisciplinaire des recherches menées sur les utilisations de l'IA comme outil de sélection et d'aide à la décision dans les domaines de la gestion des ressources humaines, des sciences informatiques et de la psychologie. Nous verrons finalement qu'au-delà des problèmes apparemment techniques, l'impact croissant des algorithmes dans les choix des recruteurs renvoie à l'enjeu fondamental de l'équité décisionnelle et de « l'explicabilité » des outils de sélection fondés sur l'IA.

1. IA et recrutement : des instruments diversifiés au service d'une promesse d'objectivité

1.1. Méthodologie de l'étude

Nous avons réalisé une étude conceptuelle basée sur revue de littérature ciblée. Cette approche est relativement répandue lorsqu'il s'agit d'interroger des concepts émergents, tout particulièrement dans le champ des NTIC (voir par exemple Woods *et al*., 2019). Pour collecter les documents pertinents pour cette recherche, nous avons procédé en trois étapes. Dans un premier temps

nous avons utilisé une requête large (requête « *artificial intelligence* ») sur les principaux journaux spécialisés dans le domaine de la GRH et du recrutement (sources : *Personnel Psychology, International journal of Selection and Assessment, Human Resource Management, Human Resource Management Journal, Personnel review, Journal of Organizational behavior, Journal of Applied Psychology, @grh, Revue de Gestion des Ressources Humaines*). Le nombre relativement limité d'articles retournés (12) apparaît comme un indicateur du caractère récent du champ analysé dans la littérature en GRH. Nous avons ensuite étendu la recherche aux recherches académiques parues dans les revues spécialisées en Management, et psychologie des organisations en précisant la requête (*« artificial intelligence » + hiring / recruiting / personnel selection*). Nous nous sommes ici appuyés sur les bases de données scientifiques Google Scholar et *EBSCO Business source complete*, ainsi que Cairn pour les articles en français. Les résultats obtenus (19 articles dans la base *EBSCO Business source complete*) nous ont amenés à élargir davantage la recherche. Dans une dernière étape, nous avons donc mobilisé des sources variées (rapports de recherche, sites professionnels, presse généraliste) dans le domaine juridique, et des sciences de l'information notamment. L'ensemble des documents récupérés à l'issue de ces trois étapes constitue une base de 198 références.

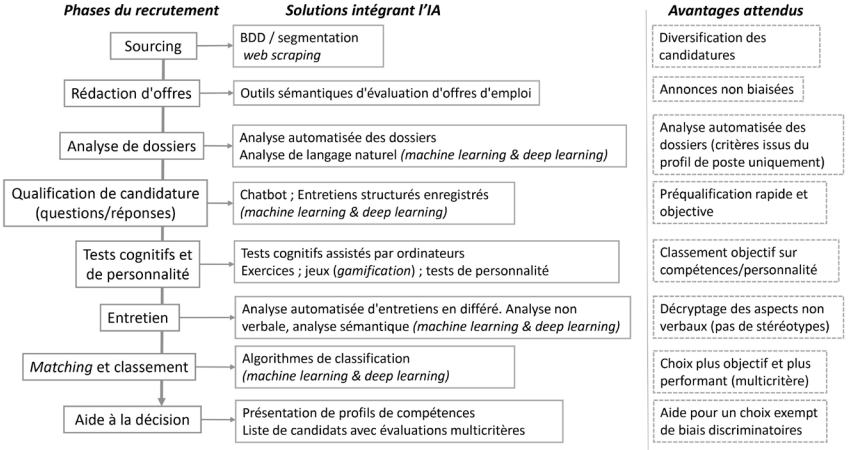
1.2. Un état des lieux des usages de l'IA dans le recrutement

Le thème du recrutement assisté par l'IA s'intègre dans le domaine plus général des outils analytiques au service des RH (HR *analytics*), dont les avantages et les limites sont de plus en plus discutés (Angrave *et al*., 2016). Nous ne sommes pas encore en présence d'un champ de recherche stabilisé, et il existe un écart important entre la place prise par ces outils dans les pratiques des entreprises et les travaux publiés dans les revues scientifiques de management (Marler & Boudreau, 2017 ; Woods *et al*., 2019). Sur le terrain, toutes les étapes des processus RH sont impactées par l'introduction d'algorithmes basés sur l'IA, et les solutions se développent de manière exponentielle (Figure 1). Selon les données recueillies en France, 50 % des professionnels RH (DRH et cabinets de recrutement) utilisaient en 2019 au moins un outil intégrant des algorithmes d'IA (logiciel de gestion automatique des candidatures[5]). Le marché du recrutement prédictif est en croissance rapide, sous l'impulsion de startups spécialisées[6]. Le projet généralement défendu par les entrepreneurs n'est pas de remplacer mais « d'augmenter » l'humain dans la tâche de recrutement, en insistant sur le fait que la décision finale est toujours prise par un recruteur.

5. https://www.hellowork.com/enquete-candidats-recruteurs-2019-2020/

6. Exemples : Easyrecrue, Kudoz, Goshaba, Assessfirst, HireSweet, Yatedo, Saven, Outmatch.

Figure 1 – La place de l'IA dans le processus de recrutement



1.3. Les promesses des créateurs des solutions de recrutement prédictif

Les arguments principaux avancés par les promoteurs des outils d'IA sont qu'ils permettent un recrutement plus rapide, plus efficace et plus inclusif (exempt de biais discriminatoires) et ce, à toutes les étapes du processus de recrutement. Les promesses sont spectaculaires : certains développeurs de solutions intégrées de recrutement promettent un gain de temps de 50 % sur un process de recrutement[7], ou encore la réduction du taux de départ annuel de 50 %[8].

Durant la phase de *sourcing*, la collecte d'informations en dehors de l'exploitation du dossier de candidature explicitement fourni par le candidat, notamment sur les réseaux sociaux est présentée comme un moyen d'optimiser la qualité de concordance entre offre et demande (*matching*). Il s'agit d'augmenter la capacité prédictive des algorithmes utilisés lors de la sélection en leur fournissant des données permettant d'inférer certains traits de personnalité ou compétences particulières pour les candidats potentiels à partir de leur activité visible sur internet. La performance et la capacité d'apprentissage des algorithmes de collecte et de tri sont dépendantes du volume de données collectées dans la base de connaissances. C'est pourquoi *machine learning* et *deep learning* sont si souvent accolés au Big Data. L'usage de robots qui explorent les réseaux sociaux à la recherche de données personnelles utilisables dans une logique prédictive (*web crawling* et *web scraping*) se généralise aujourd'hui en phase de sourcing. Plusieurs types d'information retrouvées sur les profils Linkedin sont considérés comme des prédicteurs de productivité et de résultats professionnels, comme par exemple le capital social qui serait corrélé à la productivité (Aguado *et al.*,

7. https://www.easyrecrue.com/fr/

8. https://www.cappfinity.com/case-studies/#lloyds

2019). La collecte autonome d'informations sur les réseaux sociaux permet également de s'intéresser à des salariés qui n'ont pas candidaté explicitement et de les proposer aux recruteurs. La recommandation algorithmique pour un candidat augmenterait ses chances d'être recruté de 20 % (Horton, 2017).

En ce qui concerne la rédaction d'annonces, des recherches ont montré qu'elle peut être stéréotypée (genrée), ce qui est susceptible de conduire au renoncement de candidater pour certains types de candidats (Gaucher *et al*., 2011). Des outils sémantiques permettent ainsi de diagnostiquer les risques discriminatoires dans les annonces et d'en optimiser la rédaction afin de les rendre plus inclusives[9].

En phase de présélection, l'usage du CV comme véhicule de discriminations est largement documenté (Derous & Decoster, 2017). Certains outils de qualification intégrant l'IA promeuvent donc un recrutement sans CV, basé sur des tests psychométriques et l'évaluation de compétences associées au poste offert. Cette phase est parfois menée par des robots conversationnels (*Chatbots*)[10]. En ce qui concerne les tests psychométriques, ceux-ci sont facilement « digitalisables » et sont considérés comme un outil de recrutement permettant d'objectiver certaines capacités des candidats en limitant les discriminations, avec une validité prédictive forte (Schmidt & Hunter, 1998).

La phase d'entretien, moment privilégié pour l'activation de biais affinitaires est également ciblée par les solutions d'IA. Les candidats peuvent être invités à préenregistrer une interview, qui sera traitée automatiquement à l'aide d'outils permettant d'analyser à la fois le discours (repérage de mots-clés) par analyse logicielle du langage, et certains critères non verbaux (expressions faciales, débit de parole…) grâce à des algorithmes d'analyse[11]. La digitalisation permettait dans ce cadre de réduire les biais liés à la première impression et à l'apparence physique (Suen *et al*., 2019).

D'autres outils se proposent aussi d'évaluer certaines capacités des candidats, sans que ceux-ci en soient nécessairement conscients (ex : l'évaluation de l'intelligence émotionnelle des candidats par l'analyse de leur activité sur les réseaux sociaux, Menon & Rahulnath, 2016).

La phase finale de sélection et de classement est celle dans laquelle les bénéfices anti-discrimination des algorithmes sont le plus souvent invoqués. Les nombreux biais psychosociaux individuels qui impactent le jugement du recruteur (effet de halo, effet d'attente, biais de confirmation, activation de stéréotypes, effet de primauté) seraient limités par le processus de présélection automatisé et la présentation de profils de candidats. Les algorithmes ayant

9. https://textio.com/

10. https://www.randstad.fr/randy/

11. https://www.hirevue.com/ ; https://www.easyrecrue.com/fr/solutions-RH/Entretien-video-differe

la capacité de classer en s'appuyant sur un nombre très élevé de critères, leur performance dans l'application de choix multicritères dépasse largement les capacités humaines. Il faut noter que les outils de recrutement prédictifs se présentent toujours comme des outils d'aide à la décision, et jamais comme des solutions entièrement autonomes. Le recruteur se voit proposer un classement des profils de candidats avec évaluation du degré d'adéquation par rapport au poste offert.

1.4. Des signaux d'alerte à ne pas négliger

Quelques résultats pionniers doivent cependant inviter à la prudence quant à l'efficacité d'ensemble de ces méthodes. On peut citer à ce propos l'exemple d'Amazon, qui a mis au point à partir de 2014 un modèle prédictif intégrant un algorithme de *machine learning*, entraîné sur une base comprenant les recrutements des 10 dernières années, et qui conduisait au déclassement des candidates féminines. Cet exemple, constitue une illustration exemplaire du « *garbage in, garbage out* » bien connu des spécialistes de l'IA, qui pointent des insuffisances dans le traitement efficace des biais de sélection. Certaines études (Bolukbasi *et al*., 2016) prouvent même la propension de ces modèles à reproduire fidèlement les discriminations qu'ils sont supposés gommer. Ces défaillances sont préoccupantes, car elles sont étroitement liées aux caractéristiques techniques des outils et à la méthodologie d'apprentissage supervisé ou non supervisé.

En nous focalisant sur l'exemple du recrutement, nous proposons donc dans ce qui suit d'analyser la littérature relative aux biais associés à l'usage de méthodes fondées sur l'IA en distinguant trois grandes étapes : la période antérieure au recrutement, la phase de sélection proprement dite et la phase de choix final du candidat. Nous reprenons ici une typologie qui distingue trois types de biais liés aux systèmes informatiques (Friedman & Nissenbaum, 1996) : les biais préexistants, qui touchent les données sur lesquelles les outils s'appuient pour réaliser des prédictions ; les biais techniques ou algorithmiques causés par le fonctionnement du système ; les biais émergents causés par l'utilisateur du système (biais liés au comportement du recruteur placé en situation de choix final).

2. IA et recrutement : des biais à toutes les étapes ?

2.1. Les biais préexistants : quand les données sont faussées

2.1.1. Le problème des données d'entraînement : « bias in bias out ? »

Les données d'entraînement sont les bases de données sur lesquelles les tests de validation et d'ajustement des modèles prédictifs sont menés. Dans le domaine du recrutement, il s'agit principalement des données disponibles sur les

candidats ayant postulé, dont les caractéristiques sont croisées avec celles des salariés effectivement recrutés. C'est à cette étape que les biais les plus difficiles à combattre se multiplient : il s'agit de biais institutionnels ou structurels (par opposition aux biais interpersonnels entre recruteur et candidat) qui sont « incorporés » dans les données d'entraînement sur lesquelles s'appuient les créateurs des algorithmes (Woods *et al*., 2019). Plusieurs auteurs soutiennent que, par leur fonctionnement même, les algorithmes fondés sur l'apprentissage supervisé conduisent au clonage des candidats et, dans la plupart des cas, au renforcement des stéréotypes (voir Besse, 2020 pour une simulation). Si les données d'entraînement sont biaisées, deux mécanismes peuvent conduire à des discriminations (Barocas & Selbst, 2016) :

– Si les « bons candidats » pris comme référence dans l'algorithme reflètent un préjugé (par exemple l'assertivité et les capacités de leadership associées à des traits « masculins »), ces préjugés seront reproduits dans les recommandations de recrutement.

– Si les outils de recrutement prédictif tirent des conclusions probabilistes à partir d'un échantillon biaisé de la population (par exemple un échantillon dans lequel les femmes sont sous-représentées), toute décision de recrutement qui repose sur ces conclusions va systématiquement désavantager les candidats sous-représentés dans les données d'apprentissage (la proportion de femmes recrutées sera faible).

Il faut noter que les biais liés aux données d'entraînement ne sont pas uniquement un vecteur de discrimination sexiste ou raciste : ils ont pour principale caractéristique de cloner la population de salariés existante.

2.1.2. La programmation : les algorithmes ont-ils un « genre » ?

Certains auteurs pointent à ce stade l'influence des caractéristiques sociodémographiques des programmeurs sur le fonctionnement des algorithmes. Le milieu des programmeurs informatiques est en effet spécifique, car les hommes diplômés y sont largement surreprésentés (O'Neil, 2017). Si l'on songe que le choix des critères peut être déterminant dans le calcul d'un score de recommandation, une pondération excessive mise sur le résultat de tests en logique mathématique pour le recrutement d'un manager peut surévaluer le nombre de candidats masculins ayant une culture mathématique (plus nombreux dans les écoles d'ingénieurs), même si le lien entre capacités managériales et niveau en mathématiques n'a jamais été établi. Les algorithmes de traitement du langage naturel reproduisent aussi des stéréotypes de genre qui apparaissent dans les documents écrits, notamment lorsqu'ils s'appuient sur des techniques de « prolongements de mots » (*word embedding*) qui cherchent à repérer les associations de mots (Garg *et al*., 2018). Ainsi l'exploration de la base d'articles Google news a permis de mettre en évidence une forte association entre le sexe et les co-occurrences de mots neutres en anglais (qui peuvent être en théorie associées aux deux genres) : le mot « femme » est fortement associé aux mots

« foyer, bibliothécaire » ; le mot « homme » est fortement associé aux mots « maestro, skipper, philosophe » (Bolukbasi *et al.*, 2016, p. 2).

2.1.3. Les biais liés au sourcing

Certains recruteurs discriminent sans nécessairement le savoir dès la phase de *sourcing* en choisissant des critères de diffusion de leur offre laissant de côté certains types de salariés. La plateforme Linkedin propose par exemple plus de 15 critères de diffusion d'une offre. L'algorithme *Linkedin Talent Match* utilise ces critères pour inférer les besoins des employeurs à partir leurs embauches précédentes ; il y donc de forts risques que les employeurs qui ont une tendance à la discrimination se voient proposer des candidats « adéquats », reflétant leurs choix précédents (Barocas, 2014). Concernant l'usage des réseaux sociaux comme instrument de *sourcing*, une étude récente (Van Iddekinge *et al.*, 2016) a montré que les femmes étaient mieux évaluées que les hommes à partir de leurs profils, et que les candidats issus de minorités étaient évalués plus défavorablement en termes d'adaptation, mais pas en termes de compétences et habiletés. La collecte automatique d'informations sur les réseaux sociaux pose également un problème de respect de la vie privée. Selon une étude expérimentale, lorsque les participants sont informés que le recruteur utilisait des informations disponibles sur les réseaux sociaux pour évaluer leur professionnalisme, ils sont plus enclins à entamer des poursuites et de plus, l'attractivité organisationnelle est impactée négativement (Stoughton *et al.*, 2015).

2.2. Les biais pendant le recrutement : des biais classiques aux biais algorithmiques

2.2.1. Les biais discriminatoires liés aux tests : des problèmes bien connus

Plusieurs recherches ont montré que les tests cognitifs, apparemment objectifs, peuvent être biaisés ; le média (papier ou numérique) ne change rien. L'une des distorsions les plus connues est liée à la menace du stéréotype (Steele & Aronson, 1995), qui représente l'effet psychologique qu'un stéréotype peut avoir sur un sujet qui en est habituellement victime. Face à certaines situations de test, le sujet peut avoir la sensation d'être jugé à travers un stéréotype négatif visant son groupe, ce qui peut provoquer un stress et une diminution des performances sur certains tests cognitifs utilisés en recrutement (Ng & Sears, 2010). Face à des tests à choix multiples (QCM avec pénalisation), les femmes ont par exemple tendance à s'autocensurer dans les réponses plutôt que risquer de donner une réponse fausse, ce qui conduit à une surperformance statistique des hommes plus nombreux à tenter une réponse au hasard (Baldiga, 2013). Si la ludification de la passation (*gamification*) est supposée protéger contre la menace du stéréotype en dédramatisant cette épreuve, elle est également porteuse de biais spécifiques. L'aisance face aux écrans et la rapidité de réac-

tion est par exemple affectée par l'habitude d'utiliser ce type de média, ce qui donne un avantage aux jeunes générations (Galois-Faurie & Lacroux, 2014).

2.2.2. Les entretiens différés : une fausse bonne idée ?

Les recherches montrent que les candidats préfèrent les entretiens en face à face plutôt que les entretiens différés ou asynchrones (Chapman *et al*., 2003 ; Sears *et al*., 2013). Les entretiens différés sont perçus comme moins justes, plus inquisiteurs, avec moins de contrôle perçu du candidat, et moins de présence sociale que des visioconférences (Brenner *et al*., 2016 ; Hiemstra *et al*., 2019).

2.2.3. Les biais techniques et algorithmiques : quand la machine se trompe

Les premiers types de biais sont liés à des défaillances techniques : les algorithmes peinent par exemple à décrypter le sens de certaines expressions en langage naturel. Ce type d'erreurs techniques est potentiellement vecteur de discrimination. Plusieurs recherches ont montré une mauvaise détection des expressions faciales des candidats au physique atypique (par rapport aux données d'entraînement) : par exemple, le taux d'erreur pour la reconnaissance d'une expression peut varier de 1 % pour un homme blanc à 35 % pour une femme noire (Buolamwini & Gebru, 2018).

Les biais algorithmiques peuvent aussi provenir du fonctionnement normal de l'algorithme, sans que le programmeur en ait été conscient. Il faut conserver à l'esprit que les méthodes reposant sur l'IA sont généralement « pilotées par les données » (*data driven*) dans le sens où les programmeurs cherchent à créer une combinaison optimale de prédicteurs avec pour objectif d'offrir une prévision la plus précise possible du comportement d'une variable « cible » (par exemple l'adéquation au poste proposé). Les prédicteurs sont en premier lieu sélectionnés selon leur impact statistique et non pas leur intérêt managérial ou leur validité scientifique. C'est ainsi que des critères comme le lieu d'habitation du candidat peuvent émerger comme des prédicteurs optimaux, tout simplement parce que les salariés ayant un score élevé sur la variable « cible » habitent dans telle ville. La force d'association avec la variable cible n'est absolument pas une preuve d'influence causale.

Dans le *machine learning* traditionnel, les données d'entraînement déterminent ainsi l'efficacité de l'outil. Le futur est expliqué par le passé, et il y a donc possibilité de discrimination par inférence statistique (Lohr, 2013) : les corrélations qui apparaissent dans les données d'entraînement peuvent être la source de prédictions biaisées, lorsqu'elles sont fallacieuses (*spurious correlations*). Par exemple, l'association entre l'origine géographique du candidat et la performance en poste n'apporte aucune information sur les mécanismes de causalité sous-jacents.

Dans l'apprentissage profond (*deep learning*), il y a un « effet boîte noire » : contrairement aux modèles dits explicites utilisés dans le *machine learning* qui

conduisent à des décisions explicables et reproductibles[12], il n'est pas possible de connaître et de reproduire le mode de décision de l'algorithme de *deep learning* en raison de la complexité du processus de classification. On parle dans ce cas de modèle de boîte noire car la seule information qu'il est possible de donner est la mesure de l'importance de certains prédicteurs (autrement dit le résultat du calcul).

Un dernier type de biais algorithmique est plus fondamental, puisqu'il renvoie à la logique sous-jacente à la programmation. Le cas de la détection des expressions faciales est à ce titre exemplaire. Les programmeurs (aidés par des psychologues) associent certaines expressions faciales à des traits de personnalité, qui sont à leur tour associés à certaines capacités managériales. La validité prédictive de ce genre d'inférence est un sujet très débattu. Certaines études reconnaissent un lien (Rule *et al*., 2011) : l'inférence de traits psychologiques, notamment le leadership, à partir de la simple photographie montre souvent une forte cohérence inter-juges (Pillemer *et al*., 2014), considérée comme un indicateur de validité prédictive à l'efficacité « surprenante » (Todorov *et al*., 2015). Les mécanismes à l'œuvre notamment le rôle des stéréotypes de genre et des théories implicites de la personnalité demeurent un sujet de controverse (Rule & Ambady, 2011). Les travaux récents menés sur la reconnaissance faciale des émotions fondamentales (colère, angoisse…) viennent renforcer les doutes sur la possibilité d'une reconnaissance automatique en montrant que le décodage de ces émotions est culturellement dépendant : une même mimique faciale est interprétée différemment selon les cultures (Chen, Crivelli *et al*., 2018).

2.3. Les biais en phase de choix : de l'aide à la décision à la décision automatisée ?

2.3.1. La présentation des choix : aide ou incitation ?

Les promoteurs des outils de recrutement prédictif argumentent à juste titre que la décision finale appartient toujours à un humain. Cependant, le choix des candidats à rencontrer peut être fortement impacté par les algorithmes de recommandation, qui utilisent des approches probabilistes. Les résultats de la sélection sont le plus souvent calculés en probabilités, mais sont généralement donnés sous forme de classement présentés sans marge d'erreur. Les travaux fondateurs de Tversky et Kahnemann (1974) ont popularisé une série de biais décisionnels liés à la présentation des solutions, comme le biais d'ancrage, ou l'effet de cadrage (*framing effect*) qui influencent fortement les décisions, surtout lorsque celles-ci comportent une part de risque (Kühberger, 1998). Lorsque les résultats apparaissent sous la forme d'un classement ou d'une sélection, il y a un fort risque de biais de présentation (Craswell *et al*., 2008) : les premiers candidats qui apparaissent reçoivent une très forte « surprime »,

12. Ces modèles sont issus de l'application de méthodes de régression multiple et de classification plus ou moins sophistiquées mais qui sont déterministes et paramétrables.

et ce d'autant plus que le choix multicritère est un cas typique de décision de rationalité limitée. Or, ces candidats peuvent avoir obtenu des scores agrégés proches de candidats non présentés dans le choix final lorsque le nombre de dossiers examinés a été important.

2.3.2. Le biais d'automation : quand l'humain suit aveuglément la machine

Le biais d'automation apparaît lorsque le recruteur donne un poids déterminant aux informations provenant de l'algorithme. En étendant au recrutement les résultats obtenus sur les choix dans le domaine médical ou du contrôle aérien (Parasuraman & Riley, 1997), on peut supposer que le candidat classé premier sur la liste de recommandation a une forte probabilité d'être choisi en raison de la confiance dans l'algorithme de classement et de la difficulté de la tâche (Goddard *et al*., 2011). La confiance accordée à l'algorithme en situation de prise de décision comportant une ambiguïté serait supérieure à celle accordée aux humains, sauf chez les professionnels expérimentés (Logg, 2019). Il faut toutefois garder à l'esprit que la Commission nationale de l'informatique et des libertés rappelle que « *prendre une décision à l'égard d'une personne sur le seul fondement d'un traitement automatisé des données à caractère personnel est interdit* ».

3. Discussion : Peut-on corriger les algorithmes pour les rendre plus « vertueux » ?

3.1. La correction ex-post des biais

Les promoteurs des outils de recrutement prédictifs sont pour la plupart conscients de la possibilité de biais liés aux données d'apprentissage. Plusieurs solutions sont d'ores et déjà proposées.

3.1.1. Première piste : agir sur la présentation des résultats (ex-post)

Cette première solution, applicable en fin de chaîne, consiste à intégrer une part d'aléatoire dans la présentation des résultats au recruteur afin d'éviter les biais de présentation susceptibles d'influencer le choix final au profit des profils apparaissant en première position. On peut même aller plus loin et envisager un usage plus important de l'aléatoire (*randomization*), tel que proposé par Tambe *et al*. (2019) : la proposition apparemment irrationnelle d'introduire un tirage au sort entre plusieurs candidats dont le niveau de concordance avec le poste proposé (*matching*) est élevé serait selon certaines recherches bien acceptée par les candidats dans des situations de décisions multicritères (Lind & van den Bos, 2002).

3.1.2. Deuxième piste : agir sur les critères de sélection

Une deuxième solution consiste à rendre aveugles les systèmes face à certaines caractéristiques sociodémographiques comme le sexe, l'origine (c'est par exemple la méthode utilisée dans le CV anonyme). Il s'agit par exemple de limiter les mots-clés à des termes techniques et à ne pas prendre en compte certains critères sociodémographiques. Cette approche de l'équité est toutefois qualifiée de naïve par certains auteurs, qui pointent son inefficacité face aux biais structurels, et la possibilité pour le recruteur de s'appuyer sur des indices subtils permettant d'écarter certains types de candidature (Derous & Decoster, 2017). Des simulations menées sur ce type de pratique (DeArtega *et al*., 2019 ; Besse, 2020) montrent que la suppression de certains indicateurs sociodémographiques du *pool* de variables explicatives ne modifie pas les résultats des algorithmes de *machine learning*, en raison de leur corrélation avec d'autres indicateurs sociodémographiques plus subtils (par exemple le parcours académique ou les voyages).

3.1.3. Troisième piste : redresser les données d'apprentissage

Une troisième méthode est plus efficace, mais repose sur une philosophie différente. Elle consiste à redresser et pondérer les bases d'apprentissage, avec pour but d'avantager les catégories discriminées. Cette solution est la seule qui permette de lutter efficacement contre les biais discriminatoires sur des données de simulation (Besse, 2020). Concrètement, il s'agit d'auditer les résultats obtenus sur une période donnée pour recalibrer les données d'apprentissage et permettre à l'algorithme de modifier ses prédictions. Ces techniques de redressement visent non pas à rendre la base d'apprentissage représentative de la population salariée réelle mais plutôt à la rendre conforme à un modèle idéal en pratiquant une discrimination positive. Cela permet à certains opérateurs de proposer des *pools* de candidats « divers » et favoriser ainsi un recrutement inclusif[13]. Cette forme de discrimination positive repose sur une philosophie *diversity conscious*, autrement dit la correction *ex-post* des discriminations par une action sur les données d'entraînement. Ce type de redressement ne correspond pas vraiment à la conception républicaine de l'égalité, encore majoritaire en France, où de telles pratiques sont rares et perçues comme risquées (Bender *et al*., 2010).

Le cas est plus compliqué pour les outils bâtis sur des algorithmes d'apprentissage non supervisés : comme le remarquent Villani *et al*. (2018, p. 143), les possibilités de correction des algorithmes de *deep learning* sont pour l'instant inexistantes, notamment en raison de leur opacité fondamentale (Burrell, 2016), et aussi parce que les entreprises ne sont pas prêtes à dévoiler ce qui constitue leur avantage concurrentiel. Étrange paradoxe de bâtir un avantage concurrentiel sur un outil que l'on ne maîtrise pas réellement !

13. https://www.atipicainc.com/#about-us

3.2. Le problème des critères de justice et de « l'explicabilité » des algorithmes

Comme nous l'avons vu, même s'il est difficile ou parfois impossible de rendre à la fois efficaces et équitables les algorithmes de recrutement prédictif, les recruteurs doivent *a minima* être capables de justifier et expliquer les décisions prises auprès des candidats. « L'explicabilité », néologisme apparu dans le contexte de l'IA est un critère majeur de responsabilité juridique des recruteurs qui utiliseront ces techniques. La législation française (Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique) impose d'expliquer une décision administrative obtenue par un traitement automatique lorsque la personne physique concernée en fait la demande. Comme le remarquent les auteurs du rapport Villani, « *L'explicabilité des systèmes à base d'apprentissage constitue un véritable défi scientifique, qui met en tension notre besoin d'explication et notre souci d'efficacité* » (Villani *et al*., 2018, p. 141).

La réflexion sur l'équité des critères décisionnels dans les algorithmes, particulièrement dans le cadre du recrutement, est un sujet complexe, qui mêle statistique et philosophie (Kilbertus *et al*., 2017). Cette réflexion appliquée à la procédure criminelle conduit par exemple à distinguer six types différents d'équité incorporables dans les algorithmes de sélection, qui sont parfois incompatibles (Berk *et al*., 2018). La présentation détaillée de ces formes d'équité excède le cadre du présent article, mais on peut retenir que les auteurs sont plutôt pessimistes sur les possibilités de combiner un haut niveau d'efficacité et un haut niveau d'équité pour une même décision (*ibid*, p. 17). La question des boîtes noires divise encore les chercheurs. Pour certains auteurs, qui parlent de « société de boîte noire » (Pasquale, 2015), l'impossibilité d'expliquer est source d'opacité et de danger à terme pour la démocratie. D'autres, en s'appuyant sur le parallèle avec les médicaments (dont les chercheurs ne savent pas toujours expliquer pourquoi ils fonctionnent), considèrent que le vrai problème n'est pas finalement d'expliquer mais de parvenir à des prédictions fiables. On retrouve ici la thèse positiviste de l'instrumentalisme méthodologique selon laquelle la valeur d'un modèle se mesure avant tout à sa faculté à prédire correctement la réalité (Friedman, 1984).

4. Conclusion et pistes de recherche : quels défis pour les gestionnaires RH face au risque de biais liés à l'IA dans le recrutement ?

Nous venons de voir que, loin de constituer une solution pour un recrutement exempt de discrimination, les solutions basées sur l'IA engendrent de nouveaux risques. Au niveau managérial, ceci amène à s'interroger sur la gestion de ces risques techniques et juridiques par les responsables RH. Nous proposons d'illustrer les enjeux conceptuels et pratiques de cette gestion des risques

par la présentation de quatre défis qui s'imposeront aux responsables RH, et constituent autant de pistes de recherche.

4.2.1. Premier défi : gérer le dilemme entre efficacité des algorithmes et protection des données personnelles

L'IA est souvent associée au big data, avant tout pour des raisons techniques : toutes choses égales par ailleurs la fiabilité des prédictions dans les systèmes basés sur le *machine learning* et plus encore le *deep learning*, dépend de la quantité de données collectées dans la base d'apprentissage. Il y a donc une incitation à collecter et accumuler (dans une logique technique portée par les *data scientists* en entreprise), qui conduit presque immanquablement à des problématiques de respect de la vie privée (dans une logique juridique, portée par les spécialistes RH et les juristes d'entreprise). Sur le terrain, les tensions entre *data scientists* et juristes autour des solutions fondées sur les big data RH (informations personnalisées sur les caractéristiques, les compétences et les performances des salariés) se cristallisent justement autour de cette question des données personnelles. Certaines entreprises conscientes du phénomène, comme Google ou Microsoft, tendent aujourd'hui à recruter les deux profils au sein des services RH pour essayer de résoudre ces conflits de valeurs (Tambe *et al*., 2019). Plus généralement, la dimension éthique des outils de recrutement assistés par IA est un enjeu majeur, encore mal appréhendé (notamment en ce qui concerne le respect de la vie privée et les biais algorithmiques). Ceci invite à la prudence dans leur utilisation et ouvre aussi des chantiers de recherche transdisciplinaires, intéressant le droit, le management et l'informatique (voir par exemple Coron, 2020).

4.2.2. Deuxième défi : résister à l'illusion technologique

Il faut sans doute entreprendre aujourd'hui dans les services RH des efforts de démythification (voire de démystification) de l'IA dans le recrutement. Les exemples de promesses anti-discrimination montrent bien la nécessité de résister à l'illusion technologique sous-jacente dans les discours véhiculés par certains promoteurs de l'IA dans le recrutement. On constate dans la réalité un très grand décalage dans le discours sur l'IA : il apparaît que le niveau d'optimisme des discours sur les capacités des algorithmes d'IA est inversement proportionnel au niveau d'expertise des auteurs. Les véritables spécialistes qui travaillent de longue date sur l'intelligence artificielle parlent parfois « d'incompétence » artificielle en lieu et place « d'intelligence » artificielle (Heaven, 2019). L'un des spécialistes actuels de l'IA les plus reconnus, Yann le Cum (dirigeant du laboratoire d'intelligence artificiel de Facebook), rappelle souvent que « la meilleure IA disponible à l'heure actuelle a moins de sens commun qu'un rat » (Sermondadaz, 2018).

Les solutions d'IA actuelles reposent en général sur une simple compétence d'association (mise en évidence de règles d'association basées sur les corrélations) et elle permet d'obtenir des résultats extrêmement performants dans certains

domaines (ex : la reconnaissance d'image), mais il s'agit là d'une compétence qui correspond à l'échelon le plus bas sur « l'échelle de la causalité » (Pearl & Mackenzie, 2018). Associer deux évènements est une tâche relativement simple, qui est à la portée de nombreuses intelligences animales et artificielles (raisonnement associatif : « Quand je prends de l'aspirine, mon mal de tête disparaît »). Par contre, le deuxième niveau de raisonnement causal (raisonnement interventionnel : « Est-ce que la prise d'aspirine a réduit mon mal de tête ? ») et encore plus le troisième niveau (raisonnement contre-factuel ou rétrospectif : « Que se serait-il passé si je n'avais pas pris d'aspirine ? ») sont pour l'instant hors de portée des meilleures IA… mais restent à la portée d'un enfant.

Il faut rappeler sans cesse le sens de la notion de corrélation, qui ne signifie pas causalité. En effet, l'accumulation de données dans les bases d'apprentissage des algorithmes augmente mécaniquement le risque de corrélations artificielles et fallacieuses que les utilisateurs finaux ont tendance à interpréter trop aisément comme des relations causales. Le développement d'algorithmes de nouvelle génération reposant sur l'inférence causale semble être une piste prometteuse (Tambe *et al*., 2019) : l'idée est de ne pas utiliser de simples règles d'association, mais de modéliser et d'insérer dans la base de connaissance de l'IA des analyses causales (*path analysis*) permettant d'isoler les causes précises et réduire ainsi l'effet boîte noire. De tels algorithmes existent et peuvent potentiellement être appliqués à des données RH (voir par exemple Kalish *et al*., 2012) : il s'agit encore ici d'un chantier de recherche pluridisciplinaire prometteur, intéressant la psychologie du travail et les sciences de la décision, dont les principaux enjeux sont esquissés dans l'article de Tambe *et al*. (2019).

4.2.3. Troisième défi : optimiser la transparence et l'explicabilité

L'enjeu fondamental de l'explicabilité et de la transparence n'est pas encore réellement traité, même si le rapport Villani (2018) propose quelques pistes, en recommandant notamment de reconsidérer le sujet de la validation des outils. Un outil valide selon la loi est un outil dont il est possible d'expliquer le fonctionnement, ce qui n'est pas le cas des algorithmes de *deep learning* dont la validité repose uniquement sur sa capacité prédictive. Parvenir à la transparence des algorithmes pourrait par ailleurs entraîner des conséquences paradoxales : un algorithme « transparent » et ouvert peut être détourné (ou « hacké ») par les candidats (Tambe, 2019). Ceux-ci peuvent adapter leur réponse, et mettre en place des procédures destinées à « tromper » l'algorithme de tri automatique (ex : en incluant dans le CV une grande quantité de mots-clés cachés, intégrés dans des dessins par exemple) et ainsi augmenter artificiellement le degré de concordance.[14]

14. https://www.topresume.com/career-advice/6-cheats-every-job-hunter-should-know

4.2.4. Quatrième défi : interroger la validité prédictive des solutions d'IA (le problème du « bon candidat »)

Le problème du jugement d'efficacité des prédictions demeure très délicat dans le recrutement, et ce d'autant plus que l'usage d'algorithmes basés sur des recommandations peut entraîner des phénomènes de prophéties auto-réalisatrices (Coron, 2019). Par exemple, un algorithme de sélection de CV peut produire une prédiction d'adéquation au poste proposé qui vient elle-même influer sur le comportement des recruteurs, et fait finalement advenir la réalité prédite : les méthodes d'apprentissage supervisé deviennent finalement d'autant plus performantes que les prédictions faites sont considérées par les utilisateurs victimes du biais d'automation vu précédemment comme étant le reflet de la réalité.

Comme nous venons de le voir, un algorithme est efficace si ses prédictions sont jugées bonnes (autrement dit, le « bon candidat » a été recruté) : or, ce critère qui renvoie à la performance en poste est notoirement compliqué à établir (Schmidt & Hunter, 1998), et demeure sujet à débat, puisque nous avons vu précédemment qu'il tendait à reproduire les discriminations existantes. Au niveau de la recherche académique, la question de la validité (prédictive et de construit) des outils de sélection digitaux est un sujet nouveau, et il n'existe pas suffisamment de résultats de recherches pouvant servir de guide au praticien (Woods & West, 2019). Si la méta-analyse récente de Schmidt *et al*. (2016) analyse la validité prédictive de nombreuses méthodes de sélection, la validité prédictive des nouvelles techniques issues de la digitalisation demeure inconnue.

Il faut également rester vigilant quant aux capacités d'individualisation du recrutement promises par les promoteurs de l'IA. Ces dispositifs présentent un point commun : la volonté d'individualiser et de permettre aux entreprises de recruter le « profil rare » (granularité fine). Il y a ici une tension paradoxale, déjà mise en évidence par Coron (2019) entre la promesse de rendre un service personnalisé et l'usage de méthodes statistiques basées sur la classification et la fabrication de profils-type, qui peuvent être hautement stéréotypés avec un risque de clonage discriminatoire dans le recrutement. Dans le domaine de la validité prédictive des outils d'IA, la question des réactions des candidats soumis à ce type d'outils durant le processus de recrutement mérite également d'être prise en compte : la nécessité pour les DRH de veiller à une « expérience candidat » positive et d'être vus comme des employeurs responsables devient cruciale compte tenu de la diffusion des commentaires sur les plateformes d'évaluation comme Glassdoor (Van Hoye, 2013). Dans ce cadre, l'étude fine de la réaction des candidats confrontés à un recrutement digitalisé constitue un domaine de recherche encore en devenir.

Bibliographie

AGUADO D., ANDRÉS J., GARCÍA-IZQUIERDO A.L. & RODRÍGUEZ J. (2019), "LinkedIn 'Big Four' : Job Performance Validation in the ICT Sector", *Journal of Work and Organizational Psychology*, Vol. 35, n°2, p. 53-64.

ANGRAVE D., CHARLWOOD A., KIRKPATRICK I., LAWRENCE M. & STUART M. (2016), "HR and analytics : Why HR is set to fail the big data challenge", *Human Resource Management Journal*, Vol. 26, n°1, p. 1-11.

BALDIGA K. (2013), "Gender differences in willingness to guess", *Management Science*, Vol. 60, n°2, p. 434-448.

BAROCAS S. (2014), "Data mining and the discourse on discrimination", *Proceedings of the Data Ethics Workshop, Conference on Knowledge Discovery and Data Mining*.

BAROCAS S. & SELBST A.D. (2016), "Big data's disparate impact", *California Law Review*, n° 104, p. 671. SSRN : https ://ssrn.com/abstract=2477899

BEAR S., RAHMAN N. & POST C. (2010), "The Impact of Board Diversity and Gender Composition on Corporate Social Responsibility and Firm Reputation", *Journal of Business Ethics*, Vol. 97, n° 2, p. 207-221.

BENDER A.F., KLARSFELD A. & LAUFER J. (2010), "Equality and diversity in the French context". In *International Handbook on Diversity Management at Work : Country Perspectives on Diversity and Equal Treatment*, A. Klarsfeld (Ed.), Edward Elgar, Cheltenham.

BERK R., HEIDARI H., JABBARI S., KEARNS M. & ROTH A. (2018), "Fairness in criminal justice risk assessments : The state of the art", *Sociological Methods & Research*, p. 1-42.

BESSE P. (2020), « Détecter et évaluer les risques des impacts discriminatoires des algorithmes d'IA », hal-02616963, 19.

BOLUKBASI T., CHANG K.W., ZOU J.Y., SALIGRAMA V. & KALAI A.T. (2016), "Man is to computer programmer as woman is to homemaker ? Debiasing word embeddings", *Advances in neural information processing systems*, p. 4349-4357.

BRENNER F.S., ORTNER T.M. & FAY D. (2016), "Asynchronous Video Interviewing as a New Technology in Personnel Selection : The Applicant's Point of View", *Frontiers in Psychology*, Vol. 7, p. 863.

BUOLAMWINI J. & GEBRU T. (2018), "Gender shades : Intersectional accuracy disparities in commercial gender classification", *Proceedings of Machine Learning Research*, Vol. 81, p. 77-91.

BURRELL J. (2016), "How the machine 'thinks' : Understanding opacity in machine learning algorithms", *Big Data & Society*, Vol. 3, n°1, p. 1-12.

CHAPMAN D.S., UGGERSLEV K.L. & WEBSTER J. (2003), "Applicant reactions to face-to-face and technology-médiated interviews : A field investigation", *Journal of Applied Psychology*, Vol. 88, n°5, p. 944-953.

CHEN C., CRIVELLI C., GARROD O.G., SCHYNS P.G., FERNÁNDEZ-DOLS J.M. & JACK R.E. (2018), "Distinct facial expressions represent pain and pleasure across cultures", *Proceedings of the National Academy of Sciences*, Vol. 115, n°43, p. E10013-E10021.

CORON C. (2019), « Big Data et pratiques de GRH », *Management & Data Science*, Vol. 3, n°1, p. 1-9.

CORON C. (2020), « L'utilisation des données personnelles dans les algorithmes en gestion des resources humaines », *RIMHE*, n°39, p. 95-106.

CRASWELL N., ZOETER O., TAYLOR M. & RAMSEY B. (2008), "An experimental comparison of click position-bias models", *Proceedings of the 2008 international conference on web search and data mining*, p. 87-94.

DE-ARTEAGA M., ROMANOV A., WALLACH H., CHAYES J., BORGS C., CHOULDECHOVA A., GEYIK S., KENTHAPADI K. & KALAI A.T. (2019), "Bias in bios : A case study of semantic representation bias in a high-stakes setting", *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 120-128.

DEROUS E. & DECOSTER J. (2017), "Implicit Age Cues in Resumes : Subtle Effects on Hiring Discrimination", *Frontiers in Psychology*, 8, p. 1321.

FRIEDMAN B. & NISSENBAUM H. (1996), "Bias in Computer Systems", *ACM Trans. Inf. Syst*., Vol. 14, n°3, p. 330-347.

FRIEDMAN M. (1984), "The methodology of positive economics". In *The Philosophy of Economics*, Daniel M. Hausman (Ed), Cambridge University Press, Cambridge.

GALOIS-FAURIE I. & LACROUX A. (2014), « 'Serious games' et recrutement : Quels enjeux de recherche en gestion des ressources humaines ? », *@GRH*, Vol. 10, n°1, p. 11-35.

GARG N., SCHIEBINGER L., JURAFSKY D. & ZOU J. (2018), "Word embeddings quantify 100 years of gender and ethnic stereotypes", *Proceedings of the National Academy of Sciences*, Vol. 115, n°16, p. 3635-3644.

GAUCHER D., FRIESEN J. & KAY A.C. (2011), "Evidence that gendered wording in job advertisements exists and sustains gender inequality", *Journal of Personality and Social Psychology*, Vol. 101, n°1, p. 109-128.

GODDARD K., ROUDSARI A. & WYATT J.C. (2011), "Automation bias : A systematic review of frequency, effect médiators, and mitigators", *Journal of the American Medical Informatics Association*, Vol. 19, n°1, p. 121-127.

HARDT M., PRICE E. & SREBRO N. (2016), "Equality of opportunity in supervised learning", *Advances in neural information processing systems*, p. 3315-3323.

HEAVEN D. (2019), "Why deep-learning AIs are so easy to fool", *Nature*, Vol. 574, p. 163.

HIEMSTRA A.M.F., OOSTROM J.K., DEROUS E., SERLIE A.W. & BORN M.P. (2019), "Applicant perceptions of initial job candidate screening with asynchronous job interviews : Does personality matter ?", *Journal of Personnel Psychology*, Vol. 18, n°3, p. 138-147.

HOOGENDOORN S. & VAN PRAAG M. (2012), *Ethnic diversity and team performance : A field experiment*, http ://papers.ssrn.com/sol3/papers.cfm ?abstract_id=2105284

HORTON J.J. (2017), "The Effects of Algorithmic Labor Market Recommendations : Evidence from a Field Experiment", *Journal of Labor Economics*, Vol. 35, n°2, p. 345-385.

KALISCH M., MÄCHLER M., COLOMBO D., MAATHUIS M.H. & BÜHLMANN P. (2012), "Causal Inference Using Graphical Models with the R Package pcalg", *Journal of Statistical Software*, Vol. 47, n°1, p. 1-26.

KILBERTUS N., CARULLA M.R., PARASCANDOLO G., HARDT M., JANZING D. & SCHÖLKOPF B. (2017), "Avoiding discrimination through causal reasoning", *Advances in Neural Information Processing Systems*, p. 656-666.

KÜHBERGER A. (1998), "The Influence of Framing on Risky Decisions : A Meta-analysis", *Organizational Behavior and Human Decision Processes*, Vol. 75, n°1, p. 23-55.

LIND E.A. & VAN DEN BOS K. (2002), "When fairness works : Toward a general theory of uncertainty management", *Research in organizational behavior*, 24, 181-224.

LOHR S. (2013), "Sizing Up Big Data, Broadening Beyond the Internet", *Bits Blog New York Times*. https ://bits.blogs.nytimes.com/2013/06/19/sizing-up-big-data-broadening-beyond-the-internet/

LOGG J.M., MINSON J.A. & MOORE D.A. (2019), "Algorithm appreciation : People prefer algorithmic to human judgment", *Organizational Behavior & Human Decision Processes*, Vol. 151, p. 90-103.

MARLER J.H. & BOUDREAU J.W. (2017), "An evidence-based review of HR Analytics", *The International Journal of Human Resource Management*, Vol. 28, n°1, p. 3-26.

MENON V.M. & RAHULNATH H.A. (2016), "A novel approach to evaluate and rank candidates in a recruitment process by estimating emotional intelligence through social média data", *2016 International Conference on Next Generation Intelligent Systems (ICNGIS)*, p. 1-6.

NG E.S.W. & SEARS G.J. (2010), "The effect of adverse impact in selection practices on organizational diversity : A field study", The *International Journal of Human Resource Management*, Vol. 21, n°9, p. 1454-1471.

O'NEIL C. (2017), *Weapons of math destruction : How big data increases inequality and threatens democracy*, Crown, New York.

PARASURAMAN R. & RILEY V. (1997), "Humans and automation : Use, misuse, disuse, abuse", *Human factors*, Vol. 39, n°2, p. 230-253.

PASQUALE F. (2015), *The black box society*, Harvard University Press, Cambridge.

PEARL J. & MACKENZIE D. (2018), *The book of why : The new science of cause and effect*, Basic Books.

PILLEMER J., GRAHAM E.R. & BURKE D.M. (2014), "The face says it all : CEOs, gender, and predicting corporate performance", *The Leadership Quarterly*, Vol. 25, n°5, p. 855-864.

RULE N.O. & AMBADY N. (2011), "Face and fortune : Inferences of personality from Managing Partners' faces predict their law firms' financial success", *The Leadership Quarterly*, Vol. 22, n°4, p. 690-696.

RULE N.O., ISHII K. & AMBADY N. (2011), "Cross-cultural impressions of leaders' faces : Consensus and predictive validity", *International Journal of Intercultural Relations*, Vol. 35, n° 36, p. 833-841.

SCHMIDT F.L. & HUNTER J.E. (1998), "The validity and utility of selection methods in personnel psychology : Practical and theoretical implications of 85 years of research findings", *Psychological bulletin*, Vol. 124, n°2, p. 262-274.

SCHMIDT F.L., OH I.S. & SHAFFER J.A. (2016), "The Validity and Utility of Selection Methods in Personnel Psychology : Practical and Theoretical Implications of 100 Years…", F*ox School of Business Research Paper*, p. 1-74.

SEARS G.J., ZHANG H., WIESNER W.H., HACKETT R.D. & YUAN Y. (2013), "A comparative assessment of videoconference and face-to-face employment interviews", *Management Decision*, Vol. 51, n°8, p. 1733-1752.

SERMONDADAZ S. (2018), « Yann LeCun : L'IA a moins de sens commun qu'un rat », *Sciences et Avenir* (site web).

STEELE C.M. & ARONSON J. (1995), "Stereotype threat and the intellectual test performance of African Americans", *Journal of personality and social psychology*, Vol. 69, n°5, p. 797-811.

STOUGHTON J.W., THOMPSON L.F. & MEADE A.W. (2015), "Examining Applicant Reactions to the Use of Social Networking Websites in Pre-Employment Screening", *Journal of Business and Psychology*, Vol. 30, n°1, p. 73-88.

SUEN H.Y., CHEN M.Y.C. & LU S.H. (2019), "Does the use of synchrony and artificial intelligence in video interviews affect interview ratings and applicant attitudes ?", *Computers in Human Behavior*, n°98, p. 93-101.

TAMBE P., CAPPELLI P. & YAKUBOVICH V. (2019), "Artificial Intelligence in Human Resources Management : Challenges and a Path Forward", *California Management Review*, Vol.61, n°4), 15-42. https ://doi.org/10.1177/0008125619867910

TODOROV A., OLIVOLA C.Y., DOTSCH R. & MENDE-SIEDLECKI P. (2015), "Social Attributions from Faces : Determinants, Consequences, Accuracy, and Functional Significance", *Annual Review of Psychology*, Vol. 66, n°1, p. 519-545.

TVERSKY A. & KAHNEMAN D. (1974), "Judgment under Uncertainty : Heuristics and Biases", *Science*, Vol. 185 n°4157, p. 1124-1131.

VAN HOYE G. (2013), "Word of mouth as a recruitment source : An integrative model". In *The OxfordHandbook of Recruitment*, K. Y. T.Yu & D. M.Cable (Eds.), Oxford University Press, Oxford.

VAN IDDEKINGE C.H., LANIVICH S.E., ROTH P.L. & JUNCO E. (2016), "Social média for selection ? Validity and adverse impact potential of a Facebook-based assessment", *Journal of Management*, Vol. 42, n° 7, p. 1811-1835.

VILLANI C., SCHOENAUER M., BONNET Y., BERTHET C., CORNUT A.-C., LEVIN F., & RONDEPIERRE B. (2018), « Donner un sens à l'intelligence artificielle » (236p.) [Rapport de la Mission Villani sur l'intelligence artificielle].

WOODS S.A., AHMED S., NIKOLAOU I., COSTA A.C. & ANDERSON N.R. (2019), "Personnel selection in the digital age : A review of validity and applicant reactions, and future research challenges", *European Journal of Work and Organizational Psychology*, p. 1-14.

WOODS S.A. & WEST M.A. (2019), *The Psychology of Work and Organizations*, Cengage Learning EMEA, Andover.