L’intelligence artificielle est-elle plus impartiale que l’humain ?

Les algorithmes, censés être basés sur des critères purement mathématiques, sont-ils plus objectifs que les humains remplis de préjugés ?

Temps de lecture : minute

28 avril 2018

L’humain est profondément injuste et peu fiable. Les médecins peuvent faire des erreurs de diagnostic, les juges remettent en liberté des criminels en raison d'un vice de procédure et la police effectue des contrôles au faciès. Aux États-Unis, le risque d’être condamné à la peine capitale est par exemple quatre fois plus élevé lorsque la victime est blanche que lorsqu’elle est noire, a montré le Centre d’information sur la peine de mort. Les femmes ont elles 25% moins de chances que les hommes d'obtenir une augmentation de salaire lorsqu'elles la demandent. La plupart des profs avouent avoir leurs "chouchous" en classe et les favoriser discrètement.

L’IA, étape ultime du CV anonyme

Face à de telles failles, l’intelligence artificielle nous promet un monde régi par des jugements totalement objectifs, non soumis à un quelconque préjugé ou émotion. De plus en plus de secteurs ont ainsi recours à des algorithmes pour gérer des problèmes difficiles à arbitrer. Le recrutement prédictif, qui anticipe la capacité d’un candidat à répondre aux besoins de l’entreprise sans passer par son CV, est ainsi en plein boom. À travers une série de questions, l’ordinateur parvient à détecter les capacités de raisonnement, la motivation et même la personnalité. Un système qui

évite un jugement biaisé par l'origine ou le parcours du candidat. En 2014, une étude de la Harvard Business Review a ainsi montré que les algorithmes dépassent de 25% l'instinct humain concernant la prédiction de la performance d'un candidat.

C'est pour cela que, selon une étude canadienne, un quart des employés préféreraient être dirigés par une intelligence artificielle plutôt qu'un humain dans leur entreprise. "*Les gens perdent confiance dans le management humain, et à juste titre*, explique le futurologue Nikolas Badminton. *En qui auriez-vous plutôt confiance : un être humain ayant des opinions et des préjugés personnels ou en une intelligence artificielle impartiale et équilibrée ?*" Selon lui, ce type de ressources humaines "automatisées" vont fortement se déployer d'ici 3 à 5 ans.

Mieux évaluer le risque de récidive

Mais c'est surtout dans le très délicat domaine judiciaire que l'intelligence artificielle promet de révolutionner les pratiques. La France accuse en effet un énorme retard par rapport aux pays anglo-saxons. "*Les experts psychiatres s'appuient encore largement sur leur propre jugement ou leur expérience personnelle pour évaluer le profil des prévenus ou condamnés*", regrette un rapport de deux spécialistes en droit pénal, Virginie Gautron et Émilie Dubourg.

À lire aussi

Quels sont les composants d’un bon projet d’intelligence artificielle ?

Or, il est reproché à cette approche "*de produire des estimations proches du hasard, de surévaluer les risques de récidive, de se fonder sur des concepts psychanalytiques imprécis*". Résultat : des décisions très variées selon les experts pour un même cas et des polémiques à répétition. Chacun se souvient de l’affaire Agnès Martin, la jeune fille violée et tuée en 2011 par un camarade de classe, déjà condamné pour des faits similaires, et à propos duquel la cour d'appel de Montpellier avait pourtant conclu à "*l’absence de dangerosité*".

Les pays anglo-saxons ont eux déjà largement adopté des "méthodes actuarielles", reposant sur des éléments objectifs et statistiques. Aux États-Unis, 29 tribunaux ont déjà adopté un système permettant de décider qui doit rester en prison en attente de son jugement, en fonction de sa probabilité à commettre un autre crime ou à chercher à échapper à

la justice. Le suspect se voit attribué un "score" en fonction de nombreux critères : détient-il une voiture ? Occupe-t-il un emploi ? Est-il récidiviste ? Consomme-t-il des stupéfiants ? Selon ses concepteurs, le système permet d'éviter des emprisonnements abusifs coûteux et infondés.

Quand l'IA copie-colle les défauts humains

Mais ces systèmes sont-ils réellement plus "objectifs" ? Un autre algorithme développé par la compagnie privée Northpointe, utilisé aux États-Unis pour déterminer le montant des cautions et la durée de peines, montre lui aussi des biais raciaux. Parmi les non récidivistes, 42% des Noirs étaient ainsi classés "à risque" contre seulement 22% des Blancs.

Car l'intelligence artificielle se nourrit avant tout… des données qui lui ont été fournies. Or, ces données sont elles-mêmes entachées des stéréotypes humains. Une étude publiée dans la revue Science en 2017 a ainsi montré que les algorithmes de reconnaissance du langage finissent par incorporer des biais couramment observés. Le mot "homme" est par exemple plus souvent associé à des termes évoquant la carrière et des postes de dirigeants dans les domaines de la science ou l'ingénierie, tandis que le mot "femme" renvoie vers des postes d'assistantes et les secteurs des arts et les lettres, quand ce n'est pas vers la maison et la famille. De même, les noms à consonance européenne ou américaine sont associés à des termes élogieux quand les noms afro-américains renvoient vers des termes bien plus négatifs. Lorsqu'il s'agira de choisir un candidat pour un ingénieur nucléaire, par exemple, le CV d'une femme noire aura tendance à passer directement sous la pile.

Autre exemple : en 2016, un concours de beauté jugé par un robot a désigné en grande majorité des gagnants à la peau blanche car c'est le type de fille le plus souvent jugées "belles" par leurs pairs. La question de l'échantillonnage est donc cruciale. "*Les échelles de dangerosité ont été construites sur des échantillons composés d'hommes blancs, de sorte que*

leur capacité prédictive concernant les femmes et les minorités est loin d'être assurée", note Virginie Gautron.

"Appliquer la loi de façon rigide peut aboutir à de profondes injustices"

Au-delà de ces biais statistiques se pose une question morale. "*La justice n'est pas faite pour être parfaite*", met ainsi en garde Ugo Bellagamba, maître de conférence en histoire du droit à l'université de Nice. "*Elle doit s'adapter aux cas particuliers [...] Appliquer la loi de façon rigide peut aboutir à de profondes injustices*", estime-t-il. "*Est-il légitime d'utiliser toutes les données qui ont un caractère prédictif dès lors que celles-ci sont neutres idéologiquement ?*", s'interroge de son côté Pirmin Lemberger, data scientist et fondateur du cabinet de conseil Weave. "*Les algorithmes sont des objets mathématiques, à la fois opaques et enduits d'un vernis d'objectivité scientifique. Ceci peut leur conférer un caractère intimidant propre freiner leur examen critique par des non spécialistes*", met-il en garde.

On peut s'interroger sur la pertinence d'une étude parue en septembre 2017 intitulée *Les réseaux de neurones plus fiables que les humains pour détecter l'orientation sexuelle à partir d'images de visages*. La prétendue performance des algorithmes ne doit donc pas forcément aboutir à leur généralisation, d'autant plus si elle s'applique à des domaines avec d'importants enjeux. Doit-on par exemple considérer qu'un essai de médicament contre la maladie d'Alzheimer doit d'abord être destiné aux femmes, qui représentent 60% des malades, ou que les deux sexes doivent bénéficier des mêmes "chances" d'accéder au traitement ? Il faudra bien trouver le juste milieu entre la coutumière imperfection humaine et le caractère impotoyable de la pure statistique.

Article écrit par Céline Deluzarche