Lorsque l’intelligence artificielle est discriminatoire

16 mai 2017

Auteur

Eric Lavallée

Associé, Agent de marques de commerce Associé, et Avocat

L’intelligence artificielle a connu des avancées importantes depuis quelques années, notamment grâce aux avancées de ce que l’on nomme maintenant l’apprentissage profond (deep learning)[1]. Cette méthode est le prolongement des réseaux neuroniques qui sont utilisés depuis quelques années pour l’apprentissage des machines. L’apprentissage profond, comme toute forme d’apprentissage d’une machine, requiert que le système d’intelligence artificielle soit confronté à diverses situations afin d’apprendre à réagir à des situations présentant des similitudes à ces expériences antérieures.

En entreprise, des systèmes d’intelligence artificielle sont notamment utilisés pour répondre aux besoins des clients, soit directement, soit en soutenant les employés dans leurs interventions. La qualité des services rendus par l’entreprise est donc de plus en plus tributaire de la qualité de ces systèmes d’intelligence artificielle.

Il ne faut pas, toutefois, faire l’erreur de présumer qu’un tel système informatique s’acquittera automatiquement des tâches qui lui sont confiées sans faille et dans le respect des valeurs de l’entreprise ou de sa clientèle.

Par exemple, des chercheurs de l’université Carnegie Mellon ont récemment démontré qu’un système devant présenter de la publicité ciblée à des usagers d’Internet offrait systématiquement moins de postes bien rémunérés aux femmes qu’aux hommes[2]. En d’autres termes, ce système avait un comportement que l’on pourrait qualifier de sexiste. Bien que les chercheurs n’aient pu identifier l’origine du problème, ils étaient d’avis qu’il s’agissait probablement d’une perte de contrôle du fournisseur de service de placement de publicité sur son système automatisé, et ils soulignaient

les risques inhérents aux systèmes d'intelligence artificielle à grande échelle.

Divers systèmes d'intelligence artificielle ont connu des ratés similaires, démontrant des comportements racistes et forçant même un exploitant à suspendre l'accès à son système[3]

À cet égard, l'Union Européenne a adopté en avril 2016 une règlementation relative au traitement de l'information personnelle qui, sauf dans certains cas précis, interdit la prise de décision automatisée basée sur certaines données à caractère personnel, dont « […] l'origine raciale ou ethnique, les opinions politiques, les convictions religieuses ou philosophiques ou l'appartenance syndicale, ainsi que le traitement des données génétiques, des données biométriques aux fins d'identifier une personne physique de manière unique, des données concernant la santé ou des données concernant la vie sexuelle ou l'orientation sexuelle d'une personne physique […] »[4]. Certains chercheurs s'interrogent d'ailleurs sur l'application de ce règlement, notamment alors que la discrimination apparaît de manière incidente, hors la volonté de l'exploitant du système d'intelligence artificielle[5]

Au Québec, on peut croire qu'une entreprise qui exploiterait un système d'intelligence artificielle discriminatoire au sens des lois applicables ou de la Charte des droits et libertés de la personne s'exposerait à des recours, même en l'absence de règlement précis comme celui de l'Union Européenne. En effet, le responsable du bien qu'est ce système d'intelligence artificielle pourrait voir sa responsabilité engagée à l'égard du préjudice ou du dommage causé par le fait autonome de ce bien. Qui plus est, le fait de ne pas mettre en place des mesures raisonnables pour éviter la discrimination serait fort probablement pris en compte dans l'analyse juridique d'une telle situation.

Une vigilance particulière s'impose donc lorsque le fonctionnement d'un système d'intelligence artificielle repose sur des données déjà accumulées au sein de l'entreprise, des données de tiers (notamment ce qu'on désigne souvent comme le big data), ou encore lorsque les données seront fournies au système d'intelligence artificielle par des employés de l'entreprise ou ses utilisateurs au cours d'une période « d'apprentissage ». Toutes ces sources de données, par ailleurs soumises aux obligations découlant des lois applicables à la protection des renseignements personnels, peuvent être biaisées à divers degrés.

Les effets d'un échantillonnage biaisé ne sont pas nouveaux ni cantonnés au respect des droits de la personne. Il s'agit d'un effet bien connu des statisticiens. Pendant la Seconde Guerre mondiale, la marine américaine demanda à un mathématicien nommé Abraham Wald de leur fournir des statistiques sur les parties des avions bombardiers ayant été les plus touchés dans le but de renforcer le blindage à ces endroits. Wald démontra que les données sur les avions revenant de mission étaient biaisées, car elles ne tenaient pas compte des avions abattus en mission. Il ne fallait donc pas renforcer le blindage sur les parties où les avions revenant de mission étaient endommagés, mais plutôt aux endroits où ils n'étaient pas touchés.

Dans le contexte de l'exploitation d'une entreprise, un système d'intelligence artificielle auquel on soumet des données biaisées pourrait ainsi prendre des décisions erronées, ayant des conséquences néfastes au point de vue humain, économique et opérationnel pour l'entreprise.

Par exemple, si l'on soumet un tel système à un apprentissage encadré par des employés de l'entreprise, leur façon d'agir se reflétera sans doute dans son comportement ultérieur. Ceci peut se refléter dans les jugements portés par le système d'intelligence artificielle à l'égard de demandes de clients, mais aussi directement dans leur capacité de résoudre adéquatement les problèmes techniques qui lui sont soumis. On risque ainsi de perpétuer les comportements problématiques de certains employés.

Des chercheurs du Machine Intelligence Research Institute ont proposé diverses approches pour minimiser les risques et rendre l'apprentissage machine d'un système d'intelligence artificielle

conforme aux intérêts de son exploitant [6]. Selon ces chercheurs, il pourrait notamment être opportun d'adopter une approche prudente quant aux objectifs imposés à de tels systèmes pour éviter qu'ils offrent des solutions extrêmes ou indésirables. Il serait en outre important d'établir des procédures de supervision informée, au moyen desquelles un opérateur peut s'assurer que le fonctionnement du système d'intelligence artificielle est, dans son ensemble, conforme aux attentes de son exploitant.

De ce qui précède, il faut retenir qu'une entreprise qui désire intégrer un système d'intelligence artificielle dans ses opérations doit prendre très au sérieux la phase d'implantation, au cours de laquelle se déroulera l'apprentissage du comportement désiré par le système. D'une part, il sera important d'avoir des discussions approfondies avec le fournisseur sur le fonctionnement de sa technologie et ses performances, ainsi que d'encadrer contractuellement et le plus clairement possible les attentes de l'entreprise à l'égard du système qu'elle désire implanter. Il faut également prévoir comment s'effectuera l'intégration du système d'intelligence artificielle dans l'entreprise et s'assurer que cette intégration soit confiée à des employés et consultants dignes de confiance, possédant le plus haut niveau de compétence eu égard aux tâches pertinentes.

Quant au fournisseur du système d'intelligence artificielle, il faudra généralement s'assurer que les données qui lui sont fournies ne sont pas biaisées, inexactes ou autrement altérées, de telle façon que les objectifs prévus au contrat quant aux performances souhaitées du système puissent raisonnablement être atteints, permettant ainsi de minimiser le risque de litiges devant les tribunaux découlant de comportements discriminatoires ou défaillants à d'autres égards du système d'intelligence artificielle. Non seulement de tels litiges pourraient-ils s'avérer onéreux, ils seraient de surcroît susceptibles d'entacher tant la réputation du fournisseur que celle du client utilisateur.

---

1. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.
2. Datta, A., Sen, S., & Zick, Y. (2016, May). Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. *In Security and Privacy (SP), 2016 IEEE Symposium on* (pp. 598-617). IEEE; Datta, A., Tschantz, M. C., & Datta, A. (2015). Voir aussi: Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies*, 2015(1), 92-112.
3. Reese, H. (2016). Top 10 AI failures of 2016. Le cas de Tay, le système de Microsoft, a été abondamment discuté dans les médias.
4. Règlement (UE) 2016/679 du parlement européen et du conseil du 27 avril 2016, relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant la directive 95/46/CE (règlement général sur la protection des données), art. 22.
5. Goodman, B., & Flaxman, S. (2016, June). EU regulations on algorithmic decision-making and a "right to explanation". *In ICML Workshop on Human Interpretability in Machine Learning* (WHI 2016).
6. Taylor, J., Yudkowsky, E., LaVictoire, P., & Critch, A. (2016). *Alignment for advanced machine learning systems* . Technical Report 20161, MIRI.