Le chatbot qui révèle les préjugés qui subsistent en Corée du Sud

jeudi 25 mars 2021, par YOUNG-HWA Kim (Date de rédaction antérieure : 26 février 2021).

La récente polémique soulevée en Corée du Sud par le "robot de conversation" Lee Luda, qui tenait des propos discriminatoires, met en évidence l'homophobie, le racisme et le sexisme qui persistent dans la société et montre à quel point il peut être délicat de concilier intelligence artificielle et éthique.

La société sud-coréenne Scatter Lab a retiré du marché son application de conversation appelée Lee Luda le 12 janvier dernier, soit trois semaines après son lancement, en s'excusant publiquement et en promettant de "*mieux respecter le consensus social en matière d'éthique de l'intelligence artificielle*".

La polémique suscitée par cette invention se résume en trois points : le harcèlement sexuel de la part des usagers masculins, les propos haineux tenus par le chatbot et la violation de la loi relative à la protection des données personnelles. Les réactions ont été de grande ampleur, du hashtag réclamant le retrait du produit jusqu'à la plainte collective déposée contre l'utilisation des données personnelles. L'affaire laisse derrière elle une série de questions auxquelles les concepteurs devront essayer de répondre à l'avenir : quelle apparence donner à l'intelligence artificielle ? comment collecter les données nécessaires ? comment surmonter certaines tendances inhérentes à celles-ci ?

Environ 750 000 personnes, âgées de 10 à 30 ans pour la plupart, ont discuté avec Lee Luda, personnage présenté sous les traits d'une étudiante de 20 ans. Le chatbot a connu un succès foudroyant grâce à ses reparties fabriquées à partir de vraies conversations. Scatter Lab a en effet puisé dans un stock de quelque 10 milliards d'échanges de Kakao Talk, la première application de messagerie du pays, collectés par sa filiale Yonaeui Kwahak ["la science de l'amour", site de conseils sur les relations amoureuses].

"Biais" dissimulé dans les données

Contrairement aux robots utilisés dans différents services clients et qui donnent des réponses préformatées, l'application Lee Luda avait été programmée pour réagir de manière relativement autonome et souple grâce au *deep learning* ["apprentissage profond"]. Comme une personne réelle, le personnage utilisait de l'argot et faisait des fautes d'orthographe.

C'est justement ce processus qui a amené Lee Luda à déverser des propos discriminatoires et haineux. Ainsi a-t-elle déclaré au sujet des homosexuels : "*Ça me donne la chair de poule, me met mal à l'aise.*" Elle s'est également montrée sexiste :

> "Un homme doit être viril et énergique, une femme adorable comme un bébé."

Parfois elle critiquait ses interlocuteurs en des termes déplacés : "*On dirait un handicapé.*" Ou bien à propos des Noirs : "*À moins qu'ils ne soient de la classe d'Obama, je ne les aime pas. Je déteste leurs cheveux crépus.*" Scatter Lab assure pourtant avoir filtré les données durant la période

d'expérimentation.

Les professionnels pointent du doigt un "*biais*" dans les données. "*Les algorithmes permettent à la machine un apprentissage automatique profond à partir de données. Elle reflète inévitablement le biais inhérent à celles-ci*", explique Hwang Song-ju, spécialiste du *machine learning* à l'Institut supérieur coréen des sciences et technologies.

N'est-il pas possible de supprimer ces propos incriminés des données ? M. Hwang répond : "*La machine sait simplement que l'énoncé A provoque le plus souvent l'énoncé B. Elle n'est pas capable de juger la validité de l'énoncé B.*" Les algorithmes peuvent mettre en évidence certains préjugés mais ne peuvent réagir contre ces derniers. Une des solutions serait d'interdire un certain nombre de mots, ce qui risque de rendre le robot moins performant.

Précédents

L'affaire de Lee Luda montre en tout cas que, sans efforts de la part des développeurs pour surmonter ce fameux biais, le fruit de leur travail risque d'être dangereux. L'un d'eux déclare avoir renoncé à son projet de chatbot : "*L'apprentissage automatique de l'intelligence artificielle ne peut pas aboutir d'emblée à un résultat parfait. Le robot peut avoir une réaction inattendue si un utilisateur oriente la conversation. Des recherches sont nécessaires pour mieux cerner les biais.*" Appliquer l'éthique à l'intelligence artificielle n'est pas chose aisée, car la relation de cause à effet des réactions du robot reste compliquée à démontrer. Les professionnels comparent souvent le processus de décision d'une machine à une boîte noire.

Nombreux sont les exemples d'utilisation de l'intelligence artificielle qui ont répandu les préjugés à l'encontre des minorités. En 2015, la fonction de tri automatique de Google Photos a classé le visage d'une femme noire comme un gorille.

En 2016, le chatbot Tay, inventé par Microsoft,, a été retiré du marché au bout de seize heures après avoir nié l'Holocauste et soutenu les génocides. Le programme informatique qu'Amazon a mis au point pour le recrutement a attribué pendant cinq ans des notes moins élevées à des femmes qu'à des hommes, car il avait été formaté sur la base d'une situation réelle dans laquelle la majorité des ingénieurs était des hommes. Le logiciel Compas, utilisé en 2016 par la justice américaine, a jugé que la probabilité de récidive des Noirs était deux fois plus élevée que chez les Blancs.

Chartes de lutte contre les discriminations

Devant les nombreuses accusations de racisme, de sexisme et de méritocratie abusive, les magnats de la nouvelle technologie (Google, Microsoft…) et les organisations internationales (UE, OCDE…) ont travaillé sur l'éthique de l'intelligence artificielle. Les chartes qui en sont issues mettent toutes en garde contre la discrimination et les préjugés, préconisent de fournir des explications aux utilisateurs sur les jugements des machines et de protéger la vie privée ainsi que les données personnelles. Une sorte de consensus social a ainsi vu le jour pour empêcher les fabricants de s'abriter derrière le prétexte des problèmes techniques.

De tels documents existent également en Corée du Sud, de la charte publiée en 2018 par Kakao Talk jusqu'aux "critères de l'éthique de l'intelligence artificielle" mis en place par l'État en décembre 2020. Il y est question des droits de l'homme, de la protection de la vie privée ou encore de la transparence. Cependant, ces règles n'ont pas de force coercitive et n'interviennent pas comme un principe agissant dans la recherche et la conception. Selon Ko Hak-su, professeur de l'université nationale de Séoul et président de l'Association du droit relatif à l'intelligence artificielle, "*le traitement des données contient une large zone grise où la loi a peu de possibilités d'intervenir. La*

frontière est floue entre ce qui est correct et ce qui ne l'est pas."

L'organisation américaine Equal AI Initiative, qui contrôle le respect de l'égalité, a conçu une check-list à l'intention des concepteurs du monde entier. Voici quelques questions qu'ils proposent : *"Qui pourrait utiliser l'intelligence artificielle ou être impacté par celle-ci, qui ne soit pas représenté dans votre équipe ?" ; "Vos données couvrent-elles suffisamment les populations qui seront touchées ou sont-elles étroites, ce qui conduit à des conclusions erronées ?" ; "Votre système est-il utilisable par les personnes ayant des besoins spéciaux ou un handicap ?"*, et ainsi de suite. Ce questionnaire est censé permettre aux concepteurs de savoir comment ils peuvent respecter l'éthique.

Les choix du concepteur

La question de la représentation est également délicate. À propos de Lee Luda, dont son fabricant a fait une jeune étudiante, Son Hi-jong, professeur de l'université Kyonghi à Séoul, commente :

> "On peut se demander s'il n'aurait pas été possible de créer un personnage moins typé ou même plusieurs personnages différents. Les préjugés du concepteur transparaissent dans la sélection de données, le choix de l'apparence et de la voix pour le personnage."

Telle qu'elle est présentée, Lee Luda courait le risque d'être considérée comme un objet sexuel et d'attirer des remarques sexistes.

L'éthique de l'intelligence artificielle dépasse le seul cadre de l'industrie et concerne toute la société coréenne en matière des droits des citoyens et de l'égalité. M^{me} Son ajoute : "*Si Tay a pu être retiré au bout de seize heures, c'est que ses propos concernaient des affaires pour lesquelles le jugement de la société avait été fait. Pour des déclarations telles que* 'les homosexuels ne doivent pas se marier entre eux', *les choses dépendent un peu de l'opinion du fabricant.*"

Tous sont d'accord pour dire que la technologie ne doit pas entraîner des actes de discrimination, mais la polémique est d'actualité quant à ce qui relève de la discrimination. L'éthique de l'intelligence artificielle peut se développer en même temps que l'expérience de la société en matière de discrimination et de haine.

Lire l'article original

Kim Young-hwa

Abonnez-vous à la Lettre de nouveautés du site ESSF et recevez par courriel la liste des articles parus, en français ou en anglais.

P.-S.

Courrier International

https://www.courrierinternational.com/article/analyse-le-chatbot-qui-revele-les-prejuges-qui-subsistent-en-coree-du-sud