texte

editorial

Biais dans les systèmes d?Intelligence Artificielle d'aide à la décision

Dans la cadre de la révision de la loi relative à la bioéthique, qui aura lieu en 2018 et 2019, l'Espace éthique/IDF propose une série de textes, réflexions et expertises pour animer le débat public. Chaque intervention visera à éclairer un point, une perspective ou un enjeu des révisions de la loi. Le second texte aborde les enjeux éthiques du développement de l'Intelligence artificielle.

Par: Laurence Devillers, Professeur d? Intelligence artificielle en Sorbonne, directrice de l?équipe « Dimensions affectives et sociales dans les interactions parlées », département Communication Homme-Machine, LIMSI-CNRS/Paris-Saclay, membre de la CERNA* d?Allistène /

Publié le : 09 Janvier 2018

Poursuivre la réflexion

La diffusion des progrès technologiques dépend de facteur de rentabilité économique, mais aussi de choix politique et sociétaux. Un débat scientifique émerge aujourd?hui sur les systèmes d?intelligence artificielle d?aide à la décision et sur la transparence et l?explicabilité des données et des algorithmes et pose ainsi de façon pressante la question d?une réflexion éthique dans la société.
L?intelligence artificielle (IA) permet de construire des machines capables de prendre des décisions dans un environnement incertain selon des connaissances construites à l'aide de techniques d?apprentissage. Comprendre comment les concevoir et la portée des prédictions de ces machines sont nécessaires pour bien les utiliser. Loin d?être des freins à son déploiement, les considérations éthiques permettent de construire des solutions viables qui seront mieux acceptées dans la société.

L?intelligence artificielle s?installe dans nombreux secteurs d?activités, impacte le travail et les services dans la santé, mais aussi dans de nombreux autres domaines : l?énergie, le transport, la banque, le commerce, etc. et impacte également notre bien-être dans la société. Elle automatise sur des ordinateurs, téléphones ou robots, des capacités cognitives telles que la perception, l?apprentissage, la mémoire, le raisonnement et les fonctions de communication. L'apprentissage machine est une façon de remplacer une spécification que nous ne savons pas écrire par une masse de données. Prédire un phénomène à partir d?observations passées présuppose un mécanisme causal. Expliquer ce mécanisme n?est pas toujours facile. L?apprentissage machine est une approche statistique permettant de découvrir des corrélations significatives dans une masse importante de données pour construire un modèle prédictif quand il est difficile de construire un modèle explicatif. Il est alors impossible de connaître les causes des decisions.

Récemment, ce domaine a explosé avec l?apprentissage profond, *deep learning*, ensemble d?algorithmes d'apprentissage machine appliqués à des réseaux de neurones ayant un grand nombre de couches cachées et des millions de paramètres permettant de modéliser avec un haut niveau d'abstraction des données. Par exemple, pour qu'un programme de *deep learning* puisse reconnaître un lion, on l?entraîne à apprendre des milliers d'images de lion, étiquetées comme lion. Cet apprentissage supervisé (chaque image est étiquetée) peut nécessiter beaucoup de temps d?entraînement. Ensuite le système est capable de reconnaître des lions sur de nouvelles images. Ces machines sont très performantes et peuvent même nous dépasser dans l?exécution de tâches spécifiques. Elles sont utilisées par les GAFAMI (Google Amazon Facebook Apple Microsoft IBM) ou les géants du Web chinois : Baidu, Alibaba, Tencent et Xiaomi (BATX) pour reconnaître la parole, traduire un discours, reconnaître des anomalies dans des images médicales ou beaucoup d?autres tâches. Ces systèmes d?intelligence artificielle sont souvent complexes et opaques. En 2016, AlphaGo de Google Deepmind qui a battu un champion du monde de Go comporte deux programmes de *deep learning* mais également un algorithme d?apprentissage de renforcement et de Monte Carlo. Ces systèmes appelés « boites noires » ont des capacités de classification et de prédiction importantes malgré le fait *d?apprendre sans comprendre*. La transparence et l?explicabilité des systèmes sont des défis actuels de recherche : les systèmes peuvent faire des erreurs que nous ne savons pas toujours expliquer car les connaissances sont distribuées sur un très grand nombre de paramètres. Par exemple, prenons l?exemple d?un système de reconnaissance de photo entrainé avec un algorithme de *deep learning* si vous enlevez les pixels de l??il d?un lion, l?image peut ne plus être reconnue ou encore l?image d?une savane peut être reconnue comme lion, sans doute à cause de l?arrière plan des images. Construire des corpus de données d?apprentissage en respectant équité, transparence, fidélité, diversité et neutralité permet de garantir un comportement ?loyal?des systèmes. Il est également souhaitable de pouvoir tracer et expliquer ce que fait l?algorihtme, d?autant plus si le système peut s?adapter au cours de son utilisation en apprenant de son environnement, ou grâce à l?interaction qu?il a avec les humains. Il est alors utile de verifier tout au long de son utilisation ses performances. La Commission de réflexion sur l?éthique de la recherche (CERNA) d?Allistène a produit un rapport sur l?Ethique de la recherche sur l?apprentissage machine.

Pour accompagner le changement et trouver des solutions innovantes et éthiques d?intelligence artificielle, plusieurs intiatives sont en cours :

- des missions de réflexions et de consultations autour de l?IA: la mission Villani, la

mission de la CNIL et du CNNUM, la CCNE sur la bioéthique
- mais aussi la mise en place d?écosystèmes pour outiller l?IA : des collaborations privées/public dans le HUB IA (#FranceIA), une plateforme nationale de recherche sur la transparence des algorithmes (TransAlgo) ainsi qu?un institut convergence de recherche pluridisciplinaire sur la science des données DATAIA sur Paris-Saclay.

Sommaire