Intelligence artificielle explicable – Fiche d'information

L'intelligence artificielle explicable accroît la confiance à l'égard des solutions d'intelligence artificielle (IA), ce qui permet une adoption plus large et offre un fondement pour la gouvernance de l'IA.

INTRODUCTION

Les humains ont la capacité d'expliquer leur processus décisionnel – raisonnement logique, observation, intuition ou expérience. Les algorithmes de base de l'apprentissage machine, comme les arbres décisionnels, peuvent être compris en suivant le chemin logique qui a mené à la décision. Toutefois, les systèmes d'intelligence artificielle complexes ont traditionnellement été des « boîtes noires » – même si l'on connaissait les intrants et les extrants, il n'y avait aucune compréhension des algorithmes utilisés pour en arriver à des décisions. Pour que les humains aient confiance en ces décisions, la justification et les algorithmes de l'IA doivent être explicables.

INTELLIGENCE ARTIFICIELLE EXPLICABLE

L'intelligence artificielle explicable est une intelligence artificielle qui peut décrire son but, sa justification et son processus décisionnel d'une façon que les humains peuvent comprendre.

L'intelligence artificielle explicable peut être obtenue de deux façons :

- en intégrant la capacité dans le modèle d'IA dès le début; ou
- en utilisant des outils qui décrivent la façon dont la décision a été prise après les faits.

Idéalement, l'intelligence artificielle explicable pourrait fournir des renseignements sur ce qui suit :

- intrants : la façon dont les données ont été sélectionnées, la façon dont les données utilisées aux fins d'apprentissage diffèrent des données de production, la façon dont elles ont été examinées en vue d'évaluer la partialité et la façon dont la partialité a été atténuée;
- processus : les modèles qui ont été appliqués, la raison pour laquelle plus de poids a été accordé à certains plutôt qu'à d'autres, et la façon dont ils ont été mis à l'essai;
- extrants : la façon dont les intrants et les modèles mènent aux extrants, et la façon dont les résultats ont fait l'objet d'une vérification;
- rendement : surveillance et production de rapports sur la façon dont le système fonctionne dans l'environnement de production.

POURQUOI L'INTELLIGENCE ARTIFICIELLE EXPLICABLE EST-ELLE IMPORTANTE?

La complexité et la nature exclusive de certains algorithmes peuvent obscurcir la façon dont ils prennent des décisions et peuvent, par conséquent, masquer les comportements nuisibles. La partialité des données sur la formation pourrait avoir des effets néfastes sur l'emploi, l'immigration ou les décisions financières. Les mauvaises décisions peuvent porter atteinte à la vie privée, nuire à la santé personnelle ou occasionner des incidents militaires désastreux.

Les décisions non valides prises au moyen de l'intelligence artificielle peuvent entraîner des blessures pour les humains, des défis opérationnels pour une organisation, une insatisfaction des clients, une attention médiatique négative, un taux élevé de roulement des employés ou un examen réglementaire. Lorsqu'il y a des erreurs, il est important de comprendre si elles ont été causées par des données non valides, des bogues dans le logiciel, des modèles inexacts ou une combinaison de ces éléments. Même si la décision semble plausible, les humains doivent la comprendre, notamment pour assouvir leur curiosité, assurer la tranquillité d'esprit et garantir la conformité juridique, ainsi que pour des raisons éthiques.

Pour toutes ces raisons, le manque de transparence a une incidence négative sur la confiance à l'égard de l'IA et sur son acceptation. Le nombre croissant d'aspects juridiques et d'éléments liés à la confidentialité à prendre en compte limitera également l'application de l'IA si la prise de décisions ne peut être expliquée. Puisque l'intelligence artificielle devient de plus en plus répandue, il est de plus en plus important de divulguer la façon dont la partialité et la question de la confiance sont traitées.

Gouvernance émergente

Le Règlement général sur la protection des données (RGPD) de l'Union européenne stipule que lorsqu'un humain est assujetti à une décision rendue au moyen de

l'intelligence artificielle, l'humain a le droit de recevoir une explication, de contester la décision et d'être protégé contre le profilage.

En janvier 2019, la Commission européenne a lancé un projet d'intelligence artificielle pour l'Union européenne dans le but de bâtir la première plateforme d'intelligence artificielle sur demande en Europe et de promouvoir le développement d'une IA vérifiable et explicable.

En avril 2019, le gouvernement du Canada a publié une directive sur la prise de décision automatisée afin de « veiller à ce que les systèmes décisionnels automatisés soient déployés d'une manière qui permet de réduire les risques pour les Canadiens et les institutions fédérales, et qui donne lieu à une prise de décisions plus efficace, exacte et conforme, qui peut être interprétée en vertu du droit canadien[1]. » La directive renferme un outil d'évaluation de l'incidence algorithmique (AIA) conçu pour évaluer et atténuer les risques associés au déploiement de systèmes décisionnels automatisés. L'AIA classe les décisions prises au moyen de l'intelligence artificielle en fonction de leur incidence probable sur les droits, le bien-être ou les intérêts économiques des personnes ou des collectivités, ou de la durabilité continue d'un écosystème. Celles qui sont jugées comme ayant une incidence plus élevée sont assujetties à des exigences plus rigoureuses en ce qui a trait aux explications.

ÉTAPES SUIVANTES

L'industrie reconnaît l'importance de l'intelligence artificielle explicable et prend des mesures pour assurer son déploiement à plus grande échelle :

- Les chercheurs d'IBM ont proposé le concept de fiches de présentation pour les services d'IA qui rendraient compte de leur équité, de leur robustesse, de leur explicabilité et de leur traçabilité[2]. Les chercheurs estiment que la normalisation et la médiatisation de ces renseignements sont essentielles pour bâtir la confiance à l'égard des services d'IA dans l'ensemble de l'industrie.
- La Defense Advanced Research Projects Agency (Agence pour les projets de recherche avancée de défense) des États-Unis mène des recherches sur les « systèmes d'IA de troisième vague », où les machines comprennent le contexte et l'environnement dans lesquels elles opèrent, et, au fil du temps, élaborent des modèles explicatifs sous-jacents[3].

---

[1] Section 4 : Objectifs et résultats escomptés, https://tbs-sct.gc.ca/pol/doc-fra.aspx?id=32592

[2] La traçabilité signifie l'information détaillée concernant le développement, le déploiement et la maintenance de l'intelligence artificielle.

[3] Consultez la page suivante pour obtenir de plus amples renseignements: https://www.darpa.mil/program/explainable-artificial-intelligence.

Les fournisseurs ajoutent une capacité d'explication aux plateformes d'IA de différentes façons :

- plateformes de science des données qui génèrent automatiquement des explications du modèle en langage naturel;
- visualisation des tendances pour l'analyse de modèles;
- outils permettant d'obtenir des renseignements sur le rendement du réseau neuronal;
- système automatisé et transparent d'analyse de modèles de tendances pour la prise de décisions financières;
- outils expliquant les constatations du modèle et avertissant les utilisateurs de la possibilité de partialité dans les données.

Les efforts déployés par les fournisseurs de solutions pour aborder les divers enjeux, comme les explications, la partialité et la conformité, combinés aux règlements judicieux, constituent un pas vers la création des fondements de la confiance et pavent la voie vers l'adoption efficace et généralisée de l'intelligence artificielle.

Sujets connexes : Intelligence artificielle, analytique augmentée, apprentissage machine