Comment améliorer l'explicabilité et la responsabilité des algorithmes?

Winston Maxwell

COMMENT AMÉLIORER L'EXPLICABILITÉ ET LA RESPONSABILITÉ DES ALGORITHMES ?

L'évolution des méthodes d'apprentissage basées sur de l'Intelligence Artificielle (IA) ont favorisé le développement des algorithmes d'aide à la décision tous secteurs confondus, mais ces outils informatiques restent qualifiés de « boîtes noires ». Un groupe de chercheurs pluridisciplinaire s'est penché sur cette problématique d'opacité pour y remédier et favoriser ainsi la confiance du public à l'égard des algorithmes.

« *Tout le monde souhaiterait que les algorithmes soient explicables, surtout dans les domaines les plus critiques comme la santé ou le transport aérien. Il y a un vrai consensus dans ce domaine* », souligne Winston Maxwell. Pour preuve, la Commission européenne a récemment publié un livre blanc sur sa stratégie liée à l'IA, qui promeut notamment l'explicabilité, la transparence et la responsabilisation des décisions algorithmiques. Au niveau français, le rapport Villani sur l'IA, publié en 2018, a également insisté sur la nécessaire explicabilité des algorithmes. Toutefois, cette doctrine se heurte à de nombreuses problématiques connexes au concept d'explicabilité comme l'éthique, la définition du bon niveau d'explication à fournir, les caractéristiques techniques des méthodes d'IA utilisées, la préservation du secret des affaires, les coûts supplémentaires ou encore le manque de précision juridique sur ce concept. De fait, si des législations en vigueur l'imposent, notamment pour les algorithmes utilisés par les administrations en France, elles laissent généralement de larges marges de manœuvre, rajoutant ainsi des difficultés supplémentaires pour les développeurs, utilisateurs et régulateurs de ces outils. « *Notre travail de recherche a été réalisé dans une logique pluridisciplinaire regroupant les sciences de données, les mathématiques appliquées, l'informatique, l'économie, les statistiques, le droit et la sociologie, afin d'avoir des réflexions de fond sur la définition, les techniques et les besoins d'explicabilité, qui s'intègrent dans les notions plus larges de transparence et de responsabilité* », relate David Bounie, co-auteur du rapport.

En résumé, l'explicabilité consiste, par exemple, à aider les utilisateurs à saisir le classement d'un moteur de recherche, à accompagner les enquêteurs dans la compréhension du crash d'un véhicule autonome ou encore à détecter d'éventuelles discriminations dans l'attribution d'un prêt.

L'EXPLICABILITÉ N'EST PAS UNIFORME ET DÉPEND DE PLUSIEURS FACTEURS

Pour parvenir à leurs objectifs consistant à démystifier l'explicabilité, les chercheurs ont développé une méthodologie originale, dont le point de départ est contextuel. Ils ont ainsi défini quatre facteurs importants :

– Le destinataire de l'explicabilité, c'est-à-dire le public visé par l'explication. Son niveau sera différent selon qu'il soit utilisateur ou régulateur par exemple.

– Le niveau d'importance et d'impact de l'algorithme. L'explicabilité d'un accident d'une voiture autonome n'a pas le même degré d'importance que celle d'un algorithme de publicités ou de recommandations de vidéos.

– Le cadre légal et réglementaire, qui est différent selon les zones géographiques, comme en Europe avec le règlement général sur la protection des données (RGPD).

– L'environnement opérationnel de l'explicabilité, comme son caractère obligatoire pour certaines applications critiques, le besoin de certification avant le déploiement ou la facilitation d'utilisation par les usagers.

« *La prise en compte des quatre facteurs contextuels de l'explicabilité est primordiale, car, pour les industriels, l'explicabilité est avant tout motivée par des exigences opérationnelles*

D'après *Flexible and Context-Specific AI Explainability : A Multidisciplinary Approach*, écrit par Valérie Beaudouin, Isabelle Bloch, David Bounie, Stephan Clemençon, Florence d'Alché-Buc, James Eagan, Winston Maxwell, Pavlo Mozharovskyi et Jayneel Parekh, ainsi qu'un entretien avec Winston Maxwell.

Winston Maxwell est directeur d'études en droit et numérique au département sciences économiques de Telecom Paris. Auparavant, il a été avocat associé du cabinet Hogan Lovells, spécialisé dans le droit des données. Il est diplômé de Cornell Law School et a obtenu un doctorat en sciences économiques à Télécom Paris (***Smarter Internet Regulation Through Cost-Benefit Analysis,*** publié aux Presses des Mines en 2017). Ses travaux de recherche portent principalement sur la régulation de l'intelligence artificielle.

Méthodologie

Les chercheurs ont réalisé un article, qualifié de *« position paper »* (article de positionnement), sur les problématiques d'explicabilité et de responsabilité des algorithmes. Ils ont ainsi échangé avec les différentes parties prenantes (politiques, académiques, industrielles) dans une logique pluridisciplinaire (mathématiques, informatique, sciences sociales…), afin de synthétiser l'état de l'art dans ce domaine et d'établir des pistes de recommandations scientifiques pour améliorer l'explicabilité des algorithmes.

qui se distinguent largement des aspects légaux », précise Winston Maxwell.

L'EXPLICABILITÉ AU REGARD DES COÛTS ET AVANTAGES POUR LA SOCIÉTÉ

Après la première étape liée aux différents contextes de l'explicabilité, les chercheurs ont étudié les solutions techniques des algorithmes d'IA. Sans rentrer dans des détails, ils ont effectué un inventaire des différentes méthodes utilisées pour comprendre leurs spécificités. Parmi elles figurent notamment les approches d'IA hybrides qui combinent le meilleur de plusieurs techniques et dont les développements futurs sont prometteurs pour améliorer l'explicabilité : *« Ces approches pourraient réduire le fossé entre performance et explicabilité des algorithmes. À terme, l'explicabilité sera une partie intégrante de la performance »,* anticipe Isabelle Bloch, co-auteure de l'étude.

En attendant, les chercheurs ont apporté une autre innovation majeure en intégrant des comparaisons coûts-bénéfices à la notion d'explicabilité. Autrement dit, cela consiste à chiffrer les avantages et les inconvénients de l'explicabilité pour la société. À ce titre, l'exemple comparatif précédemment cité entre l'explication d'un accident d'une voiture autonome et celle d'un moteur de recherche est pertinent, car ces deux catégories d'algorithmes n'ont pas le même impact sur la société. Les chercheurs ont ainsi identifié plusieurs catégories de coûts liés à l'explicabilité, en particulier celles relatives au stockage des données dans des registres dédiés, qui s'avèrent très importantes. Néanmoins, au-delà de ces coûts, la réglementation RGPD limite la possibilité de stocker des données personnelles. *« Cette question des données devra faire l'objet d'un choix politique, car le RGPD rentre en contradiction frontale avec l'exigence d'explicabilité, en particulier sur les données biométriques et de reconnaissance faciale. Les réflexions sur ce sujet ne sont qu'à leurs prémices et les régulateurs devront certainement décider en fonction des applications et de leurs impacts sur la société »,* analyse Winston Maxwell.

L'EXPLICABILITÉ LOCALE ET GLOBALE

Outre les facteurs contextuels et les analyses coûts-bénéfices, les chercheurs ont également observé qu'un bon niveau d'explicabilité doit tenir compte des dimensions globale et locale. Le premier cas implique la description de l'algorithme et la manière de l'utiliser ou pas. *« C'est comme une notice d'emploi et de mise en garde dans laquelle figurent notamment le type de données utilisées et les situations où il doit être employé. D'ailleurs, la Commission européenne a adopté ce type d'approche dans son livre blanc dédié à sa stratégie d'IA »,* précise Winston Maxwell. Quant à l'explicabilité locale, elle consiste à expliquer les décisions algorithmiques particulières, comme le refus d'un prêt bancaire. *« Ces deux dimensions de l'explicabilité sont nécessaires même si elles visent des choses totalement différentes et qu'elles dépendent des quatre facteurs contextuels de départ »,* confirme Winston Maxwell. Nul doute que, dans les prochains mois, ce sujet de l'explicabilité de l'IA, en particulier des algorithmes, reviendra sur le devant de la scène parallèlement aux réflexions actuelles et futures de Bruxelles. ●

À retenir

- Le bon niveau d'explicabilité d'un algorithme dépend de 4 facteurs contextuels : le destinataire de l'explication, le niveau d'importance de l'application algorithmique, l'environnement légal et réglementaire, ainsi que le cadre opérationnel. En outre, le niveau d'explicabilité global (fonctionnement général de l'algorithme) ou local (décision particulière) est également à prendre en compte.
- L'explicabilité d'un algorithme doit être effectuée au regard des coûts et avantages induits pour la société. Le stockage des données devra notamment faire l'objet d'un choix politique, car il est coûteux et n'est pas compatible avec tous les types de données en raison de la réglementation RGPD.
- L'explicabilité s'oppose généralement à la performance des algorithmes, car ils sont souvent construits dans une logique opérationnelle. Toutefois, avec le développement de techniques d'IA hybride, l'explicabilité fera partie intégrante du bon fonctionnement des algorithmes.