Intelligence Artificielle : comment anticiper les nouvelles réglementations européennes

L'Union européenne s'est saisie de la question de l'explicabilité des algorithmes. Le Dr Sara Meftah alerte les entreprises sur la nécessité d'anticiper les changements de réglementation à venir.

Depuis quelques années, l'Intelligence Artificielle (IA) est bien comprise comme un enjeu majeur par l'Union européenne. En Europe, plusieurs documents ont matérialisé les réflexions en cours visant à encourager la mise en œuvre des conditions favorables au développement d'un terrain européen d'excellence en matière d'IA. Dans le même temps, les textes reflètent les réflexions autour des enjeux de régulation de son utilisation afin de chercher à atténuer les risques dérivés possibles.

Ainsi, dans le cadre de l'annonce officielle de la commission européenne sur le projet de réglementation en matière d'intelligence artificielle (paru le 21 avril 2021), les applications d'IA sont classées en quatre catégories selon le niveau de risque : risque minimal, risque limité, risque élevé et risque inacceptable. En pratique, une grande partie des applications de l'IA sont catégorisées selon le troisième niveau de risque (risque élevé). Parmi celles-ci, nous retrouvons les applications de diagnostic médical, les services publics et privés essentiels (comme l'attribution des crédits), l'accès à l'emploi, l'éducation, le tri des CVs, le classement des candidats et l'administration de la justice. Ces applications

ont vocation à être soigneusement évaluées avant d'être mises sur le marché et tout au long de leur cycle de vie. La Commission européenne recommande tout particulièrement l'obligation de la transparence de ces applications pour faciliter l'évaluation de leur conformité. Pour être appliquées, ces suggestions devront désormais être adoptées par le Parlement européen.

Rendre les boîtes noires plus transparentes

Le fait est que la majorité des entreprises s'orientent de plus en plus vers des modèles d'IA complexes et opaques comme le Deep Learning, en raison des perspectives très prometteuses qu'ils apportent en termes de performances prédictives. Ces modèles sont effectivement des « boîtes noires » puisque les représentations internes et les décisions produites par ces modèles sont difficiles à interpréter. En outre, les raisons qui expliquent pourquoi un modèle a donné une certaine décision plutôt qu'une autre sont inconnues.

À lire aussi

Comment reprendre le contrôle sur les algorithmes ?

Dès lors, si l'obligation de transparence de ces modèles s'impose dans les prochaines années, la plupart de ces entreprises ne parviendront pas à répondre aux exigences des régulateurs et ne pourront donc pas mettre leurs produits sur le marché. Ils devront alors travailler précipitamment sur l'explicabilité de leurs modèles, ce qui peut s'avérer complexe, coûteux, voire contre-productif. Anticiper dès à présent les effets de ces réglementations pourrait donc être une priorité pour les entreprises où l'explicabilité des modèles doit progresser concurremment au développement des modèles d'IA de plus en plus performants (et opaques). Plus précisément, ces entreprises pourraient entreprendre dès maintenant des actions concrètes pour s'assurer que leurs systèmes n'enfreignent pas les réglementations futures.

Une étape après l'autre

De prime abord, les experts métiers de chaque entreprise pourraient recenser les risques potentiels que les modèles d'IA utilisés ou développés par l'entreprise peuvent engendrer, comme par exemple la liste des

discriminations liées à des biais éthiques. Les experts au sein des entreprises pourraient aussi contribuer au développement des méthodes d'explicabilité, qui permettent en premier lieu d'investiguer si les risques recensés par les experts métiers sont encodés dans les représentations internes des modèles et en deuxième étape d'expliquer les risques présents.

À titre d'exemple, une application de recrutement développée (apprise à partir de millions de CV) et utilisée par Amazon en 2014 s'est avérée biaisée vis-à-vis des femmes. Ainsi, pour les entreprises qui développent ce type d'applications de recrutement basées sur l'apprentissage automatique à partir d'exemples, le risque de discriminations liées aux données biaisées est élevé. Il faudra alors, avant la mise sur le marché, mettre en œuvre des algorithmes d'explicabilité qui permettent de vérifier si ce type de biais est présent dans le modèle. Et par la suite, en utilisant une autre famille d'algorithmes d'explicabilité, justifier la source de ce biais (par exemple en détectant les exemples d'apprentissage qui ont causé ce biais) et le corriger.

À lire aussi

« Les algorithmes ne sont pas plus sexistes que la société actuelle »

Enfin, une étape essentielle d'ailleurs souvent délaissée dans le monde industriel repose dans l'évaluation des explications produites par les méthodes d'explicabilité. Ainsi, l'établissement de métriques automatiques sophistiquées qui combinent la robustesse, la plausibilité et la fidélité de ces modèles est primordial. Pour cela, un débat public incluant les secteurs privés et public mériterait d'être engagé pour définir les attendus de l'explicabilité et ainsi pouvoir définir comment on peut objectiver l'évaluation de l'explicabilité de ces modèles et satisfaire les exigences réglementaires.

Le Dr Sara Meftah est chercheuse au Square Research Center.

Article écrit par SARA MEFTAH