Consultation publique du NIST sur les 4 grands principes pour une IA explicable (17 août au 15 octobre)

Quatre principes fondamentaux pour une IA explicable

Dans ce rapport de trente pages, les chercheurs du NIST présentent quatre grands principes qui sont des propriétés jugées fondamentales que doivent avoir les systèmes d'IA pour être qualifiés d'explicables.

1. **Explication** *(Explanation):* le système produit des indications ou raison(s) pour chaque résultat généré en sortie.
2. **Compréhension** *(Meaningful)* : le système fournit des explications compréhensibles aux utilisateurs.
3. **Précision de l'explication** *(Explanation Accuracy)* : les explications fournies reflètent correctement la logique selon laquelle le système a généré ses résultats.
4. **Limites de connaissance du système** *(Knowledge Limits)* : le système fonctionne uniquement dans les conditions pour lesquelles il a été conçu ou fonctionne uniquement s'il a une confiance suffisante dans les résultats générés en sortie.

Différents types d’explication et d’algorithmes explicables

Après avoir établi ces quatre principes, le document revient sur cinq catégories d’explication que peut générer un algorithme en fonction du but recherché.

1. **Explication au profit de l’utilisateur** : le système fournit une explication qui permet à l’utilisateur final de comprendre comment le résultat a été généré.
2. **Explication pour générer de la confiance et de l’acceptation par la société :** similaire au type d’explication précédent, ces explications se distinguent par une conception particulière destinée à créer de la confiance et à rendre l’acceptation sociale plus aisée notamment dans les cas où le système arrive à des résultats inattendus.
3. **Explication pour la régulation et la conformité :** ces explications sont conçues pour, entre autres, assister dans la conduite d’audits, pour s’assurer de la conformité à des réglementations ou à des standards de sécurité. Il peut s’agir notamment d’explications détaillées à destination d’un développeur ou certificateur de systèmes véhiculés autonomes.
4. **Explication pour le développement des systèmes :** on parle ici d’explications qui permettent aux programmeurs de faciliter le développement du système, de son amélioration, son débogage ou sa maintenance.
5. **Explication au profit de l’opérateur du système :** certaines explications peuvent également être utiles pas uniquement pour l’utilisateur final mais également pour la personne qui fournit le service. On peut imaginer le cas d’un algorithme de recommandations de films permettant à un utilisateur de recevoir des suggestions de nouveaux films à visionner. Une explication possible faite par le système serait de dire « nous vous recommandons ces nouveaux films car vous avez aimé ces autres films ». Si l’utilisateur final accepte l’explication fournie, celle-ci bénéficie au fournisseur de service et opérateur du système si l’utilisateur continue à regarder de nouveaux films sur la plateforme.

Outre le public à qui est destiné l'explication et l'objectif recherché, les chercheurs du NIST soulèvent également deux autres dimensions qui peuvent permettre de catégoriser les explications générées par les algorithmes : le niveau de détails de l'explication et le temps de réponse (c'est-à-dire la contrainte de temps au bout de laquelle le système est censé fournir une explication).

Par la suite, plusieurs concepts sont utilisés pour décrire différents types d'algorithmes qui permettent d'expliquer ce que fait un système d'IA. On peut notamment parler (1) des algorithmes comme les arbres de décision ou les régressions linéaires pour lesquels le modèle lui-même représente l'explication (*self-explainable models*), (2) des algorithmes d'IA dit d'explication globale (*Global Explainable AI Algorithms*), (3) des algorithmes d'IA d'explication par décision (*Per-Decision Explainable AI Algorithms*).

Et nous humains, respectons-nous ces principes *d'explicabilité* ?

Dans la dernière partie du document, les auteurs mettent à l'essai ces quatre principes *d'explicabilité* pour voir s'ils pourraient être applicables à des personnes physiques.

Pour le premier principe *Explication*, ils s'interrogent notamment pour savoir si pour chacune des décisions que nous prenons, nous produisons une explication et si ce processus est bénéfique pour le preneur de décision. Pour le second *Compréhension*, les auteurs ont ensuite cherché à caractériser la capacité d'une personne à interpréter la logique d'autrui ou autrement dit, interpréter comment une autre personne est arrivée à une conclusion.

A travers divers exemples, les chercheurs tracent différents parallèles entre la façon que l'homme et la machine ont de prendre des décisions et cherchent à montrer qu'en réalité nous ne respectons, nous-mêmes, que partiellement les principes édictés.

Selon les auteurs, ce genre de réflexions peut être très utile et ouvrir la voie à de nouvelles interfaces entre homme et machine qui tirent le meilleur parti des

compétences de chacun.

Rédacteur : Kévin KOK HEANG, Attaché-adjoint pour la Science et la Technologie