NIST : Nouvelle consultation publique sur les biais en IA jusqu'au 5 août 2021

Le 22 juin, le NIST a publié un nouveau document qui propose une **méthodologie pour l'identification et la gestion des biais dans les algorithmes d'intelligence artificielle** (IA).

Ce travail spécifique sur les biais s'inscrit dans un cadre plus large au NIST de développement d'une IA de confiance et responsable *(trustworthy and responsible* AI), qui passe notamment par la création de standards établis sur la base d'un consensus entre les différentes parties prenantes et d'une approche basée sur les risques *(consensus sandards and risk-based framework*). Parmi les caractéristiques techniques identifiées qui permettent de *cultiver la confiance* envers les systèmes d'IA, le NIST liste les suivantes :

- Précision (*accuracy*)
- Explicabilité et interprétabilité (*explainability and interpretability*)
- Protection de la vie privée (*privacy*)
- Fiabilité (*reliability*)
- Robustess (*robustness*)
- Sûreté (*safety*)
- Sécurité au sens résilience (*security*)

- Gestion des biais nuisibles (*harmful biases are mitigated*)

Comme pour l'**établissement de ses principes pour une IA explicable** (consultation publique qui était ouverte du 17 août au 15 octobre 2020), le NIST encourage tout organisme à commenter ou à faire des suggestions de modification à ce document initial et ce avant la date butoir du 5 août 2021 **[date butoir repoussée jusqu'au 10 septembre 2021], en écrivant à ai-bias@list.nist.gov.**

Malgré une conscience plus grande des biais en IA, leur gestion reste difficile

Le document à commenter débute par une remise en contexte de ce que l'on entend lorsque l'on parle de biais en lien avec les systèmes d'IA et indique que les travaux à venir se concentrent sur ces biais dans les algorithmes qui peuvent avoir des conséquences nuisibles pour la société. On entend par là notamment les biais liés aux problèmes de discrimination de certaines populations, d'impact disproportionné d'une décision prise à l'aide de l'IA ou simplement de résultat injuste obtenu par l'IA.

Avec la prolifération d'approches utilisant l'apprentissage statistique ou qui font appel à des larges bases de données, l'IA a permis de mettre en lumière de nombreux biais qui existent, pour des raisons historiques, culturelles ou autres, dans différents systèmes qui structurent nos sociétés. Couplée à l'automatisation des prises de décision, cette situation a produit une prise de conscience du grand public sur les risques potentiels pour la société induite par une dépendance excessive à l'outil IA.

Au-delà de la détection de ces biais, qui représente un défi en soi, il est également important de réussir à quantifier le degré d'erreur qu'introduit le biais par rapport à l'objectif initial. Même dans ce cas, une autre difficulté qui se présente est la réponse à apporter une fois qu'un biais est identifié et quantifié.

Malheureusement, l'état de l'art actuel ne permet pas d'offrir des principes qui soient à la fois suffisamment généraux pour couvrir l'ensemble des impacts négatifs que peuvent engendrer les biais tout en pouvant être adaptés à différents contextes.

Pour le moment, les biais sont surtout classifiés en fonction de leur type (statistique, cognitif ou autre) ou bien par cas d'usage ou secteur industriel spécifique.

Une approche en trois étapes qui suit le cycle de développement de l'IA

En réponse à cette problématique, le NIST propose d'adopter une approche en trois temps qui suit des étapes clés du développement d'un système d'IA (cf. **cycles de vie en IA du *General Services Administration*** et **recommandations du Conseil sur l'IA de l'OCDE**) .

Cette approche est fondée sur un travail bibliographique préliminaire dans lequel près de 313 articles, extraits de livres, rapports et autres publications en rapport avec les biais en IA ont été passés en revue pour voir les différentes définitions existantes de biais et les catégorisations proposées dans la littérature. Des discussions entre experts et leaders d'opinion se sont ensuite tenus lors d'ateliers ouverts à tous et organisés par le NIST (cf. **workshop en ligne du 18 août 2020**).

Pour chacune des étapes (a) pré-conception, (b) conception et développement puis (c) déploiement, le document fournit des éléments clés à prendre en considération et des examples qui expriment comment :

- des biais statistiques peuvent se présenter,
- ces biais statistiques peuvent refléter et interagir avec des biais cognitifs et des biais de la société qui présents dans les données, dans les modèles et dans les pratiques associés à l'usage de l'IA.

À titre d'exemple, pendant l'étape de pré-conception où l'on cherche à définir le problème à résoudre et proposer une ébauche de la solution à concevoir, la façon de formuler le problème et savoir qui est en charge de cette tâche sont des éléments qui peuvent trahir certains biais.

Il est possible que le développement de l'algorithme soit porté par des personnes convaincues de l'apport positif de l'IA au problème identifié, ou bien également que

ces personnes surestiment les capacités et performances possibles de l'IA pour la résolution du problème donné.

En conclusion de ce document préliminaire, le NIST estime que l'approche proposée serait la plus efficace pour une gestion efficace des biais dans les systèmes d'IA.

Le NIST prévoit divers nouveaux ateliers et événements en 2021-2022 pour continuer les travaux sur les différentes briques constituantes d'une IA de confiance.

Rédacteur :

Kévin KOK HEANG, Attaché adjoint pour la Science et la Technologie, deputy-ntics@ambascience-usa.org

Autres liens :

https://www.nist.gov/news-events/news/2021/06/nist-proposes-approach-reducing-risk-bias-artificial-intelligence