IA Explicable

Synthèse Atelier

Bérard Thibault p2005244
Colmant Axel p2408624
Guillaume Jean-Luc p2114105
Savigny Henri p1935618

Présentation

L'intelligence artificielle (IA) se développe rapidement et s'intègre dans de nombreux domaines, mais l'utilisation de modèles complexes, comme les réseaux neuronaux profonds, pose des problèmes de transparence. Qualifiés de "boîtes noires", ces modèles offrent peu de visibilité sur leur raisonnement, rendant difficile l'interprétation de leurs décisions.
Ce manque de transparence est d'autant plus problématique lorsqu'il s'agit d'applications critiques en santé, finance ou justice. La transparence de l'IA, toutefois, n'est pas absolue : elle varie selon le contexte et les besoins d'interprétation.
L'IA Explicable (XAI) vise ainsi à rendre l'IA plus accessible, permettant une meilleure compréhension et acceptation des décisions automatisées.
Durant cette présentation, nous nous sommes les questions suivantes :

- Comment rendre l'IA compréhensible pour les humains ?
- Comment garantir la confiance dans des modèles d'IA opaques ?
- Comment éviter les biais et assurer des décisions éthiques et équitables ?

Les techniques de transparence se répartissent en deux catégories principales : les techniques d'explications post-hoc et les méthodes de transparence inhérentes au modèle.

Les explications post-hoc interviennent après la prédiction et comprennent l'analyse des données, du modèle et des résultats. Par exemple, analyser les données d'entrée permet de déterminer les caractéristiques influentes, tandis que l'étude des prédictions met en lumière des tendances ou biais.

À l'inverse, certaines méthodes, comme les arbres de décision, intègrent directement la transparence dans leur structure hiérarchique. Chaque nœud représente une caractéristique, et chaque branche une condition, permettant une visualisation simple et intuitive du processus décisionnel.
En outre, d'autres méthodes, comme les modèles linéaires et la régression logistique, établissent une relation linéaire entre les variables d'entrée et la sortie, facilitant l'interprétation.
Le modèle de régression linéaire est particulièrement prisé pour sa simplicité : chaque coefficient associé aux variables d'entrée révèle leur influence directe sur la prédiction. La clarté apportée par ces modèles linéaires les rend incontournables, notamment dans les contextes où la compréhension de l'impact individuel de chaque variable est essentielle.

Les approches post-hoc comme LIME et SHAP aident aussi à interpréter les modèles complexes.
LIME génère des instances proches de l'exemple d'origine et ajuste un modèle simple pour expliquer la prédiction initiale, tandis que SHAP calcule la "valeur de Shapley" de chaque caractéristique, évaluant son importance moyenne dans différentes combinaisons.

Les explications contrefactuelles sont un autre exemple, fournissant des scénarios hypothétiques pour comprendre les variations de décision : si un individu modifie certaines caractéristiques (comme le revenu ou l'âge), cela pourrait changer l'issue d'une prédiction, par exemple, un prêt approuvé ou refusé.

Les applications de la XAI sont variées.
En santé, elle aide les médecins à interpréter les diagnostics, comme lorsqu'une IA détecte une tumeur et explique les caractéristiques suspectes.
Dans la finance, elle assure l'équité des décisions de crédit, évitant des biais sur des critères comme l'origine ethnique.
En éducation, elle personnalise l'apprentissage, et dans la justice, elle favorise des décisions transparentes, identifiant les biais et assurant une impartialité.

En conclusion, la XAI est indispensable pour favoriser une IA digne de confiance. En rendant les décisions des algorithmes compréhensibles, elle répond aux besoins de transparence et d'équité. Elle permet une collaboration équilibrée entre l'homme et la machine, assurant une IA au service de l'humain dans les domaines clés de notre société.

Débats

L'IA Explicable (XAI) suscite des débats quant à sa capacité réelle à résoudre des problèmes de transparence et de compréhension des décisions.
D'un côté, certains estiment que la XAI peut effectivement clarifier les décisions en cas de problème, à condition que le modèle soit bien conçu et ses failles correctement identifiées. Si les défauts sont connus mais laissés inchangés, la responsabilité retombe alors sur le créateur du modèle, renforçant l'idée que la XAI peut aussi rendre compte des erreurs de conception.
En revanche, d'autres soutiennent qu'il est souvent difficile de corriger des failles connues, et que même avec une XAI, la responsabilité repose sur la transparence des concepteurs quant à ces limites.
Par exemple, dans le cas d'un modèle d'IA appliqué à un scénario comme le naufrage du Titanic, la XAI pourrait révéler quelles caractéristiques ont influencé les décisions, mais cela n'explique pas entièrement pourquoi ces décisions ont été prises — l'interprétation finale reste entre les mains des humains. En outre, les informations fournies par la XAI sont souvent complexes ou fragmentaires, risquant de provoquer une surcharge cognitive si toutes les règles doivent être assimilées, ce qui limite la clarté qu'elle est censée apporter.

Bien que la XAI ait pour objectif de fournir des explications objectives et sans biais, elle ne peut que partiellement pallier les biais provenant des données d'origine.
De plus, une confiance aveugle dans l'IA, favorisée par la promesse de transparence de la XAI, pourrait conduire à des dérives, notamment si des entreprises utilisent cette "certification" de transparence pour renforcer leur crédibilité auprès de clients non avertis.

Ainsi, malgré ses apports, la XAI ne doit pas être vue comme la solution absolue aux biais et limites de l'IA, mais plutôt comme un outil qui, utilisé de manière responsable, peut contribuer à une meilleure compréhension de ces systèmes.