2 offres de stage
sur le projet CODEX-AIM

L'XAI ou *Explainable Artificial Intelligence* est un vaste champ de recherche en plein développement. L'une des questions centrales, parmi les nombreux sujets de recherche ouverts de ce domaine, est de déterminer quelle est la meilleure façon d'expliquer à l'utilisateur le résultat produit par la machine.

Le micro-projet CODEX-AIM vise à étudier l'effet de différents types d'explications sur l'utilisateur final de systèmes d'aide à la décision basés sur des algorithmes d'IA. Il se basera sur un état de l'art pour déterminer les types d'explications les plus fréquemment employés. Il s'appuiera sur des travaux antérieurs du GIS Cormorant dans le domaine de l'XAI et la confiance (METRAU et CATARI), pour développer une plateforme d'expérimentation afin de comparer l'effet du type d'explication sur la relation entre l'utilisateur et la machine.

Pour répondre aux problèmes posés par l'utilisation croissante des modèles IA dans les applications à forts enjeux, l'intelligence artificielle explicable (XAI) a connu un essor important durant les dernières années. Initialement dévolue essentiellement à la recherche de solutions techniques permettant de produire automatiquement des explications, la discipline s'est heurtée à plusieurs difficultés, en particulier lorsque ces solutions ont été confrontées aux utilisateurs finaux, en raison notamment de leur manque d'expertise en IA. L'XAI s'est alors attachée à s'inspirer des sciences sociales pour produire des explications plus faciles à comprendre. Produire la « meilleure » ou du moins une « bonne explication » est une tâche complexe, dépendant à la fois de l'utilisateur, de la tâche et de l'environnement. Parmi ces multiples facteurs, CODEX-AIM se propose d'étudier l'effet de différents types d'explications, qui est actuellement mal compris.

Selon leur type, les explications peuvent être agnostiques du modèle, ou au contraire dépendre de la technologie d'IA utilisée par celui-ci. Les explications agnostiques sont en général plus simples à appréhender par l'utilisateur car elles ne font pas appel à des connaissances techniques. En contrepartie, elles peuvent être plus complexes à générer. Mais leur plus grand avantage est qu'elles sont par nature applicables à n'importe quel modèle IA. Les explications non-agnostiques sont parfois plus simples à générer, mais elles demandent de plus grande compétences techniques pour être comprises. Par exemple, pour un algorithme de forêts aléatoires, on pourra montrer l'arbre de décision ayant le plus contribué à la décision de l'IA.

Il est intéressant de comparer l'effet de différentes sortes d'explication, avec des explications agnostiques telles que celles basées sur des exemples ou des contre-exemples, et des explications plus techniques et dépendant de la technologie du modèle IA.

Le projet pourra se baser en partie sur des travaux réalisés en 2023 dans le cadre du GIS Cormorant : micro-projet METRAU et étude CATARI. Ces travaux ont conduit à la mise en place de deux environnements pour l'expérimentation dans le domaine de la confiance et de l'XAI :

- Une plateforme d'instrumentation et d'expérimentation (CUPCAX)
- Des modules et fonctions complémentaires pour l'application GoldenAI développée par Thales, utilisant des modèles IA pour analyser et classifier des interceptions en Guerre électronique.

L'un ou l'autre de ces deux environnements pourront être utilisés pour mettre en place l'expérimentation.

Le projet sera encadré par Gilles Coppin à IMT Atlantique, et Mathias Bollaert à Thales DMS. Les ressources nécessaires à la réalisation du projet sont de l'ordre de 2 x 6 mois (stage), une partie des ressources étant dédiée aux développements informatiques du projet, une autre à la préparation et au suivi des expériences.

Le projet se décompose de la façon suivante :

- Compléments sur l'état de l'art à propos des différents types d'explications ;
- Définition des scénarios d'expérimentation ;
- Développement d'un prototype enrichi à partir des travaux réalisés dans les projets METRAU et CATARI menés en 2023 ;
- Expérimentation et comparaison de l'effet des différents types d'explications.

La définition plus précise des 2 sujets de stage pourra dépendre des profils des candidatures reçues.