Réseau de recherche Good in Tech (partenariat Institut Mines-Télécom Sciences Po) – Programme Confiance IA Ingénieur en Deep Learning

Ingénieur de recherche / Post-doctorat

Contexte : le projet autopsie de l'IA

Face aux nombreux incidents techniques et sociaux posés par l'IA, ces dernières années ont connu la naissance de communautés académiques en IA, telles que *Explainable AI* (XAI), l'équité en ML (*FairML*) ou encore *Privacy-preserving ML*.

Cependant les méthodes issues de ces communautés rencontrent des limites. Bien qu'il existe aujourd'hui des méthodes pour interpréter le *Machine Learning* tabulaire et catégoriel (Lime, Shap, etc.), il n'existe aujourd'hui aucune méthode générique pour produire des explications stables et contextuelles du Deep learning. De même, concernant l'équité et les biais des algorithmes, bien qu'il existe des métriques pour mesurer et corriger les biais du ML (disparate impact, equal opportunity), il est difficile d'appliquer ces méthodes sur du Deep learning sans variables catégorielles stables (sensitive variable). Aujourd'hui la littérature académique s'accorde à dire que la production d'explication du Machine Learning est contextuelle, c'est à dire qu'elle dépend des raisons de l'explication, de la personne destinataire de l'explication, de la situation de l'explication, du type d'incident qui rend nécessaire l'explication, etc. Il est alors crucial en XAI de ne pas proposer des outils d'explications top-down indépendants du contexte mais, au contraire, des méthodes d'investigation *bottom-up* qui partent d'un cas contextuel en situation pour le mettre en récit.

Or aujourd'hui, les incidents générés par l'IA sont largement diffusés dans la presse ou la littérature scientifique et il existe des bases de données recensant les incidents posés par l'IA. Ces incidents offrent des situations idéales, non pas pour tester des méthodes d'explications préexistantes, mais pour *co-construire* des méthodes d'investigation contextuelle et bottom-up. Dans le projet *autopsie de l'IA*, nous nous intéressons aux incidents d'ordre éthique et social de l'IA. Beaucoup de ces incidents ont la particularité d'être générés par des IA semi-ouvertes, c'est à dire qui mobilisent des bases d'apprentissages publiques, utilisent des méthodes d'apprentissage faisant objet de papier de recherche en accès libre, utilisent des réseaux pré-entrainés en accès libre, etc. Cette relative ouverture de l'IA offre la possibilité de rétro-engineering certains algorithmes afin de simuler les incidents éthiques et les mettre en récit.

Méthodologie : XAI en Deep Learning

Il s'agit donc dans ce projet pluridisciplinaire de créer un environnement de mise en récit d'incidents éthiques liés au Deep Learning non catégoriel (image, texte) par la simulation d'incidents. Dit autrement, nous cherchons à développer les outils et méthode d'investigation pour un laboratoire public d'autopsie des incidents éthiques posés par l'IA. La conception de ce laboratoire sera également un dispositif qui vise à explorer la compréhension et l'explicabilité des algorithmes par le (grand) public par la mise à disposition d'outils accessibles de diagnostic d'un incident d'IA. Pour cela, nous proposons la méthode suivante :

1. Identifier dans des répertoires publiques (Github ou paper with code) les bases de données d'apprentissage, les réseaux pré-entraînés utilisés, etc.
2. Reproduire le classifieur de Deep Learning
3. Simuler l'incident éthique
4. Développer des outils d'autopsie de l'incident pour permettre l'investigation et la production d'explication. Ex : Méthode de visualisation des réseaux de neurone par simulation d'images[1], utilisation de classifieurs annexes
5. Scénariser l'investigation, proposer des mises en récit de l'incident
6. Organiser des ateliers de confrontation avec les utilisateurs
7. Répertorier l'investigation dans une plateforme d'autopsie de l'IA

Proposer des méthodes et outils d'autopsie de l'IA permettra de générer une base de connaissance très utile dans les phases de tests avant la mise en production de l'IA. Les méthodes d'investigation proposés dans ce projet peuvent en effet être utilisées pour simuler des scénarios de tests avant la mise en production, diagnostiquer les incidents et générer des explications robustes pendant la phase de production. Il s'agit à la fois de proposer une méthode d'investigation mais aussi des outils accessibles pour réaliser une enquête face à un incident de l'IA.

Nous pensons en effet qu'une IA de confiance suppose d'élargir le champ de l'explication des experts vers le grand public par la création d'outils de visualisation d'incidents accessibles tout en prenant en compte la complexité du problème. Un tel laboratoire d'autopsie des incidents est aujourd'hui un dispositif manquant dans l'écosystème IA. Pourtant, un organe de contrôle ex-post, transparent, accessible et disposant de moyens techniques précis, est fondamental pour assurer une confiance durable dans l'IA.

Voici ci-dessous quelques exemples d'incidents publics simulables par la plateforme d'autopsie :

1. L'algorithme de rognage de Twitter
    - Une controverse dans la presse : lien presse, réponse Twitter
    - Une méthode publique : lien
    - Une base de données publique : lien arXiv
2. Background matting sur Zoom
    - Une controverse dans la presse : lien presse
    - Une méthode publique : github
    - Une base de données publique : paper with code
3. Recrutement avec reconnaissance faciale :
    - Une controverse dans la presse : lien presse
    - Une méthode publique : paper with code
    - Une base de données publique : paper with code

[1] https://distill.pub/2017/feature-visualization/

Présentation de Good In Tech

Le réseau de recherche Good in Tech (www.goodintech.org) est un partenariat entre l'Institut Mines-Télécom et Sciences Po, sous l'égide de la Fondation du Risque, en partenariat avec CGI et Artefact. Ce réseau regroupe des chercheurs des deux institutions, des doctorants et post-doctorants pour mener des recherches pluridisciplinaires sur deux enjeux :

1. Données, algorithmes et société
2. Responsabilité Numérique des Entreprises

L'Institut Mines-Télécom (www.imt.fr) est le premier groupe d'écoles d'ingénieurs et de management en France, avec plus de 1120 chercheurs, 8 grandes écoles et plus de 13 000 étudiants.

Le Medialab de Sciences Po est un laboratoire de recherche interdisciplinaire réunissant sociologues, ingénieurs et designers, le Medialab mène des recherches thématiques et méthodologiques exploitant et interrogeant la place prise par le numérique dans nos sociétés.

Le projet Autopsie IA est financé par le programme Confiance IA (https://www.confiance.ai/), le plus gros programme de recherche technologique du plan #AIforHumanity qui doit faire de la France un des pays leader de l'intelligence artificielle (IA).

Profil souhaité

- ✓ Vous avez un **master** dans lequel vous avez étudié le Machine Learning avancé (Deep Learning en computer vision et NLP). Idéalement vous avez un **doctorat** où vous avez appliqué les techniques de Machine Learning avancé.
- ✓ Vous avez des facilités en informatique, notamment en **Python**. Des connaissances en **javascript** sont un plus.
- ✓ Vous aimez la **transdisciplinarité** : vous serez amené à travailler avec des designers, des développeurs *front-end*, des sociologues, ingénieurs, etc.
- ✓ Vous avez une appétence pour les **projets de recherche** : lecture de la littérature académique, mise en place d'un protocole de recherche, analyse de données, rédaction d'articles de recherche, communication académique, etc.
- ✓ Vous savez gérer un projet en autonomie. Vous avez déjà utilisé des outils de **gestion de projet** comme *Notion,* outils de *Kanban* et autres outils *DevOps.*

Modalités du poste

- Rémunération attractive
- Poste disponible au plus vite

Contact :

Envoyer CV + Lettre de motivation à :

christine.balague@imt-bs.eu