Thèse en informatique : « **Intelligence artificielle explicable et non biaisée : vers une compréhension et représentation des phénomènes de sécurité urbaine »**

IMT Atlantique : Campus Brest, département LUSSI

Collaboration: département *SSG* (IMT Atlantique, Nantes)

Laboratoire de recherche: LAB-STICC (UMR CNRS 6285), équipe DECIDE

École doctorale : SPIN (Sciences pour l'ingénieur et le numérique)

Financement : projet ANR IAAP

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Mots clés : intelligence artificielle explicable ; machine learning ; représentation des connaissances ; détection de biais, safe city ; smart city (ville intelligente) ; sécurité ;

Environnement académique

IMT Atlantique, reconnue internationalement pour la qualité de sa recherche, est une grande école d'ingénieur généraliste dépendant du ministère en charge de l'industrie et du numérique. Sur 3 campus, Brest, Nantes et Rennes, IMT Atlantique a pour ambition de conjuguer le numérique et l'énergie pour transformer la société et l'industrie, par la formation, la recherche et l'innovation. Avec 290 chercheurs et enseignants-chercheurs permanents, 1000 publications et 18 M€ de contrats, elle encadre chaque année 2300 étudiants. Ses formations s'appuient sur une recherche de pointe, au sein de 6 unités mixtes de recherche dont elle est tutelle : Lab-STICC, GEPEA, IRISA, LATIM, LS2N et SUBATECH.

Le poste sera basé au département LUSSI dans le campus de Brest d'IMT Atlantique. Le département compte environ 65 personnels dont 21 enseignants chercheurs permanents. Le département propose une approche pluridisciplinaire des transitions numérique et technologique des systèmes sociotechniques. Il est partie prenante du laboratoire de recherche Lab-STICC (UMR CNRS 6285). Le ou la futur-e doctorant-e fera partie de l'équipe DECIDE du Lab-STICC qui fournit des solutions d'aide à la décision aux décideurs confrontés à des données hétérogènes et complexes. L'équipe travaille sur 3 axes de recherche : données, décision et information. Pour cela elle développe des travaux en fouille de données, apprentissage machine, théorie des graphes, optimisation, aide à la décision et fusion de données.

Description du projet

Dans de plus en plus de secteurs, les techniques d'intelligence artificielle sont proposées afin de comprendre, prédire, représenter de nombreux phénomènes et connaître les relations de cause à effet (la causalité) entre les phénomènes étudiés, notamment pour assister les experts d'un domaine dans leur prise de décision.

Cependant, l'opacité et la complexité de certains modèles dits « boites noires » ont été récemment largement dénoncées (Miller, 2018). Par conséquent, le domaine de l'Intelligence Artificielle eXplicable (XAI) a pris de l'ampleur face aux modèles d'IA de plus en plus complexes et opaques, et surtout suite au règlement européen général de protection de données (RGPD) qui donne droit aux individus d'exiger une explication des processus de traitement automatique de données les concernant (Goodman et Flaxman, 2017). Ces nouvelles méthodes de XAI permettent aux experts de mieux comprendre la décision proposée par le modèle d'IA, d'être guidés dans le choix des actions à

effectuer, et de développer un certain degré de réflexivité vis-vis du modèle et d'apporter une interaction optimale avec celui-ci (Fahed et al., 2018 ; Chraibi Kaadoud et al., 2021 ; Chraibi Kaadoud, et al., 2022 ; Saeed et al., 2023).

Le sujet de thèse s'inscrit dans le domaine de l'intelligence artificielle explicable, de l'apprentissage automatique et de la science des données. Le travail de recherche sera mené au sein du projet ANR intitulé « Les effets de l'Intelligence Artificielle sur l'Activité Policière : nouveaux régimes de quantification, diversification du marché et redéfinition des dispositifs de sécurité urbaine (ANR IAAP, ANR-21-CE26-0023-01) ». Le ou la futur-e doctorant-e fera partie de l'équipe projet composée de chercheurs en informatique et en sociologie.

De manière globale, ce projet étudie plusieurs cas de « safe city » et de « smart city » (Paris, Marseille, Montpellier, Montréal et Toronto) utilisant notamment des technologies de vidéosurveillance automatisée et de prédiction de phénomènes criminels et délictuels (Castagnino, 2019), notamment par de l'apprentissage automatique. L'objectif général du projet est d'analyser les effets concrets de l'IA dans le travail policier. Deux grands questionnements articulent la recherche :

- Comment la mise en place des systèmes d'IA change les modes de connaissance et de représentation des phénomènes de délinquance ? Comment comprendre et représenter ces phénomènes dans un système d'IA ? Comment l'intégration des connaissances des experts améliore la transparence et l'explicabilité du système IA ?
- Comment détecter et intégrer les préoccupations politiques et sociales dans les productions scientifiques et techniques, i.e. un système d'IA pour la sécurité urbaine, liées notamment aux risques de biais et de discriminations ?

Objectifs de la thèse et contributions attendues

Données hétérogènes disponibles : Le ou la futur-e doctorant-e aura à disposition une première source de données : les analyses sociologiques issues des enquêtes de terrain réalisées par les chercheurs en sociologie qui font partie du projet ANR IAAP. Ces analyses sociologiques représentent une source de données « non numériques » qui peuvent servir à corriger, informer et compléter les sources de données numériques, c-à-d. les données issues de capteurs et autres données publiques (comme celles disponibles sur data.gouv.fr) sur les phénomènes de délinquance. En permettant de les contextualiser et d'en faire ressortir certains biais, l'apport attendu est d'améliorer la compréhension de ces phénomènes à partir de l'analyse, de la compréhension et de l'extraction de connaissances pertinentes des donnés à la fois qualitatives et quantitatives. Cependant, la nature et hétérogénéité de ces deux sources de données rendent la tâche de fusion très complexe. Afin de représenter de telles données hétérogènes, des approches à base de graphes de connaissances temporels seront à étudier (Xu et al., 2020) vue leur apport dans des systèmes explicables (Goebel et al., 2018).

Approches à proposer : l'objectif de la thèse est de proposer un système d'IA à base d'apprentissage automatique non-biaisé et explicable. Cela sera réalisé en deux étapes :

- Détection de biais : les sources potentielles de biais et les préjugés potentiels doivent être identifiés en confrontant différentes sources(expertises professionnelles, enquêtes empiriques, etc.). Les biais algorithmiques doivent également être détectés : biais des données, biais statistiques, biais de traitement, biais des experts (Mehrabi et al., 2021). Le défi réside dans la définition d'une équité adaptée au cadre des activités policières et l'intégration des résultats d'analyses sociologiques dans une mesure de détection de biais.
- Système IA explicable : un système à base d'apprentissage automatique, principalement non-supervisé sera proposé afin d'extraire les connaissances et de les représenter de manière transparente. Nous proposons de représenter une explication sous formes multiples : un ensemble de statistiques, visualisations, règles, et termes sémantiques. Des techniques à base de graphes de connaissances temporels seront étudiées (JI et al., 2021 ; Tiddi et al., 2022). Le

défit ici réside (i) dans la définition d'un équilibre entre transparence (i.e. explicabilité) et performance afin de s'assurer que le modèle d'apprentissage fonctionne conformément aux attentes et ne propage pas de biais, et (ii) dans la validation quantitatives et qualitatives des formes d'explication .

Un état de l'art sur ces sujets sera à réaliser.

Le ou la doctorant-e pleinement intégré-e au projet ANR IAAP sera amené-e à :

- Contribuer à la réflexion collective de l'équipe projet via une participation active aux réunions et séminaires du projet.
- Contribuer à la rédaction d'articles scientifiques.
- Participer aux actions de valorisation et de diffusion des résultats obtenus (séminaires, conférences, …).

Profil du candidat

- La ou le candidat(e) doit avoir un diplôme de Master et/ou Ingénieur dans des domaines liés à l'informatique, science des données, mathématiques appliquées, statistique ou traitement de signal.
- Avoir une aptitude au développement de méthodes d'intelligence artificielle, machine learning, statistique, analyse des données. Des connaissances en traitement automatique des langues sera appréciée. Une appétence pour le dialogue interdisciplinaire et la sociologie sera valorisée.
- Être familier avec certains outils informatiques/langages : python (scikit-learn, Pandas, NumPy), …
- Avoir un bon niveau d'anglais écrit et oral. Avoir la capacité de communiquer en français (niveau A2 minimum).

Modalités de candidature

Le dossier de candidature doit comprendre, en un seul PDF,

- CV
- lettre de motivation
- relevés de notes de L3, M1, M2 (ou années équivalentes)
- copie de diplômes ou attestation de réussite
- noms/coordonnées de 1-2 personnes référentes à contacter ou éventuellement des lettres de recommandation.

L'ensemble du dossier (un seul PDF) doit être adressé par mail le plus tôt possible aux chercheurs suivants avec pour sujet du mail [candidature-doctorat-IAAP] (les entretiens se feront en visio-conférences) :

- Lina Fahed (co-encadrant, informatique) lina.fahed@imt-atlantique.fr
- Florent Castagnino (co-encadrant, sociologie) florent.castagnino@imt-atlantique.fr
- Philippe Lenca (directeur, informatique) philippe.lenca@imt-atlantique.fr

Pour toute question, merci de contacter Lina Fahed : lina.fahed@imt-atlantique.fr

Informations complémentaires

- Date limite de candidature : jusqu'à ce que le poste soit pourvu
- Début du contrat : Automne 2023
- Nature et durée du contrat : CDD de 3 ans (36 mois)
- Salaire : Le salaire prévu est d'environ 1700€ net par mois (financement via le projet ANR IAAP). Des vacations d'enseignement seront éventuellement possibles en supplément.
- Le ou la futur-e doctorant-e sera inscrit-e à l'école doctorale SPIN
- Localisation Géographique : IMT Atlantique, campus de Brest (possibilité de télétravail partiel) avec bureau au sein du département LUSSI
- Les postes offerts au recrutement sont ouverts à toutes et tous avec, sur demande, des aménagements pour les candidats en situation de handicap
- Autres avantages : prise en charge des transports en commun, forfait mobilité durable (pour le covoiturage ou les trajets à vélo), possibilité de logement dans la résidence étudiants, accès aux activités sportives au campus.

Références

CASTAGNINO, Florent, « Rendre « intelligentes » les caméras : déplacement du travail des opérateurs de vidéosurveillance et redéfinition du soupçon », Working papers de la Chaire Villes et Numériques, Sciences Po, , p. 1-32, 2019.

CHRAIBI KAADOUD, Ikram, FAHED, Lina, TIAN, Tian, HARALAMBOUS, Yannis, LENCA, Philippe. « Automata-based Explainable Representation for a Complex System of Multivariate Times Series ». In Proceedings of the 14th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2022) KDIR, 2022.

CHRAIBI KAADOUD, Ikram, FAHED, Lina, LENCA, Philippe. . « Explainable AI: a narrative review at the crossroad of Knowledge Discovery, Knowledge Representation and Representation Learning. », MRC@IJCAI 2021: Twelfth International Workshop Modelling and Reasoning in Context. Vol. 2995, 2021.

FAHED, Lina, BRUN, Armelle, et BOYER, Anne, « DEER: Distant and Essential Episode Rules for early prediction », Expert Systems with Applications (ESWA), 93, p. 283-298, 2018.

GOEBEL, Randy, CHANDER, Ajay, HOLZINGER, Katharina, et al., « Explainable AI: The New 42? », Machine Learning and Knowledge Extraction, p. 295-303, 2018.

GOODMAN, Bryce et FLAXMAN, Seth, « European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation" », AI Magazine, 38, 3, p. 50-57, 2017.

JI, Shaoxiong, PAN, Shirui, CAMBRIA, Erik, et al.. « A survey on knowledge graphs: Representation, acquisition, and applications », IEEE transactions on neural networks and learning systems, vol. 33, no 2, p. 494-514, 2021.

MEHRABI, Ninareh, MORSTATTER, Fred, SAXENA, Nripsuta, et al. « A survey on bias and fairness in machine learning. », ACM Computing Surveys (CSUR) 54.6 : 1-35, 2021.

MILLER, Tim, 2018, « Explanation in Artificial Intelligence: Insights from the Social Sciences », arXiv:1706.07269 [cs], 2018.

SAEED, Waddah et OMLIN, Christian. « Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities. », Knowledge-Based Systems : 110273, 2023.

TIDDI, Ilaria et SCHLOBACH, Stefan. « Knowledge graphs as tools for explainable machine learning: A survey », Artificial Intelligence, vol. 302, p. 103627, 2022.

XU, Chenjin, NAYYERI, Mojtaba, ALKHOURY, Fouad, et al. , « Temporal Knowledge Graph Completion Based on Time Series Gaussian Embedding », The Semantic Web – ISWC 2020, p. 654-671, 2022.