Titre de la thèse : Cadre de raisonnement pour une explicabilité adaptative, intelligible et centrée sur l'humain des systèmes d'intelligence artificielle

Laboratoire d'accueil : Connaissance et Intelligence Artificielle Distribuées (CIAD)

http://www.ciad-lab.fr

Spécialité du doctorat préparé : Informatique **/** Intelligence Artificielle

Mots-clefs : Explicabilité de l'IA (XAI), Explicabilité adaptative, IA centrée sur l'humain, IA symbolique, Raisonnement symbolique, ontologie

Encadrement et Contact Scientifique :

Vincent Hilaire vincent.hilaire@utbm.fr

Amel Hidouri amel.hidouri@ube.fr

Ouassila Labbani Narsis ouassila.narsis@ube.fr

Descriptif détaillé de la thèse

Contexte

L'explicabilité des systèmes d'intelligence artificielle (IA) est devenue un enjeu scientifique, éthique et sociétal majeur, en particulier dans des domaines centrés sur l'humain (*human-centric*) et sensibles tels que la santé, la finance ou l'éducation, où les prédictions algorithmiques influencent directement des décisions humaines à fort impact. Dans ces contextes, l'enjeu ne réside plus uniquement dans la compréhension interne des modèles, mais dans la capacité des systèmes d'IA à communiquer leurs raisonnements de manière intelligible, pertinente et contextualisée [1].

Cette évolution s'inscrit dans un changement de paradigme, passant d'une *approche model-centric* à une vision *user-centric*, centrée sur les besoins et le contexte des utilisateurs. L'explication ne doit plus être considérée comme une donnée brute, mais comme un processus de communication adaptatif entre le système d'IA et l'humain. Elle joue ainsi un rôle déterminant dans la justification, la confiance et l'appropriation des décisions.

Les travaux existants en *Explainable Artificial Intelligence* (XAI) se sont largement concentrés sur des méthodes d'explication post-hoc appliquées à des modèles de type *boîte noire* [2]. Bien que largement diffusées, ces approches présentent des limites importantes : manque de fidélité au raisonnement réel du modèle, incohérences entre méthodes d'explication, absence de mesure explicite de la confiance, faible actionnabilité pour des utilisateurs non experts, ainsi que des coûts computationnels parfois incompatibles avec des usages en temps réel. Dans des contextes à fort enjeu, ces limites peuvent induire une illusion de compréhension et de confiance, rendant l'utilisation de telles explications potentiellement problématique, voire dangereuse [3].

Ces constats mettent en évidence un problème fondamental : expliquer *a posteriori* des modèles opaques ne garantit pas une compréhension fiable et cohérente. Il devient dès lors nécessaire d'adopter des approches intégrant l'interprétabilité directement dans la conception des modèles d'IA (*XAI by design*), afin de concilier performance prédictive, transparence et fidélité explicative [4].

Par ailleurs, les explications doivent être adaptées aux profils et aux contextes d'usage et s'inscrire dans une approche centrée sur l'utilisateur, capable de produire des explications fidèles, naturelles, intelligibles, interactives et flexibles (*explanation facilities*) [5,6].

Objectifs

L'objectif de cette thèse est de concevoir un cadre formel d'interprétabilité adaptative, capable de produire des explications différenciées en fonction du profil de l'utilisateur et du contexte d'usage. Dans une approche centrée sur l'utilisateur, il s'agira de formaliser explicitement les profils des utilisateurs (niveau d'expertise, rôle, objectifs, contraintes, préférences), ainsi que leurs contextes d'usage, au sein d'un modèle symbolique et formel, sous la forme d'une ontologie. Cette ontologie permettra d'orienter le processus d'explication, d'assurer l'interopérabilité entre différents modèles d'IA, et d'élargir la notion classique d'explicabilité en intégrant des concepts complémentaires centrés sur l'utilisateur.

En structurant les différents types d'explications possibles, ainsi que les relations entre profils, contextes et formes d'explication (abductives, contrastives, etc.), le modèle symbolique constituera un support de raisonnement permettant de définir des règles logiques capables d'orienter dynamiquement le processus d'explication du raisonnement de l'IA. Ce cadre sera intégré au sein d'un système explicatif couplé à différents types de modèles d'intelligence artificielle (réseaux de neurones, forêts aléatoires, etc.). Le système sera également capable d'apprendre à partir des retours des utilisateurs, permettant une mise à jour de l'ontologie en temps réel. Enfin, l'impact de ces explications adaptatives sera évalué au travers des études de cas menées dans différents domaines d'application et auprès de profils utilisateurs variés, afin d'analyser leurs effets sur la compréhension, la confiance et l'appropriation des décisions produites par les systèmes d'IA.

Cette thèse vise à contribuer à une vision renouvelée de l'XAI dans laquelle l'explicabilité n'est plus un ajout a posteriori, mais une composante centrale de la conception de systèmes d'IA fiables, responsables et réellement centrés sur l'humain.

Références bibliographiques

1. Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning*.
2. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why should I trust you?" Explaining the predictions of any classifier*.
3. Rudin, C. (2019). *Stop explaining black box machine learning models for high stakes decisions*. Nature Machine Intelligence.
4. Caruana, R. et al. (2015). *Intelligible models for healthcare*. KDD.
5. Miller, T. (2019). *Explanation in artificial intelligence: Insights from the social sciences*. Artificial Intelligence.
6. Abdul, A. et al. (2018). *Trends and trajectories for explainable, accountable and intelligible systems*.

Profil demandé :

- Un diplôme de master recherche ou équivalent dont la spécialité principale est l'informatique
- Une bonne maîtrise de l'anglais (oral et écrit) est exigée
- Des connaissances en intelligence artificielle, en explicabilité, en ingénierie des connaissances et en ontologie seront appréciées
- Un bon esprit de travail en équipe

Les dossiers de candidature (CV, lettre de motivation, relevés de notes, lettre de recommandation) doivent être envoyées par email.

PhD title: A Reasoning Framework for Adaptive, Intelligible, and Human-Centered Explainability of Artificial Intelligence Systems

Laboratory: Connaissance et Intelligence Artificielle Distribuées (CIAD)
http://www.ciad-lab.fr

Specialty of PhD: Computer Science / Artificial Intelligence

Keywords: Explainable AI (XAI), Adaptive Explainability, Human-Centered AI, Symbolic AI, Symbolic Reasoning, Ontology

Supervision and Scientific Contact:

Vincent Hilaire vincent.hilaire@utbm.fr

Amel Hidouri amel.hidouri@ube.fr
Ouassila Labbani Narsis ouassila.narsis@u-bourgogne.fr

Description:

Context

Explainability in artificial intelligence (AI) systems has become a major scientific, ethical, and societal challenge, particularly in human-centric and sensitive domains such as healthcare, finance, and education, where algorithmic predictions directly influence high-stakes human decisions. In these contexts, the challenge no longer lies solely in understanding the internal mechanisms of models, but in the ability of AI systems to communicate their reasoning in an intelligible, relevant, and contextualized manner [1].

This evolution reflects a fundamental paradigm shift, moving from a model-centric approach to a user-centric vision focused on users' needs and contexts. Explanations should no longer be viewed as raw outputs, but rather as an adaptive communication process between the AI system and the human user. As such, explanations play a decisive role in the justification, trust, and appropriation of automated decisions.

Existing work in Explainable Artificial Intelligence (XAI) has largely focused on post-hoc explanation methods applied to black-box models [2]. Although widely adopted, these approaches exhibit significant limitations, including limited fidelity to the model's actual reasoning, inconsistencies

across explanation methods, lack of explicit confidence measures, low actionability for non-expert users, and computational costs that are sometimes incompatible with real-time applications. In high-stakes settings, these limitations may create an illusion of understanding and trust, making the use of such explanations potentially problematic, or even dangerous [3].

These observations highlight a fundamental issue: explaining opaque models *a posteriori* does not guarantee reliable or coherent understanding. It therefore becomes necessary to adopt approaches that integrate interpretability directly into the design of AI models (*XAI by design*), in order to reconcile predictive performance, transparency, and explanatory fidelity [4]. Furthermore, explanations must be adapted to user profiles and usage contexts and embedded within a user-centered approach, capable of producing explanations that are faithful, natural, intelligible, interactive, and flexible (*explanation facilities*) [5,6].

Objectives

The objective of this PhD thesis is to design a formal framework for adaptive interpretability, capable of producing differentiated explanations according to the user profile and the usage context. Within a user-centered approach, the work aims to explicitly formalize user profiles (level of expertise, role, objectives, constraints, preferences), as well as their usage contexts, within a symbolic and formal model, in the form of an ontology. This ontology will guide the explanation process, ensure interoperability across different AI models, and extend the classical notion of explainability by integrating user-centered concepts.

By structuring the different types of explanations, as well as the relationships between profiles, contexts, and explanation forms (abductive, contrastive, etc.), the symbolic model will serve as a reasoning support for defining logical rules capable of dynamically steering the explanation process of AI reasoning. This framework will be integrated into an explanatory system coupled with various types of artificial intelligence models (neural networks, random forests, etc.). The system will also be able to learn from user feedback, enabling real-time updates of the ontology. Finally, the impact of these adaptive explanations will be evaluated through case studies conducted across different application domains and involving diverse user profiles, in order to analyze their effects on understanding, trust, and appropriation of decisions produced by AI systems.

This PhD thesis aims to contribute to a renewed vision of XAI, in which explainability is no longer *a posteriori* add-on, but a core component in the design of reliable, responsible, and genuinely human-centered AI systems.

References

1. Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning*.
2. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why should I trust you?" Explaining the predictions of any classifier*.
3. Rudin, C. (2019). *Stop explaining black box machine learning models for high stakes decisions*. Nature Machine Intelligence.
4. Caruana, R. et al. (2015). *Intelligible models for healthcare*. KDD.
5. Miller, T. (2019). *Explanation in artificial intelligence: Insights from the social sciences*. Artificial Intelligence.

6. Abdul, A. et al. (2018). *Trends and trajectories for explainable, accountable and intelligible systems*.

Candidate Profile:

- A research-oriented Master's degree or equivalent, with a primary specialization in Computer Science
- Strong proficiency in English (both written and spoken) is required
- Knowledge in Artificial Intelligence, Explainable AI, Knowledge Engineering, and Ontologies will be an asset
- Strong ability to work in a team

Application files (CV, cover letter, academic transcripts, letter(s) of recommendation) must be submitted by email.