SORBONNE UNIVERSITÉ

**Proposition de projet de recherche doctoral**
**LIP6 - ISIR - SCAI**

# Explicabilité des systèmes d'intelligence artificielle hybride dans un contexte d'interactions avec des humains


**Directrice de thèse porteuse du projet :**

NOM : BLOCH
Prénom : Isabelle
Titre : Professeure
e-mail : isabelle.bloch@sorbonne-universite.fr

**Unité de Recherche :** Laboratoire d'Informatique de Sorbonne Université (LIP6), UMR7606

**École Doctorale** **ED130 - EDITE**

**Co-directeur :**

NOM : CHETOUANI
Prénom : Mohamed
Titre : Professeur
e-mail : mohamed.chetouani@sorbonne-universite.fr
**Unité de Recherche :** ISIR

**Co-directrice :**

NOM : LESOT
Prénom : Marie-Jeanne
Titre : MCF, HDR
e-mail : Marie-Jeanne.Lesot@lip6.fr
**Unité de Recherche :** LIP6

**Co-directrice :**

NOM : PELACHAUD
Prénom : Catherine
Titre : DR CNRS
e-mail : catherine.pelachaud@sorbonne-universite.fr
**Unité de Recherche :** ISIR

SORBONNE UNIVERSITÉ

**Description du projet de recherche doctoral :**

1. **Contexte, objectif scientifique**
   Cette thèse s'inscrit dans le cadre des recherches émergentes sur l'explicabilité des systèmes d'intelligence artificielle (IA), en particulier dans des situations d'interaction avec des humains. Le domaine de l'*eXplainable Artificial Intelligence* (XAI), qui vise à enrichir les systèmes d'IA pour permettre à leurs utilisateurs de mieux comprendre les traitements réalisés, est en plein essor et très actif actuellement ; il se situe à la croisée de plusieurs disciplines et peut faire intervenir, outre la communauté de l'IA au sens large, des travaux en philosophie, sciences cognitives, sociologie, sciences de l'éducation ou psychologie par exemple.
   Les systèmes d'interaction humain-machine impliquent de comprendre les comportements de l'humain, de doter la machine de capacités de prendre des décisions (par exemple pour choisir quelle action prendre face à une action d'un utilisateur) et d'y répondre. La réponse peut prendre différentes formes, aussi bien logicielles que verbales ou non verbales (par exemple lorsque l'interaction se fait par un agent conversationnel animé ou par un robot humanoïde). Ces systèmes exploitent des modèles d'IA à des étapes importantes de l'interaction telles que la détection de signaux socio-émotionnels, le dialogue, la décision et la planification d'actions. Afin que l'humain puisse mieux comprendre le choix des actions d'un système d'interaction, il est important de pouvoir expliquer les modèles de décision et d'IA.
   Dans le domaine de l'IA symbolique, la recherche de "meilleures" explications a été formalisée en particulier par le raisonnement par abduction, notion introduite par Pierce dans les années 1950. Dans le domaine de l'apprentissage automatique, les méthodes d'apprentissage symbolique fournissent des résultats naturellement explicables, comme les arbres de décision, les règles ou les motifs fréquents, les implications d'attributs en analyse formelle de concepts, etc., à condition toutefois que les espaces de représentation restent de petite taille. Les méthodes d'apprentissage par réseaux de neurones sont au contraire longtemps apparues comme des boîtes noires, et la recherche d'explications est un sujet de recherche très actuel. Par exemple en analyse d'images, plusieurs méthodes ont été proposées pour l'interprétation de réseaux de classification, beaucoup moins pour des réseaux de segmentation ou dédiés à d'autres tâches. Le but de cette thèse est de formaliser cette notion d'explication lorsque l'on manipule à la fois des données et des connaissances, comme c'est typiquement le cas dans des situations d'interaction avec des humains. Très peu de travaux portent sur ce domaine, et ils utilisent le plus souvent des formes spécifiques de connaissances, comme des graphes causaux ou des règles. Des sources d'inspiration peuvent être trouvées dans les travaux en philosophie, sciences cognitives, sociologie, psychologie (Miller, 2019). Il s'agira alors de trouver des modèles mathématiques et computationnels permettant de mettre en œuvre les mécanismes identifiés dans ces domaines, en prenant en compte les humains, qui peuvent être à la fois des observateurs mais aussi des acteurs dans les systèmes d'interaction et de décision.
2. **Approche scientifique**
   La richesse des recherches du domaine XAI vient de la multiplicité des formes que peut prendre une explication (image, schéma, règle, formule logique, texte, graphe...), selon la définition qu'on lui donne, le destinataire considéré (à la fois selon son expertise ou son mode de communication privilégié) ou l'expression qui lui est associée. Elle est illustrée par le

nombre et le large spectre des articles publiés dans le domaine, ainsi que la diversité des taxonomies proposées pour l'organiser.
A un niveau différent, l'approche envisagée dans la thèse pour aborder la question de l'explication en situation d'interaction portera d'abord sur une structuration de l'espace des explications par la proposition de mesures de comparaison d'explications : il s'agira de mesurer la similarité d'explications, par exemple pour évaluer leur complémentarité, leur redondance ou leurs contradictions, en tenant compte de leur granularité et de leur forme, et éventuellement de la complexité de ces mesures en fonction de la taille et de la structure de l'espace des explications. Ces problématiques sont à relier à celle de la robustesse des explications, ainsi qu'à celle de leur cohérence, en particulier d'un point de vue temporel. La notion de contexte de l'explication joue un rôle crucial et devra être intégrée dans les propositions. Il s'agira de plus d'aller au-delà de simples corrélations, et d'introduire des notions de causalité, en particulier pour permettre d'agir ou intervenir (par exemple si on a appris qu'un comportement donné suscite une certaine réaction, la corrélation entre les deux ne permet pas de dire quel comportement il faudrait adopter pour que cette réaction ne se produise pas). L'évolution temporelle de ces causalités sera également importante pour prendre en compte par exemple la portée d'une action dans le temps et la variation de son impact (cas d'une action pouvant engendrer une réaction négative immédiate, mais impliquer des retours positifs à plus long terme).
Savoir ce qu'est une "bonne" ou une "meilleure" explication est une autre question que nous souhaitons aborder. Cela nécessite de définir une relation d'ordre partiel sur l'espace des explications, qui pourra représenter des degrés (qualitatifs) de préférence, de robustesse, de pertinence, et pourra dépendre du contexte d'application et de la question d'explication posée. Par exemple la meilleure explication pour un individu n'est pas forcément la plus vraisemblable dans un sens statistique. Là encore les travaux en sciences sociales pourront nous éclairer.
La thèse se focalisera sur un type d'interaction humain-machine : les agents conversationnels animés (ACA). Ceux-ci ont une forme humanoïde, peuvent communiquer verbalement et non verbalement avec leurs interlocuteurs humains, et montrer toute une palette d'expressions d'émotions et d'attitudes sociales. Le comportement multi-modal des humains est détecté, décrit et interprété en termes d'état émotionnel, signaux sociaux, et aussi langage naturel. De nombreuses études ont montré l'impact du comportement des agents conversationnels sur les performances des humains à mener une tâche et sur la qualité d'interaction. On peut imaginer des applications de type jeux sérieux, par exemple pour l'entraînement de soignants pour interagir avec des patients Alzheimer, ou encore la préparation à des entretiens d'embauche. On se placera dans le cas où le comportement des agents est modélisé et réalisé par des méthodes d'IA hybride, associant apprentissage à partir de données et formalisation de connaissances. Plusieurs types d'explication peuvent alors être envisagés. Il est important de comprendre et donc d'expliquer le choix de montrer tels comportements plutôt que tels autres pour une interaction donnée, et on pourra donc chercher à expliquer le comportement de l'agent (en fonction du contexte donc sûrement le comportement de l'humain), à expliquer le comportement de l'humain par exemple dans une tâche éducative (expliquer pourquoi il est ou n'est pas adéquat pour l'humain de faire telle ou telle chose dans une situation interactive donnée). Dans ce contexte, il s'agira donc de générer les explications, en utilisant la commu-

nication verbale et non verbale, structurer ces explications, les ordonner, les évaluer, comme décrit plus haut. Les explications pourront servir ensuite à améliorer les interactions (comme dans les deux exemples cités ci-dessus), ce qui suppose d'identifier dans les comportements et les interactions ce qui a conduit à la décision et à l'explication.
A plus long terme, ce travail pourra également donner des pistes pour, inversement, permettre un apprentissage de l'agent par explication.

L'IA explicable est au cœur des enjeux actuels de la recherche en Intelligence Artificielle et figure explicitement parmi les défis que SCAI (le centre en IA de Sorbonne Université, auquel les deux équipes encadrantes participent) se propose de relever. Cette thèse abordera la question sous un angle multidisciplinaire original, à la croisée des communautés d'apprentissage automatique et d'agents intelligents, avec des applications potentielles dans des domaines variés.

3. **Rôle et compétences des encadrants**
Les encadrants, venant de deux laboratoires différents, ont des expertises complémentaires qui couvrent tous les aspects du projet.
Isabelle Bloch travaille sur des méthodes d'IA hybride, combinant logique, ensembles flous, graphes et apprentissage, en particulier pour le raisonnement spatial et l'interprétation d'images. L'IA explicable a été abordée dans ses travaux d'une part en logique, dans le cadre du raisonnement abductif, et d'autre part dans le cadre de réseaux de neurones profonds de classification et segmentation.
Mohamed Chetouani mène des activités de recherche dans les domaines du traitement du signal social, de la robotique sociale et de l'apprentissage automatique interactif avec des applications en psychiatrie, psychologie, neurosciences sociales et éducation.
Marie-Jeanne Lesot a travaillé sur les questions d'interprétabilité des algorithmes d'apprentissage automatique, principalement des classifieurs, notamment sous la forme d'explications contrefactuelles. Ses travaux exploitent également le cadre de la théorie des sous-ensembles flous et la logique floue, qui fournissent des outils et qui visent intrinsèquement à être proches des modes de réflexion des êtres humains, favorisant naturellement les questions d'interprétabilité.
Catherine Pelachaud travaille sur la modélisation des agents conversationnels animés. Ses travaux ont porté sur la modélisation du comportement non verbal, de l'expression des émotions et des attitudes sociales. Pour cela, elle a appliqué différentes techniques d'IA allant du développement de règles issues de la littérature en sciences sociales à l'apprentissage profond à partir de larges corpus.

4. **Quelques publications en lien avec le projet**

   - M. Aiguier, J. Atif, I. Bloch, and R. Pino Pérez. Explanatory relations in arbitrary logics based on satisfaction systems, cutting and retraction. *International Journal of Approximate Reasoning*, 102:1–20, 2018.
   - J. Atif, C. Hudelot, and I. Bloch. Explanatory reasoning for image understanding using formal concept analysis and description logics. *IEEE Transactions on Systems, Man and Cybernetics: Systems*, 44(5):552–570, 2014.
   - J. Broekens and M. Chetouani. Towards Transparent Robot Learning through TDRL-

based Emotional Expressions, *IEEE Transactions on Affective Computing*, (in press).

- V. Couteaux, O. Nempont, G. Pizaine, and I. Bloch. Towards interpretability of segmentation networks by analyzing deepdreams. In *iMIMIC Workshop at MICCAI 2019: Interpretability of Machine Intelligence in Medical Image Computing*, volume LNCS 11797, pages 56–63, Shenzhen, China, 2019.
- V. Couteaux, S. Si-Mohamed, O. Nempont, T. Lefevre, A. Popoff, G. Pizaine, N. Villain, I. Bloch, A. Cotten, and L. Boussel. Automatic knee meniscus tear detection and orientation classification with Mask-RCNN. *Diagnostic and Interventional Imaging*, 100:235–242, 2019.
- S. Dermouche, C. Pelachaud, Leveraging the Dynamics of Non-Verbal Behaviors For Social Attitude Modeling, *IEEE Transactions on Affective Computing*, 2020
- R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti and D. Pedreschi. A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5):1-42, 2018.
- T. Laugel, M.-J. Lesot, C. Marsala, X. Renard and M. Detyniecki. The Dangers of Post-hoc Interpretability: Unjustified Counterfactual Explanations. In *Proc. of IJCAI'19*.
- T. Miller. Explanation in artificial intelligence: insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.
- R. Niewiadomski, S. Hyniewska, C. Pelachaud, Constraint-Based Model for Synthesis of Multimodal Sequential Expressions of Emotions, *IEEE Transactions on Affective Computing*, vol. 2, no. 3, 134-146, July 2011.
- A. Pirovano, H. Heuberger, S. Berlemont, S. Ladjal, and I. Bloch. Improving interpretability for computer-aided diagnosis tools on whole slide imaging with multiple instance learning and gradient-based explanations. In *Workshop iMIMIC at MICCAI*, Lima, Peru.
- S. Wallkötter, S. Tulli, G. Castellano, A. Paiva, and M. Chetouani. Explainable embodied agents through social cues: a review. *ACM Transactions on Human-Robot Interaction*, 2021.

5. **Profil du doctorant recherché**
   Diplôme de master ou équivalent en informatique et intelligence artificielle, avec un intérêt pour les systèmes interactifs, entre humains et agents.