Concevoir et intégrer des IA explicatives de confiance : conditions sociotechniques et d'usage pour des technologies centrées sur l'humain en contextes professionnels et domestiques

Ranya Bennani, Marc-Eric Bobillier-Chaumon, Myriam Frejus

▶ **To cite this version:**

Ranya Bennani, Marc-Eric Bobillier-Chaumon, Myriam Frejus. Concevoir et intégrer des IA explicatives de confiance : conditions sociotechniques et d'usage pour des technologies centrées sur l'humain en contextes professionnels et domestiques. Rencontre Cifre "Ma recherche j'en parle" - Technologies centrées sur l'humain au service d'une industrie responsable, ANRT, Dec 2025, Paris (France), France. hal-04813734

HAL Id: hal-04813734
https://hal.science/hal-04813734v1

Submitted on 2 Dec 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Communication ANRT

Thématique : « Technologies centrées Humain au service d'une industrie responsable »

« Concevoir et intégrer des IA explicatives de confiance : conditions sociotechniques et d'usage pour des technologies centrées sur l'humain en contextes professionnels et domestiques »

AUTEURS

Ranya BENNANI[1], Marc-Eric Bobillier-Chaumon[2], Myriam Frejus[3].

[1]CNAM/CRTD /EDF R&D (France)

[2]CNAM /CRTD(France)

[3]EDF R&D (France)

Résumé

Cette communication s'inscrit dans un projet de thèse Cifre en partenariat avec EDF R&D, visant à faciliter l'intégration des outils d'intelligence artificielle (IA) dans des contextes professionnels et sociodomestiques. Plus précisément, elle examine les critères d'explicabilité et d'acceptabilité nécessaires à la conception de ces dispositifs.

Depuis les travaux fondateurs d'Alan Turing dans les années 1950, l'intelligence artificielle a connu une évolution significative. Alors que les premières approches reposaient sur des systèmes experts avec des règles explicites, l'avènement des techniques de machine learning et de deep learning a accru l'autonomie des systèmes. Toutefois, cette autonomie pose des défis majeurs en matière de transparence et de compréhension des décisions algorithmiques, suscitant des inquiétudes concernant la confiance et l'acceptabilité de ces technologies par les utilisateurs.

Dans le domaine des sciences humaines et sociales (SHS) et dans d'autres disciplines techniques (computer science), des recherches ont exploré les attentes des utilisateurs concernant l'explicabilité, la confiance et l'adoption des systèmes d'IA, en particulier pour les outils d'aide à la décision (Vuarin & Steyer, 2023 ; Goodman & Flaxman, 2016).

Au sein d'EDF R&D, une étude exploratoire (Turpin & Frejus, 2023) a été réalisée auprès du personnel de la centrale nucléaire du Blayais, chargé de la gestion des colmatants (des déchets susceptibles de bloquer les canalisations d'eau refroidissant les réacteurs). Cette recherche a permis d'identifier les besoins d'explicabilité auprès des futurs utilisateurs d'un logiciel d'aide à la décision, destiné à évaluer le risque de colmatage. Elle a également souligné l'importance d'adopter une approche anthropocentrée et contextuelle (Bobillier-Chaumon, 2023), intégrant les besoins réels et les variations professionnelles, tout en impliquant les utilisateurs dès les phases initiales de conception (co-conception).

Notre thèse prolonge cette réflexion en considérant l'explicabilité comme un processus flexible, impliquant des utilisateurs aux besoins variés et s'articulant autour d'une « trajectoire d'activité » mobilisant divers acteurs, communautés professionnelles, organisations, règles de travail, et dispositifs sociotechniques. L'objectif de cette recherche est de favoriser une meilleure compréhension des décisions et des raisonnements sous-jacents aux systèmes d'IA, afin de co-

concevoir des outils plus adaptés (à l'activité qui se fait) et appropriables (par les divers profils d'usagers). L'explicabilité inclut donc les notions de transparence, d'interprétabilité, de confiance et d'appropriation.

Les principales questions de recherche que nous aborderons dans cette communication sont les suivantes : Comment concevoir des explications adaptées à des contextes spécifiques ? Comment intégrer ces dispositifs dans des environnements collectifs où la prise de décision est réfléchie et distribuée collectivement ? Quelles conditions d'acceptabilité et de confiance doivent être prises en compte lors de la conception des dispositifs basés sur l'IA ?

Nous proposons d'explorer ces questions à travers trois dimensions de l'explicabilité : l'explicabilité instrumentale, qui se concentre sur les questions liées à l'instrumentation et à l'utilisation des artefacts ; l'explicabilité interactionnelle, qui examine les interactions entre l'humain et l'artefact dans leur cadre d'action et la dynamique collective qui en découle ; enfin, l'explicabilité systémique, qui interroge les systèmes socio-organisationnels impliqués dans la conception et le déploiement de ces outils.

Références

Bobillier - Chaumon, M -E. (2023). *Psychologie du travail digitalisé*. Dunod. https://www.dunod.com/sciences-humaines-et-sociales/psychologie-du-travail-digitalise-transformations-et-clinique-usages

Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation." AI Magazine, 38(3), 50-57. https://doi.org/10.1609/aimag.v38i3.2741

Vuarin, L., & Steyer, V. (2023). Le principe d'explicabilité de l'IA et son application dans les organisations: *Réseaux*, *N° 240*(4), 179-210. https://doi.org/10.3917/res.240.0179

Turpin, M. & Frejus, M. (2023). Comment faciliter l'appropriation de systèmes opaques : Etude sur l'eXplanable Artificial Intelligence et préconisations ergonomiques sur l'explicabilité d'un système de Machine Learning (Rapport interne, EDF). Non publié.

Mots-clés : *Explicabilité, Acceptation située, Intelligence artificielle, approche qualitative, Facteurs humains*