Quelles conditions sociotechniques et d'activité pour concevoir et intégrer des IA explicatives de confiance en situations professionnelles et socio-domestiques ?

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

L'INTELLIGENCE ARTIFICIELLE DANS LA RECHERCHE EN PSYCHOLOGIE
9 DÉCEMBRE 2024

Thématique 2 : L'apport de la recherche psychologique à l'évaluation des outils et interventions basés sur l'IA dans des contextes diversifiés

« Quelles conditions sociotechniques et d'activité pour concevoir et intégrer des IA explicatives de confiance en situations professionnelles et socio-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 le cadre d'un projet de thèse Cifre en collaboration avec EDF R&D, ayant pour objectif d'accompagner l'intégration des outils d'IA dans des environnements professionnels et socio-domestiques. Plus précisément, elle porte sur la compréhension et la détermination des critères d'explicabilité et d'acceptabilité dans la conception de ces dispositifs.

L'intelligence artificielle (IA) a considérablement évolué depuis les travaux pionniers d'Alan Turing dans les années 1950. Alors que les premières approches étaient fondées sur des systèmes experts basés sur des règles explicites, l'émergence des techniques de machine learning et de deep learning a rendu les systèmes de plus en plus autonomes. Cependant, cette autonomie accrue engendre des défis majeurs en matière de transparence et de compréhension des décisions algorithmiques, ce qui suscite des préoccupations quant à la confiance et à l'acceptabilité de ces technologies par les utilisateurs finaux.

Dans le domaine des sciences humaines et sociales (SHS) ainsi que dans d'autres disciplines plus techniques (computer science), quelques recherches ont exploré les besoins des utilisateurs vis-à-vis de l'explicabilité, de la confiance et de l'adoption des systèmes d'IA, notamment lorsqu'il s'agit de systèmes d'aide à la décision (Vuarin & Steyer, 2023 ; Goodman & Flaxman, 2016).

Aussi, au sein d'EDF R&D, une étude exploratoire (Turpin & Frejus, 2023) a été menée auprès du personnel de la centrale nucléaire du Blayais, impliqué dans la gestion du phénomène d'arrivée des colmatants (déchets qui peuvent bloquer les canalisations d'arrivée d'eau qui refroidissent les réacteurs, avec des menaces potentielles sur la fiabilité des installations). Cette étude a permis de mieux cerner les besoins en termes d'explicabilité exprimés par les futurs utilisateurs d'un logiciel d'aide à la décision (IA) permettant de diagnostiquer la gravité de ce risque de colmatage. Elle a également mis en lumière la nécessité d'adopter une approche anthropocentrée et située (Bobillier-Chaumon, 2023), tenant compte à la fois des besoins réels et des spécificités de l'activité professionnelle, tout en impliquant les utilisateurs dès les phases initiales de la conception (démarche de co-conception). Notre thèse prolonge cette réflexion en considérant l'explicabilité comme un processus flexible, impliquant des utilisateurs aux besoins diversifiés et s'articulant à un/une « cours/trajectoire d'activité » mobilisant divers

acteurs et communautés professionnelles, organisations et règles de travail, environnements et dispositifs sociotechniques d'action afin d'arriver à prendre la meilleure décision. L'objectif final de cette recherche sur l'explicabilité étant de permettre une meilleure compréhension des décisions et des raisonnements sous-jacents aux systèmes d'IA afin de co-concevoir des outils plus appropriés (à l'activité qui se fait) et appropriables (par les divers profils d'usagers). L'explicabilité englobe ainsi les notions de transparence, d'interprétabilité, de confiance, et d'appropriation.

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

Nous proposons d'examiner ces questions à travers trois dimensions de l'explicabilité : L'explicabilité instrumentale; cette dimension se concentre sur les questionnements liés à l'instrumentation et à l'usage des artefacts, l'explicabilité interactionnelle; elle traite des interactions entre l'humain et l'artefact dans leur cadre d'action, ainsi que de la dynamique collective qui en découle et 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*