L'explicabilité et l'appropriation des IA :
Cas d'une IA de prévision et d'aide à la décision en centrales nucléaires

Ranya BENNANI, ranyabennani2@gmail.com
Marc –Eric Bobillier-Chaumon, marc-eric.bobillier-chaumon@lecnam.net
Myriam Fréjus, Myriam.frejus@edf.fr

INTRODUCTION

Qu'est-ce que l'explicabilité ?

« pour un public donné, une IA explicable produit des détails ou des Raisons rendant son fonctionnement clair ou facile à comprendre » (Barredo Arrieta et al., 2020)

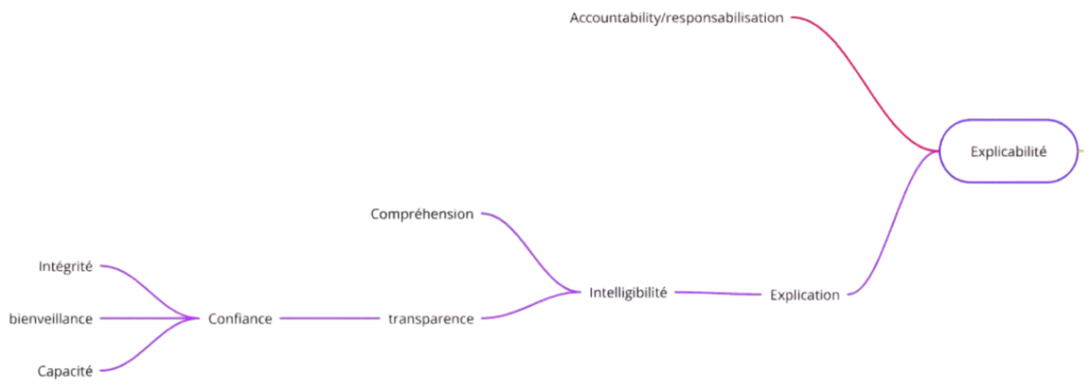
==> *U*ne des définitions de l'explicabilité vue des datascientists

==> Qui est ce public cible ?

==> Est-ce que tous les utilisateurs ont besoin de comprendre le fonctionnement de l'IA ?

==> De quelles informations auront-ils besoin pour comprendre les décisions prises par une IA ?

Où se situe le concept d'explicabilité au sein de tous les termes employés par les datascientists ?



Qu'est qu'on sait ?

- Kirsch (2017) souligne l'importance d'adapter l'explication aux besoins spécifiques des utilisateurs (variations intraindividuelles)
- Ras et al. (2018) Wang et Ying (2021) distinguent experts et non-experts
- Biran et Cotton (2017) expliquent que des dispositifs explicatifs renforcent la confiance des utilisateurs et leur appropriation du système.
- Amershi et al. (2019) constatent l'inadaptation des informations fournies par les systèmes, négligeant plusieurs aspects (contexte, variations, etc.)
- Bouzekri et Rivière (2022) pointent un manque d'outils garantissant que les systèmes autonomes soient compréhensibles et fiables pour les non-experts

ETUDE EXPLORATOIRE (Turpin et Frejus, 2023)

Contexte : Projet de conception d'un outil à base d'IA de prévision et d'aide à la décision (MACADAM)

Etude exploratoire : menée par Turpin et Frejus (2023) au sein d'une des centrales nucléaires d'EDF

Phénomène rare : Accumulation massive de colmatants

Objectif : identifier les utilisateurs et leurs besoins de conception et d'explicabilité

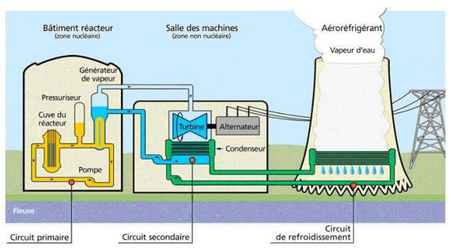


Figure 1 : fonctionnement d'une centrale nucléaire

Figure 2 : débris végétaux

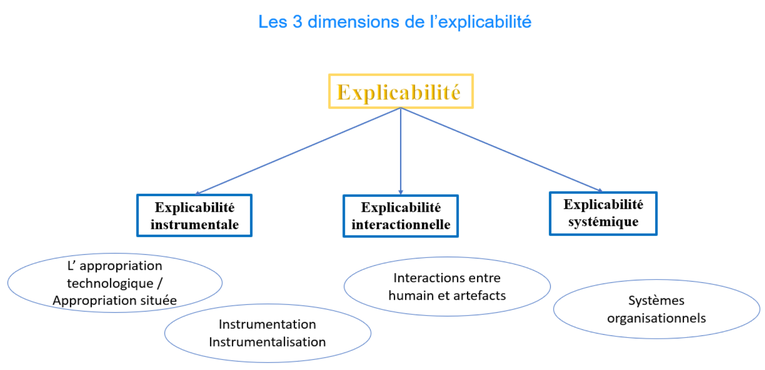
Méthodologie : Observations libres et entretiens semi-directifs

Population : Ingénieurs experts de la source froide

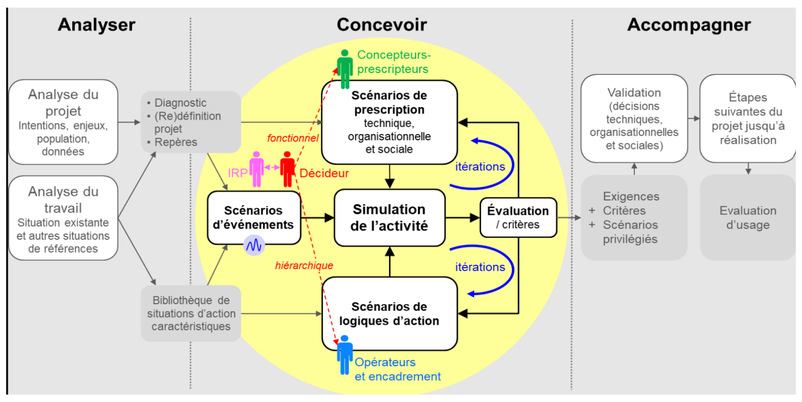
Résultats :

- Les utilisateurs principaux ont divers profils : experts et non experts, experts et novices, la direction, etc.
- Importance de la prise en compte du contexte
- Importance d'étudier cette question d'explicabilité de manière située et systémique.

APPROCHES ET METHODOLOGIE



L'explicabilité comme processus dynamique : 3 dimensions de l'explicabilité



Démarche de l'ergonomie dans la conception (Van Belleghem, 2018)

BIBLIOGRAPHIE

Amershi, S., D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen, J. Teevan, R. Kikin-Gil, et E. Horvitz (2019). Guidelines for human-ai interaction. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1–13.

Barredo Arrieta, A., N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, et F. Herrera (2020). Explainable artificial intelligence (xai) : Concepts, taxonomies, opportunities and challenges toward responsible ai. Information Fusion 58, 82–115.

Biran, O. et C. Cotton (2017). Explanation and justification in machine learning : A survey.

Bouzekri, E. et G. Rivière (2022). Choosing a questionnaire measuring connectedness to nature for human–computer interaction user studies. In IHM '22 : Proceedings of the 33rd Conference on L'Interaction Humain-Machine.

Gornet, J. et W. Maxwell (2023). Understanding ai decisions : A human-centered approach. AI Ethics 3, 45–67.

Frejus, M. (1999). Analyser l'activité d'explication pour concevoir en terme d'aide : application à la formation et à la négociation commerciale. Ph. D. thesis, Université Paris 5.

Kirsch, A. (2017). Explain to whom ? putting the user in the center of explainable ai. In Proceedings of the First International Workshop on Comprehensibility and Explanation in AI and ML. CEUR Workshop Proceedings

Van Belleghem, L. (2018). Activity simulation in design : achievements and perspectives. Activites, 15(1). https://doi.org/10.4000/activites.3129