PEPR-eNSEMBLE – Proposition de Thèse

Title of the PhD Proposal: CoDEX - **Co**llaborative **D**esign and **E**valuation of e**X**plainable AI systems

- **Name(s) of PhD Advisor(s):**
  Kathia Marçal de Oliveira, PR 1cl kathia.oliveira@uphf.fr
  Rafik Belloum, MdC, rafik.belloum@uphf.fr

- **Host Laboratory:** Univ. Polytechnique Hauts-de-France, LAMIH CNRS UMR 8201

- **Short abstract:**

With the widespread use of machine learning for algorithmic decision-making, Explainable Artificial Intelligence (XAI) systems are increasingly important. However, they are not always understandable and trustworthy for end-users, who should know when to rely on AI's advice to make informed decisions. We advocate that to better address this problem the design of the XAI system should be done in a collaborative way among the different stakeholders (AI experts, Human-computer iteraction designers, domain experts and end-users. Moreover, the decision-making should be done in collaboration during the interaction of end-user and the XAI system.

- **Short description of hosting research group / lab:**

This thesis will be developed in the Informatics Department of the LAMIH UMR CNRS 8201 and more precisely in the InterA (Interaction et Agents) team. This team is made up of 6 full professors and 7 conference leaders. The team works along two main axes:
1) The design of coordination, reasoning and knowledge models to build systems adaptable to different situations.
2) The design and evaluation of interactive systems. Human-computer interaction issues are studied in an original way through tangible, implicit and multimodal modes of interaction.

This thesis is in the context of axis 2, with regard to the interaction of end users with XAI Systems in a collaborative way for the decision-making process.

Description of the PhD proposal

Context:

Explainable Artificial Intelligence (XAI) is critical to ensuring that users develop appropriate reliance on AI systems [1]. However, studies show that even with explanations, users often either over-trust incorrect AI outputs or under-utilize correct ones. The challenge is not just about generating explanations but about ensuring that they help users make **better-informed decisions** [2]. The design of XAI systems involves two main aspects: (i) XAI design, which focuses on selecting AI techniques (e.g., LIME, SHAP) to generate explanations [3], and (ii) **explainable user interface (XUI) design**, which defines how these explanations are presented to users. The way explanations are structured in the UI directly influences user reliance, meaning that **the collaboration between AI experts, UI designers, and end-users** is crucial. Once the system is developed, a collaboration between end users and the AI system should be applied to strengthen analytical skills of the users rather than replacing human expertise. However, these collaborations context are often poorly defined, with little guidance on how these different stakeholders should work together and with the AI system to ensure that explanations truly benefit decision-making.

Problem and objectives:

The main problem this thesis aims to address is concerning the need of collaboration between stakeholders that work in the engineering of XAI systems, and between the end-users and the final XAY system in order to support them in decision making.

The goal of this thesis is to propose a human-centered design methodology for the collaborative design and evaluation of XAI systems composed of:

i. techniques and tools to facilitate an iterative co-design process;
ii. interactive mechanisms allowing users to collaborate with the AI system for the decision-making process.
iii. metrics to evaluate the collaboration established between end-users and AI systems.

A key goal is to move beyond static explanations by exploring adaptive interaction mechanisms that allow users to engage with explanations dynamically in a collaborative way with the system improving their analytical skills. We will investigate how explanation strategies such as progressive disclosure, counterfactual reasoning, and contrastive explanations can help informed decisions. This methodology will be validated in an educational setting, studying how XAI can support students and educators in decision-making tasks.

Positioning in relation to the state of the art:

Rather than treating AI models, UI design, and user needs separately, our methodology fosters continuous collaboration between all stakeholders and between end-users and the XAI system. Additionally, while many XAI solutions provide static explanations, we explore interactive and adaptive **techniques** to force the user engage cognitively with AI and optimize reliance calibration [6]. Cognitive forcing strategies to encourage users to critically assess AI recommendations have been studied, but not in the context of collaboration [1]. Regarding measures for evaluation, most of the works on literature seeks to evaluate trust [7] or the appropriate reliance [8] the user has in the XAI system not focusing on how the user interact with the system to clarify, better understand before taking a decision. In this thesis we will focus in measuring that, that means, how the user and AI system collaborate together to reach informed decision, which implies in a better user reliance.

Research questions:

- How to support the collaboration of stakeholders during the engineering of XUI of XAI systems?
- How to support the collaboration between end-users and XAI systems for the decision-making?
- How to evaluate the level of collaboration?

Theoretical foundation:

This thesis is founded in two main theorical elements:

- We build on existing research on reliance calibration in AI-supported decision-making as well as cognitive science theories that explain how users process explanations, trust AI, and integrate AI outputs into their decision-making. Cognitive forcing strategies to encourage critical engagement will also be used.
- We will leverage software engineering approaches to define structured metrics for assessing collaboration, interactivity, and user reliance. We will draw from usability and UX evaluation methods to develop new ways of measuring collaboration in XAI interactions.

Approach and methods:

Our approach combines human-centered design based on ISO 9241-210 with agile principles, ensuring iterative refinement through short design-evaluation cycles. The originality of this work lies in its dual focus on collaboration and interaction. While existing research has explored user-centered XAI, most approaches remain either too theoretical, focusing only on model transparency [4], or too domain-specific, limiting generalizability [5]. Our methodology will establish structured collaboration procedures for AI experts, UI designers, and end-users, that constitute a major challenge. It investigates how users can dynamically interact with explanations and provides concrete guidelines and tools that will be validated through real-world case studies.

This PhD project will span three years, structured around the following key milestones:

- T0 – T0+12 months: Literature review on human-AI collaboration in XAI systems, analyzing stakeholder roles and existing methodological approaches for collaborative design.
- T0+6 – T0+24 months: Definition of key user-AI interaction scenarios to guide the development of the proposed methodology.
- T0+9 – T0+24 months: Development of techniques and tools to facilitate an iterative co-design process and the implementation of interactive mechanisms enabling users to collaborate with AI in decision-making.
- T0+18 – T0+30 months: User studies to evaluate the proposed methodology, focusing on the impact of dynamic and adaptive interactions on user.
- T0+12 – T0+36 months: Scientific dissemination, including research publications and thesis writing.

Evaluation of the contributions:

We will evaluate how interactive and collaborative explanations impact users' ability to make informed decisions. This includes experimental studies with end-users in educational contexts. We will define and validate metrics for measuring collaboration in XAI systems, focusing on interaction, engagement, and decision quality. The proposed human-centered co-design methodology will be tested with AI experts, UI designers, and end-users (students) through iterative design cycles and real-world case studies.

Presentation and role of co-supervisors:

Kathia Marçal de Oliveira, Full Professor in Computer Science. She works on human computer interaction, more specifically in user-centered design and the evaluation of interactive systems through measurement. She coordinated the Interreg Project ParkinsonCom, applying user-centered approaches for design of interactive systems and evaluation based on measurement. She has also worked on defining quality measures in various domains, including ubiquitous applications, web applications, and legacy systems. She is currently supervising a thesis in XAI domain that aims to define guidelines for XUI design in order to provide a better user reliance. This thesis is in its second year and has the co-supervision (co-encadrement) of Rafik Belloum.

Rafik Belloum, Associate Professor in Computer Science. He works on Human-Computer Interaction (HCI) and software engineering, more specifically regarding explainability. He contributed to the national project XAI4AML on explainability at Télécom Paris, developing methods, tools, and frameworks for designing interactive systems aligned with user needs. His research focuses on eXplainable AI. He will contribute to defining evaluation tools and planning and conducting user studies.

References

[1] de Souza Filho, J. C., Belloum, R., de Oliveira, K. M.: Where Are We and Where Can We Go on the Road to Reliance-Aware Explainable User Interfaces? *2024 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)* (2024), 282–288.

[2] Broniatowski, D. A.: *Psychological Foundations of Explainability and Interpretability in Artificial Intelligence*. NIST Pubs (2021). https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf

[3] Lundberg, S. M., Lee, S.-I.: *A Unified Approach to Interpreting Model Predictions*. Advances in Neural Information Processing Systems 30 (2017). https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

[4] Sy, D.: *Adapting Usability Investigations for Agile User-Centered Design*. Journal of Usability Studies 2(3), 112–132 (2007). https://uxpajournal.org/wp-content/uploads/pdf/JUS_Sy_May_2007.pdf

[5] Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., Kankanhalli, M. *Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda*. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (2018). https://dl.acm.org/doi/10.1145/3173574.3174156

[6] Bertrand, A., Viard, T., Belloum, R., Eagan, J. R., Maxwell, W.: On selective, mutable and dialogic XAI: A review of what users say about different types of interactive explanations. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (2023), 1–21.

[7] Naiseh, M., Al-Thani , D., Jiang, N., Ali, R.. How the different explanation classes impact trust calibration: The case of clinical decision support systems. International Journal of Human-Computer Studies, Volume 169, (2023), 102941.

[8] Schemmer, M., Kuehl, N., Benz, C., Bartos, A. and Satzger, G. Appropriate Reliance on AI Advice: Conceptualization and the Effect of Explanations. In Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23). Association for Computing Machinery (2023), New York, NY, USA, 410–422. https://doi.org/10.1145/3581641.3584066.

Nature of digital collaboration

This thesis explores collaboration at two levels:

1. **Collaboration between stakeholders in designing XAI systems**
    - **Function**: Coordination and shared decision-making in designing explanations.
    - **Type**: Asynchronous (iterative design reviews) and synchronous (co-design workshops, discussions).
    - **Time scale**: Over months, structured in iterative cycles.
    - **Group size**: Typically small groups (AI experts, UI designers, domain experts, and end-users).
    - **Space**: Hybrid, mixing remote collaboration (e.g., virtual design sessions) and co-located workshops.
2. **Collaboration between end-users and XAI systems**
    - **Function**: Interactive decision support, explanation refinement, and trust calibration.
    - **Type**: Synchronous (real-time AI-user interaction).
    - **Time scale**: Seconds to minutes, depending on task complexity.
    - **Group size**: Individual user interacting with an AI system (or small groups, educators and students)
    - **Space**: Primarily remote or hybrid (e.g., educational settings).

By studying these two forms of collaboration, this thesis aims to improve how AI explanations are designed and how users interact with them in a way that fosters user analytical skills and informed decision-making

Contribution to digital collaboration: Expected results and Impact

This thesis will contribute to digital collaboration in several ways:

Methodological

- A systematic human-centered design methodology for the co-design of XAI systems, ensuring that explanations are collaboratively defined and iteratively refined.
- Guidelines on how different stakeholders (AI experts, UI designers, end-users) should collaborate during the development of explainable AI systems.

Empirical Contributions

- User studies examining how collaboration between users and AI systems affects reliance and decision quality**.**
- Experimental validation of different interactive explanation strategies to support collaboration in decision-making.

Positioning in the eNSEMBLE program (½ page max)

This proposal aligns with PC3 - Theme 3 by examining how AI collaboration impacts user expertise and decision. It addresses the risks of inappropriate reliance and cognitive disengagement while promoting interaction strategies that preserve analytical skills,

- **Inappropriate reliance on AI**: The thesis aims to prevent blind trust or unjustified skepticism by fostering critical engagement through interactive explanations.
- **Cognitive disengagement**: Many XAI systems provide static explanations. This work promotes interactive strategies that require users to actively refine and question AI outputs.
- **Long-term impact on user expertise**: By integrating human-centered methodologies into XAI design, the project ensures that users are not just passive recipients of AI recommendation but active participants in the decision-making process.