DESIGN STRATEGIES FOR EXPLAINABLE AI:

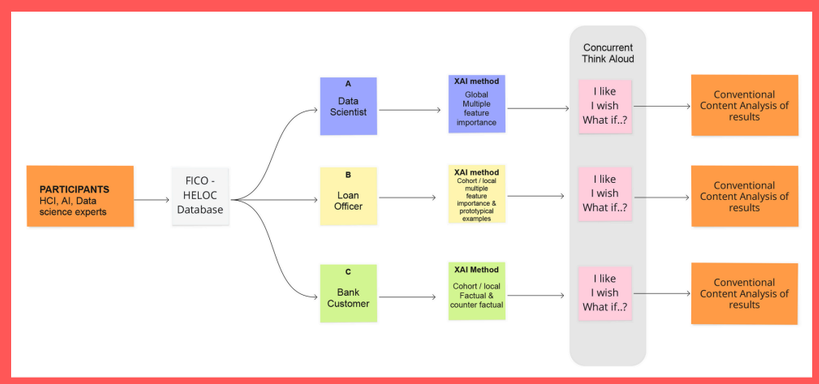
12 guiding design principles for human-centred XAI

Helen Sheridan, Dympna O'Sullivan & Emma Murphy
School of computer science, TU Dublin, Ireland

BACKGROUND

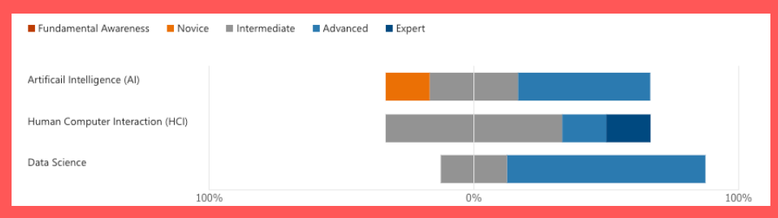
With the growing demand for transparent AI systems, the EU's AI Act emphasizes the need for accessible explanations to foster user trust and ethical AI use [1].

Our research explores the "gulf of explanation" in XAI, evaluating how well current systems align with users' mental models and engaging experts to improve human-centred XAI design [2]. This work particularly considers the application of a new and novel HCI framework, a set of 12 key design principles for human-centred XAI.



PARTICIPANTS

We recruited seven experts in AI, HCI, and data science, spanning various age groups. Participants self-assessed their expertise, with one novice and others intermediate or above.



METHODOLOGY

- Two expert evaluations with participants from HCI, AI, and data science
- Concurrent think-aloud method and "I like, I wish, What if?" discussion framework
- Evaluated XAI interfaces from IBM's Explainability 360 toolkit in FinTech context [3]
- Data collected via audio recordings, transcribed, and analyzed using conventional content analysis

01 OBVIOUS BUT UNOBTRUSIVE INTERACTIVITY

Interactive elements should be obvious but not intrusive and should enhance user understanding.

02 USE NATURAL LANGUAGE & REAL-WORLD SCENARIOS

Text / Language descriptions should use natural language and/or real-world scenarios.

03 BE CONSISTENT

Information should be consistent across all XAI representations.

04 VISUALS AND TEXT SHOULD REINFORCE EACH OTHER

Written descriptions should enhance understanding of visuals and visuals should enhance understanding of written descriptions.

05 USE NATURAL LANGUAGE FOR DATA AND FEATURES

Data used, and descriptions of features should be explained in natural language.

06 SHOW FEATURE IMPORTANCE

Features should give option to show importance related to the system decision

07 SHOW COUNTERFACTUAL EXPLANATIONS

Features should show counterfactuals, so users know what they need to improve or what they are able to improve to gain a different result.

08 VISUALS SHOULD AID UNDERSTANDING

The format of visuals should enhance explanation.

09 USE MIXED MODALITIES FOR ACCESSIBILITY

visuals and text should use mixed modalities and should be accessible.

10 CLEAR RELATIONSHIPS SHOULD BE SHOWN WITH EXAMPLE BASED XAI

Example based XAI are useful when well executed showing clearly how the examples relate to each other and the reason for the matched example.

11 DATA USAGE AND ITS CONNECTION TO FEATURES SHOULD BE SHOWN

The user should know what data was used in AI decision and how the data is linked to features.

12 EXPLAIN IF THE DATA CHANGES OVER TIME

The user should know what data was used in the AI decision and if the data changed over time.

THEMES

- Interactive elements should be clear and enhance understanding.
- Descriptions must use natural language and complement visuals.
- Charts/graphs should clarify explanations and align with text.
- Users need transparency on data used in AI decisions.
- Example-based XAI should clarify reasoning behind examples.
- Features should use natural language, highlight importance, and provide counterfactuals.

CONCLUSION

- Study evaluates and informs the design of human-centred XAI.
- Introduces a novel HCI framework with 12 key design principles for XAI.
- Highlights the importance of XAI for non-technical users, especially in high-risk domains.
- Identifies gaps between current XAI design and user understanding.
- Advances HCI practice with actionable guidelines for enhancing AI transparency.
- Acknowledges limitations and plans further validation of design principles through iterative design with non-technical users.

[1] Madiega, T., 2021. Artificial intelligence act. European Parliament: European Parliamentary Research Service.
[2] Sheridan, H., Murphy, E. and O'Sullivan, D., 2023, July. Exploring Mental Models for Explainable Artificial Intelligence: Engaging Cross-disciplinary Teams Using a Design Thinking Approach. In International Conference on Human-Computer Interaction (pp. 337-354). Cham: Springer Nature Switzerland.
[3] IBM, AI Explainability 360 – Demo. 2024 Retrieved October 1, 2024 from https://aix360.res.ibm.com/data

TU DUBLIN OLLSCOIL TEICNEOLAÍOCHTA BHAILE ÁTHA CLIATH TECHNOLOGICAL UNIVERSITY DUBLIN | COMPUTER SCIENCE