# Unlocking the Black Box: Evaluating XAI Through a Mixed Methods Approach

## Combining Quantitative Standardised Scales & Qualitative Techniques

Helen Sheridan, Dympna O'Sullivan and Emma Murphy
School of Computer Science, TU Dublin, Grangegorman, Dublin, Ireland

## AI & Society

AI is impinging on all areas of life often making life changing decisions for people. Concerns have been expressed related to biased decisions by AI systems around the processing of personal data [1]. However, legislation such as the EU's AI act has called for greater regulation of artificial intelligence in the EU including catagorising according to risk, developing systems with greater transparency and the right to an explanation of AI systems' decisions [2].

## Why are explanations for AI Important?

- Bias in data which is amplified by the model [3]
- Transparency & trust [4]
- Black-box models are inherently opaque [5]
- Transparent models are difficult to understand by lay users [6]
- The Eu's AI act legislates for explanations [2]

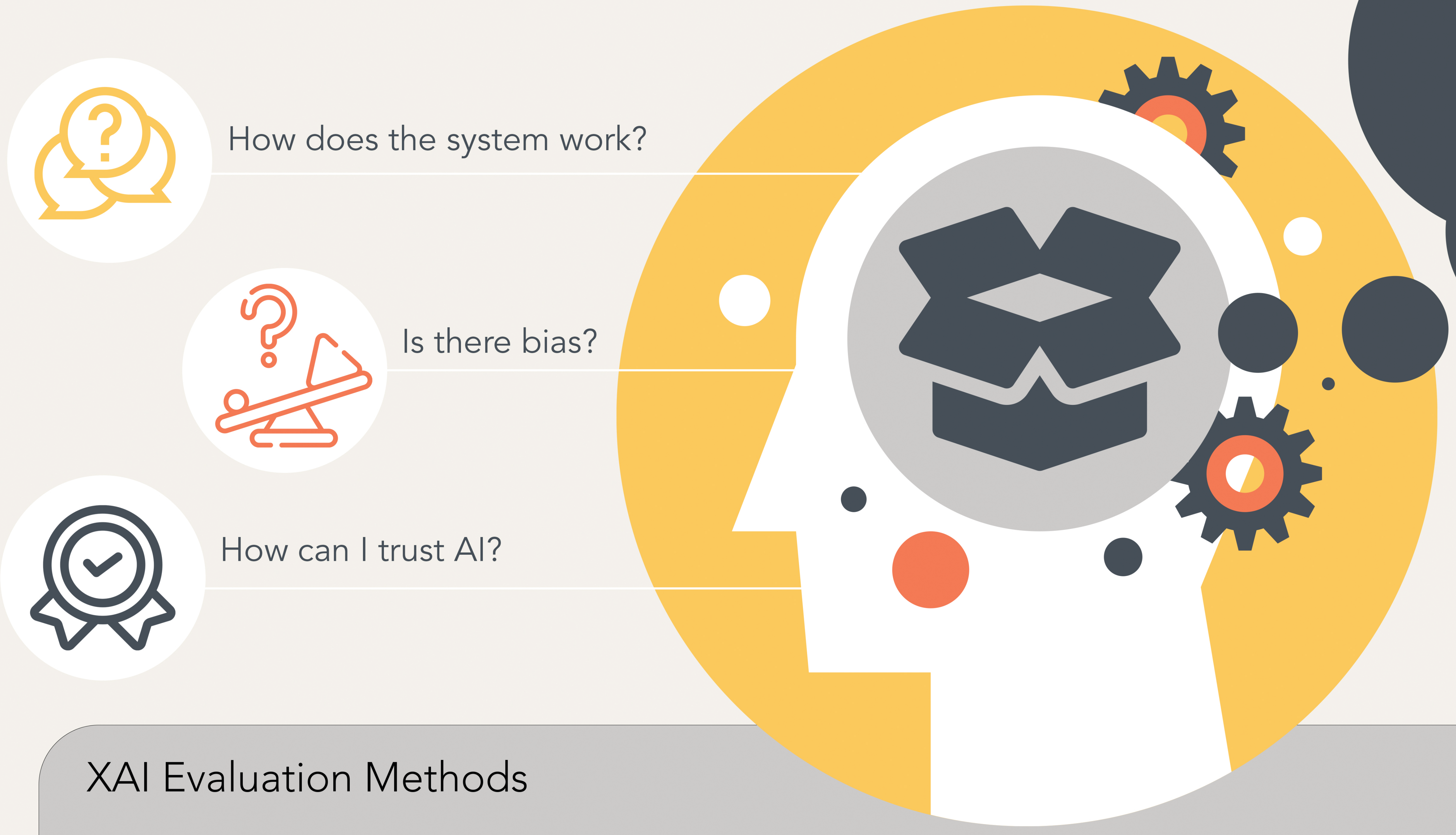
## Aim: How can we evaluate XAI from a user's perspective?

## Method

We investigated the leading practitioners of non-algorithmic XAI, those that explore methods used to evaluate XAI from a user perspective.

Evaluation methods which have been validated have been included.

An overview of not only how these methods of evaluation should be implemented but also what they evaluate has been examined.



How does the system work?

Is there bias?

How can I trust AI?

## XAI Evaluation Methods

| XAI | Non-XAI | XAI | Non-XAI |
|---|---|---|---|
| SCS: System Causability Scale [7]<br>Goodness Checklist [8]<br>Satisfaction Scale [8]<br>Curiosity Check [8]<br>XAI Trust Scale [8] | SUS: System Usability Scale [9]<br>SUMI: Software Usability Measurement Instrument [10]<br>WEBQUAL: Website quality measurement [11] | AAR/AI: After Action Review/AI [12]<br>CTAI: Cognative Tutorial For AI [13]<br>DoReMi [14]<br>AIQ Toolbox [13]<br>Mental Model Matrix [13]<br>Self Explaining Scorecard [13] | Design Thinking [15]<br>Wizard of Oz [16]<br>A/B Testing [17]<br>I like, I wish, What if? [18]<br>Shadowbox task [13] |
| Quantitative | | Qualitative | |

## Conclusion

When we evaluate XAI how do we know a user understands that explanation? What if we want to evaluate other factors apart from users' understanding? We have presented a comprehensive overview of XAI specific evaluation scales which assess understanding, usability, goodness, satisfaction, trust and curiosity and XAI specific qualitative methods to evaluate XAI with users. Along with non XAI methods these can be utilised to evaluate XAI, improve XAI methods and formats and ultimately increase users' trust and understanding of AI.

Our initial research investigated users' mental models for XAI using a design thinking approach [15]. We followed with an expert evaluation of XAI formats and methods using an I like, I wish, What if? method. These form our initial, exploratory investigations which will inform our next steps; a more controlled series of evaluations using XAI scales and XAI evaluation methods. This work will lead to the iterative design and evaluation of XAI formats and methods for end users of AI systems.

## Next Steps

Following from our evaluation of users' mental models and expert evaluation we will implement further XAI evaluation methods.

- A/B Testing of XAI formats & methods with lay users
- Goodness check of XAI methods with AI experts
- Satisfaction scale & XAI trust scale with lay users
- I like, I wish, What if? evaluation
- Iterative design and evaluation of XAI methods and formats for the design of an XAI pattern library

[1] Boyd, K., 2022, June. Designing up with value-sensitive design: Building a field guide for ethical ML development. In 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 2069-2082)
[2] European Commission, 2021. Artificial Intelligence Act. Available at: https://artificialintelligenceact.eu/the-act/. Accessed: 01/11/2023
[3] Lloyd, K., 2018. Bias amplification in artificial intelligence systems. arXiv preprint arXiv:1809.07842.
[4] von Eschenbach, W.J., 2021. Transparency and the black box problem: Why we do not trust AI. Philosophy & Technology, 34(4), pp.1607-1622.
[5] Wischmeyer, T., 2020. Artificial intelligence and transparency: opening the black box. Regulating artificial intelligence, pp.75-101.
[6] Zhu, J., Liapis, A., Risi, S., Bidarra, R. and Youngblood, G.M., 2018, August. Explainable AI for designers: A human-centered perspective on mixed-initiative co-creation. In 2018 IEEE conference on computational intelligence and games (CIG) (pp. 1-8). IEEE.
[7] Holzinger, A., Carrington, A. and Müller, H., 2020. Measuring the quality of explanations: the system causability scale (SCS) comparing human and machine explanations. KI-Künstliche Intelligenz, 34(2), pp.193-198.
[8] Hoffman, R.R., Mueller, S.T., Klein, G. and Litman, J., 2023. Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. Frontiers in Computer Science, 5, p.1096257
[9] Brooke, J., 1996. Sus: a "quick and dirty'usability. Usability evaluation in industry, 189(3), pp.189-194.
[10] Kirakowski, J., 1996. The software usability measurement inventory: background and usage. Usability evaluation in industry, pp.169-178.
[11] Loiacono, E.T., Watson, R.T. and Goodhue, D.L., 2002. WebQual: A measure of website quality. Marketing theory and applications, 13(3), pp.432-438
[12] Mai, T., Khanna, R., Dodge, J., Irvine, J., Lam, K.H., Lin, Z., Kiddle, N., Newman, E., Raja, S., Matthews, C. and Perdriau, C., 2020, March. Keeping it" organized and logical" after-action review for AI (AAR/AI). In Proceedings of the 25th International Conference on Intelligent User Interfaces (pp. 465-476
[13] Mueller, S., Tan, Y.Y., Linja, A., Klein, G. and Hoffman, R., 2021. Authoring Guide for Cognitive Tutorials for Artificial Intelligence: Purposes and Methods
[14] Schoonderwoerd, T.A., Jorritsma, W., Neerincx, M.A. and Van Den Bosch, K., 2021. Human-centered XAI: Developing design patterns for explanations of clinical decision support systems. International Journal of Human-Computer Studies, 154, p.102684.
[15] 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.
[16] Kelley, J.F., 1984. An iterative design methodology for user-friendly natural language office information applications. ACM Transactions on Information Systems (TOIS), 2(1), pp.26-41.
[17] King, R., Churchill, E.F. and Tan, C., 2017. Designing with data: Improving the user experience with A/B testing. " O'Reilly Media, Inc.".
[18] Rekonen, S., 2017. Unlocking the potential of interdisciplinary teams. In Passion-based co-creation (pp. 90-101). Aalto University