May 13, 2024

Comparing Implementations of Explainable Artificial Intelligence using End-User Evaluations: Protocol for a Systematic Review

DOI
dx.doi.org/10.17504/protocols.io.n92ld89e8v5b/v1

Mieke Ronckers[1,2], Rianne Conijn[1,2], Chris Snijders[1]

[1]Human-Technology Interaction group, Eindhoven University of Technology;
[2]Eindhoven Artificial Intelligence Systems Institute

Mieke Ronckers
Human-Technology Interaction group, Eindhoven University of ...

DOI: dx.doi.org/10.17504/protocols.io.n92ld89e8v5b/v1

Protocol Citation: Mieke Ronckers, Rianne Conijn, Chris Snijders 2024. Comparing Implementations of Explainable Artificial Intelligence using End-User Evaluations: Protocol for a Systematic Review. **protocols.io** **https://dx.doi.org/10.17504/protocols.io.n92ld89e8v5b/v1**

Protocol status: Working

Created: April 29, 2024

Last Modified: May 13, 2024

Protocol Integer ID: 98935

Keywords: Explainable AI, Systematic review, End-user evaluation

Abstract

In this protocol, we outline the steps for a systematic review on different implementations of Explainable Artificial Intelligence (XAI) evaluated by end-users. We aim to answer the research question: How do different implementations of XAI influence human-AI interaction? Therefore, we search for studies that compare different implementations of XAI to each other on human-AI interaction outcomes, such as user-perception measures and collaboration performance measures. We aim to create an overview of empirical studies that involve humans in the evaluation of XAI and compare different XAI implementations to each other, to gain more understanding on the factors contributing to a good explanation of AI systems and how this could be potentially different between application domains.

Guidelines

In this protocol, we followed the PRISMA guidelines for systematic reviews (Shamseer et al., 2015).

Introduction

1 **Rationale**

The field of Artificial Intelligence (AI) is rapidly expanding, with AI technologies increasingly integrated into everyday life. Concerns about the lack of transparency and interpretability in many opaque AI systems have led to a growing interest in Explainable AI (XAI), aimed at making AI systems and their outcomes understandable to users. Current studies in XAI often focus on algorithm and expert perspectives. In these studies, explanations are often meant to help developers improve the AI system, or the design of the explanation is based on the researcher's intuition of a 'good' explanation (Miller, 2019).

However, AI systems are often applied in non-technical domains, such as healthcare (Bussone et al., 2015; Cai et al., 2019), criminal justice (Dodge et al., 2019), and education (Farrow, 2023). This means that end-users are often laymen with no technical background. They need explanations to understand the underlying principles of AI systems and to make effective decisions supported by the AI. Therefore, XAI solutions should be developed more for end-users and XAI research needs to shift towards human-centered approaches, involving end-users in the evaluation process (Laato et al., 2021; Wang et al., 2019). Currently, only a small proportion (20%) of XAI research involves human participants in evaluating explanations (Nauta et al., 2023).

A first step in the direction of more human-centered XAI evaluations is a literature review by (Rong et al., 2024). They give an overview of existing user studies in XAI, that compared various types of AI explanations to AI systems with no explanations. While most findings indicated that explanations improved different human-AI interaction outcomes, such as trust, understanding, usability, and collaboration performance, some did not. The challenge underlying these mixed results lies in the absence of a clear understanding of what elements in explanations contribute to positive human-AI interaction. Research should investigate which elements of the explanations contribute to a good explanation of AI. This requires studies that test the influence of one element in the explanation while keeping all other factors equal across conditions. In this systematic review, we search for studies that compare different implementations of XAI to each other on human-AI collaboration outcomes, not only comparing an explanation to a control condition without an explanation to identify studies that test the specific influence of separate elements in an explanation.

Additionally, it remains uncertain whether elements of the explanations have the same influence on human-AI interaction across different target groups or application domains. For example, (Haque et al., 2023) mention that XAI solutions are presented in different ways in different domains; in health-care-related systems explanations often need textual and visual(graphical) formats, whereas in law-related systems hybrid explanations are optional.

Therefore, in this systematic review, we intend to create an overview of the findings of existing research on different XAI implementations for different target groups and application domains.

2 **Objectives**

Research question:

With this systematic review, we aim to answer the following research question: *How do different implementations of explainable Artificial Intelligence (XAI) influence human-AI interaction, both subjectively (perceptions of humans on AI advice) and objectively (performance of human-AI interaction)?*

Objective:

We aim to create an overview of empirical studies that involve humans in the evaluation of XAI and compare different XAI implementations to each other, to gain more understanding on the factors contributing to a good explanation of AI systems and how this could be potentially different between
application domains.

Methods

3 **Elegibility criteria**

Studies will be selected according to the criteria outlined below.

Study designs

We will include all types of studies that involve human participants: e.g., user studies and experiments. Studies can be both qualitative and quantitative. We will exclude cross-sectional studies, case series, and case reports. Additionally, evaluation frameworks and literature studies will be excluded.

Participants

The population of interest is all types of users of AI systems, in all application domains. We will include empirical studies including human participants.

Interventions

The phenomenon of interest is AI-systems that present an explanation to realize “explainable” or “interpretable” AI, for any application domain. Recommender systems will be included as AI-systems as well. Participants should be aware that they are interacting with an AI system and they should be actively trying to understand the AI system and/or its output. Therefore, we will exclude studies where users are not clearly aware that recommendations are generated by AI or studies where AI provides recommendations without the user or use case specifically requesting it, such as music and series recommendations on streaming platforms.

Comparators

Studies should compare different XAI implementations with each other, which can be a comparison between XAI algorithms or methods (e.g., feature-based, example-based, counterfactuals, etc.) and/or comparison between the explanation and presentation methods (e.g., size, text or visual, tone, level of detail, completeness, etc.). We will exclude comparisons between different black-box AI-models without different XAI implementations or comparisons between factors that are associated to the original black-box AI-model, such as model accuracy, and timing of the errors.

Outcomes

Studies will be included if they measure any type of human-AI interaction outcome. This could be subjective experiences of the users, such as trust, likeability,
preference, etc., or objective measures, such as effectiveness, performance of the human-AI collaboration, decision time, anchoring and adjusting decisions,
etc.

Report characteristics

We will include articles published in English. Other languages will be excluded.
Articles should be published and peer-reviewed. Other types of articles, such as commentaries, letters, editorials, etc. will not be included. Additionally, papers only published in archive will be excluded. There will be no restriction on the year of publication, papers from all years will be included up until the day of the final search.

4 **Information sources**

We will search the electronic databases WebOfScience, Scopus, IEEE explore, and ACM DL. To ensure literature saturation, we will scan the reference lists of included studies or relevant reviews identified through the search. Additionally, we will use the AI-tool “Connected papers” to identify related papers of the included studies, in a similar manner to scanning the reference lists.

5 **Search strategy**

We will search for words related to XAI and user-studies. The general search strategy will be as follows:

```
(   XAI
   OR
   (   (explanation OR explainable OR explanatory OR interpretable OR intelligible OR explainability OR  interpretability OR intelligibility)
      NEAR
      (AI OR “Artificial Intelligence” OR “black-box” OR “machine learning” OR “recommender system*”)
   )
)

AND
(“user study” OR participant* OR “human subject*” OR “empirical study” OR “lab study” OR “user
```

evaluation” OR “human evaluation” OR “end-user*” OR “user experiment*”)

NOT
(“model calibration” OR “validation data” OR “model performance” OR “cross validation”)

Limits
The search will be limited to title, abstract, and keywords. If not possible for one of the databases, we will use a full-text search.

Example of search query in Web Of Science:
TS=(XAI OR ((explanation OR explainable OR explanatory OR interpretable OR intelligible OR explainability OR interpretability OR intelligibility) NEAR3 (AI OR “Artificial Intelligence” OR “black-box” OR “machine learning” OR “recommender system*”))) AND TS=(“user study” OR participant* OR “human subject*” OR “empirical study” OR “lab study” OR “user evaluation” OR “human evaluation” OR “end-user*” OR “user experiment*”) NOT TS=(“model calibration” OR “validation data” OR “model performance” OR “cross validation”)

The NEAR operator is used to check if these terms appear within a 15 words distance.

6 **Study records**

Data management
The results from the literature search will be downloaded from the databases and uploaded to Rayyan.ai, an online tool for organizing systematic reviews (Ouzzani et al., 2016). We will use the function for automatically resolving duplicates, for duplicates with a similarity of 97% of higher. Other duplicates will be manually screened by MR and removed if necessary. For screening, Rayyan.ai will be used for both the first screening round (title and abstracts), and the second round (full-text), making use of the “blind mode”. Since Rayyan does not offer an automatic full-text search, we will manually search and download the full-text versions and upload it to Rayyan for full-text screening.

Selection process
Prior to the formal screening process, we will perform a pilot to refine the screening criteria and to reach consensus among all authors. Then, MR will screen all search results on titles and abstracts, taking a lenient perspective. All studies that are clearly not meeting the criteria will be excluded by MR. Other studies will be presented to RC and CS as well for title-abstract screening. All authors will screen these
papers independently and blinded to other’s decisions.

After the title-abstract screening, we will obtain full reports for studies that are not excluded after the first round. Then, full-text screening will be done with a procedure similar to the title-abstract screening. First, MR will screen all papers and exclude papers clearly not meeting the criteria. Then, all authors will screen the full-text of the remaining papers independently and blinded to other’s decisions.

For all steps, disagreements will be resolved through discussion between all authors. We will record the reasons for excluding trials. The authors will not be blinded to the journal titles, study authors, or institutions.

Data collection process

The data of the included papers will be extracted by MR. To reduce bias or errors in the data extraction, RC and CS will check the extracted data from a random sample of the papers. Disagreements will be resolved through discussion. Extracted data will include research aim, methodology, participant demographics, XAI details, and all reported human-AI interaction outcomes. All extracted data will be collected in a table (through Excel or similar software). If (part of) the information is not presented in the paper and not available through supplementary materials, we will contact authors for additional data, with a maximum of three email attempts.

Data items

We will extract the following information from included papers:

- Research question and aim, hypotheses
- Information about the AI system: type of AI model, accuracy, confidence, etc.
- Design of XAI implementations, e.g., visuals, text, XAI method/algorithm, exact wording.
- Application domain and the presented scenario/script
- Study design, such as number of conditions, procedure, etc.
- Measurements of human-AI interaction (e.g., questionnaires, time measurements, performance measurments, interview questions)
- Sample sizes and participant characteristics
- Results of XAI implementations on human-AI interaction outcomes (statistical tests, p-values, effect sizes, themes and codes)
- Results of other factors influencing or mediating the results (e.g., individual characteristics)

7 **Outcomes and prioritization**

The primary outcome is to create an overview of the XAI implementations that are compared in the included studies, the different factors presented in the explanation, and how these influence human-AI interaction, both subjectively and objectively.

As secondary outcome, we will create an overview of the findings for different domains (health, law, education, etc.) and types of users (experts/laymen, personalities), and how the influence of XAI implementations on human-AI interaction outcomes might be differ.

8 **Risk of bias in individual studies**

For risk of bias assessment, we will use the Mixed Methods Appraisal Tool (MMAT; Hong et al., 2018). This tool is most appropriate, since we include all types of user-studies (qualitative and quantitative, with different types of design). We will visually represent and critically discuss the overall scores and the scores per criteria. If a study has an extreme overall score, and can be seen as an outlier (threshold still to be discussed), we can exclude this study from the review.

9 **Data synthesis**

We will not be performing a meta-analysis, because we expect that the included studies will lack the level of similarity needed to perform a meta-analysis. Therefore, we will provide a systematic narrative synthesis with information presented in both text and tables to summarise and explain the characteristics and findings of the included studies. The narrative synthesis will be used to create an overview of all empirical studies that involve humans and compare different XAI implementations to each other, to gain more understanding on the factors contributing to a good explanation of AI systems and how this could be potentially different between application domains.

10 **Confidence in cumulative evidence**

We do not plan to perform an assessment of the confidence in cumulative evidence, because we are

not performing a meta-analysis. However, we plan to create an overview of the distribution of application domains, AI-methods, user populations, etc. We aim to critically review the state-of-the-art of user-studies in XAI, and to identify potential gaps in the targeted populations or domains and to identify potential differences between different target groups.

Acknowledgements

11 **Contributions**

MR is the guarantor, drafted the manuscript, developed the search strategy, and chose the risk of bias assessment strategy. All authors contributed to the development of the selection criteria and data extraction criteria. Additionally, all authors read, provided feedback and approved the final manuscript.

12 **Financial support**

The PhD position, and therefore this project, has been partly funded by Eindhoven AI Systems Institute (EAISI) and partly by the Human-Technology Interaction Group.

Protocol references

Bussone, A., Stumpf, S., & O'Sullivan, D. (2015). The role of explanations on trust and reliance in clinical decision support systems. *Proceedings - 2015 IEEE International Conference on Healthcare Informatics, ICHI 2015*, 160–169. https://doi.org/10.1109/ICHI.2015.26

Cai, C. J., Winter, S., Steiner, D., Wilcox, L., & Terry, M. (2019). "Hello Ai": Uncovering the onboarding needs of medical practitioners for human–AI collaborative decision-making. In *Proceedings of the ACM on Human-Computer Interaction* (Vol. 3, Issue CSCW). Association for Computing Machinery. https://doi.org/10.1145/3359206

Dodge, J., Vera Liao, Q., Zhang, Y., Bellamy, R. K. E., & Dugan, C. (2019). Explaining models: An empirical study of how explanations impact fairness judgment. *International Conference on Intelligent User Interfaces, Proceedings IUI*, *Part F147615*, 275–285. https://doi.org/10.1145/3301275.3302310

Farrow, R. (2023). The possibilities and limits of XAI in education: a socio-technical perspective. *Learning, Media and Technology*, *48*(2), 266–279. https://doi.org/10.1080/17439884.2023.2185630

Haque, A. B., Islam, A. K. M. N., & Mikalef, P. (2023). Explainable Artificial Intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research. *Technological Forecasting and Social Change*, *186*. https://doi.org/10.1016/j.techfore.2022.122120

Hong, Q. N., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M.-C., Vedel, I., & Pluye, P. (2018). The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. *Education for Information*, *34*(4), 285–291. https://doi.org/10.3233/EFI-180221

Laato, S., Tiainen, M., Najmul Islam, A. K. M., & Mäntymäki, M. (2021). How to explain AI systems to end users: a systematic literature review and research agenda. *Internet Research*, *32*(7), 1–31. https://doi.org/10.1108/INTR-08-2021-0600

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. In *Artificial Intelligence* (Vol. 267, pp. 1–38). Elsevier B.V. https://doi.org/10.1016/j.artint.2018.07.007

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., Van Keulen, M., & Seifert, C. (2023). From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *ACM Computing Surveys*, *55*(13). https://doi.org/10.1145/3583558

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan—a web and mobile app for systematic reviews. *Systematic Reviews*, *5*(1), 210. https://doi.org/10.1186/s13643-016-0384-4

Rong, Y., Leemann, T., Nguyen, T. T., Fiedler, L., Qian, P., Unhelkar, V., Seidel, T., Kasneci, G., & Kasneci, E. (2024). Towards Human-Centered Explainable AI: A Survey of User Studies for Model Explanations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*(4), 2104–2122. https://doi.org/10.1109/TPAMI.2023.3331846

Shamseer, L., Moher, D., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., Stewart, L. A., Altman, D. G., Booth, A., Chan, A. W., Chang, S., Clifford, T., Dickersin, K., Egger, M., Gøtzsche, P. C.,
Grimshaw, J. M., Groves, T., Helfand, M., … Whitlock, E. (2015). Preferred reporting items for systematic review and meta-analysis protocols (prisma-p) 2015: Elaboration and explanation. In *BMJ (Online)* (Vol. 349). BMJ Publishing Group. https://doi.org/10.1136/bmj.g7647

Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019, May 2). Designing theory-driven user-centric explainable AI. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3290605.3300831