XAI for Group-AI Interaction: Towards Collaborative and Inclusive Explanations

Mohammad Naiseh[1*], Catherine Webb[2], Tim Underwood[2], Gopal Ramchurn[2], Zoe Walters[2], Navamayooran Thavanesan[2], Ganesh Vigneswaran[2]

[1]Bournemouth University, Poole, United Kingdom)
[2]University of Southampton, Southampton, United Kingdom

Abstract

The increasing integration of Machine Learning (ML) into decision-making across various sectors has raised concerns about ethics, legality, explainability, and safety, highlighting the necessity of human oversight. In response, eXplainable AI (XAI) has emerged as a means to enhance transparency by providing insights into ML model decisions and offering humans an understanding of the underlying logic. Despite its potential, existing XAI models often lack practical usability and fail to improve human-AI performance, as they may introduce issues such as overreliance. This underscores the need for further research in Human-Centered XAI to improve the usability of current XAI methods. Notably, much of the current research focuses on one-to-one interactions between the XAI and individual decision-makers, overlooking the dynamics of many-to-one relationships in real-world scenarios where groups of humans collaborate using XAI in collective decision-making. In this late-breaking work, we draw upon current work in Human-Centered XAI research and discuss how XAI design could be transitioned to group-AI interaction. We discuss four potential challenges in the transition of XAI from human-AI interaction to group-AI interaction. This paper contributes to advancing the field of Human-Centered XAI and facilitates the discussion on group-XAI interaction, calling for further research in this area.

Keywords

Explainable AI, Group-AI Interaction, Interaction Design

1. Introduction

eXplainable AI (XAI) has emerged as a research direction in response to the lack of explainability and interpretability of AI models [6]. XAI aims to enhance the transparency of AI models, ML in particular, by providing human decision-makers with insights into the inner workings of ML models [21]. XAI makes ML outputs more interpretable and comprehensible by demystifying the complex processes within ML models. Approaches such as feature importance, example-based and counterfactual explanations have been developed for that purpose [1]. XAI seeks to bridge the gap between the technical complexity of these models and the need for human-understandable outputs by unravelling the intricacies of ML model decisions.

Late-breaking work, Demos and Doctoral Consortium, colocated with The 2nd World Conference on eXplainable ArtificialIntelligence: July 17–19, 2024, Valletta, Malta

*Correspondint author

mnaiseh1@bournemouth.ac.uk (M. Naiseh)

https://orcid.org/0000-0002-4927-5086 (M. Naiseh),

Despite the potential benefits of XAI, many existing XAI models lack practical usability and fail to improve human-AI performance [1]. Decision-makers often perceive explanations as tools designed for data scientists and ML engineers [7], leading to disinterest and reluctance to engage with XAI interfaces [8]. Decision-makers also report not seeing explanations as motivating to learn and solve problems; they show a lack of interest and curiosity unless these explanations align with their initial expectations [1]. This is consistent with findings from cognitive psychology showing that users tend to focus on features that have apparent value for their decision-making process [8]. Research in Human-Computer Interaction (HCI) has identified several user issues in human-XAI interaction, such as misinterpreting explanations, highlighting the need for improved design solutions [9]. Explanations might also be ignored if they are overly abstract, as people tend to prioritise concrete information instead [8].

Researchers have suggested various approaches to operationalise XAI in human-AI settings. For instance, contextualising XAI design by incorporating domain-related information and empirical knowledge has shown promising results in enhancing user satisfaction and understanding [10]. Additionally, employing contrastive explanations and juxtaposing features can help develop an expert ability to notice salient features and anomalous events [11]. To enhance critical engagement with explanations and mitigate over-reliance on AI recommendations, incorporating cognitive forcing [12] and nudging [8] design elements into the XAI interface has been empirically tested. These elements have been shown to discourage decision-makers from blindly accepting AI suggestions and instead prompt them to evaluate recommendations thoughtfully. Furthermore, methodological approaches have been suggested to help UX designers operationalise XAI methods on the XAI user interface level [12]. These approaches have been shown to improve human-XAI interaction and increase participant engagement with AI explanations.

Interestingly, much of the existing research in Human-Centered eXplainable AI focuses on one-to-one interactions between the XAI and human decision-makers. In other words, it focuses on how individual decision-makers interact with AI explanations and use them to make decisions in human-AI settings. However, much of real-world scenarios often involve many-to-one relationships in a group-AI interaction, where a group of individuals collaborate to make collective decisions using the XAI interface. *Figure 1* explains this relationship. Group-AI interaction refers to the collaboration and interaction between groups of human decision-makers and AI systems in decision-making processes [17]. In this context, groups may include executive committees, boards, teams of professionals, or any collective of individuals tasked with making decisions within an organisation or context [15]. Group-AI interaction has the potential to leverage the capabilities of AI technologies alongside collective human expertise to enhance decision-making outcomes. It has also been shown to exceed the accuracy of human-AI collaboration by bringing different expertise and perspectives of humans involved in the decision-making process [7]. Individuals in the group can be assigned specialised roles and responsibilities related to interacting with AI systems, interpreting AI-generated insights, and integrating them into decision-making processes [1]. This division of responsibilities ensures that each member contributes their expertise effectively, mitigating the risks associated with human-AI collaboration [1].

In this paper, we argue that designing XAI for group-AI interaction requires distinct approaches and careful considerations compared to the ones used in human-AI interaction. This late-breaking contribution synthesises insights from collective decision-making and Human-Centered XAI literature to discuss challenges inherent in transitioning XAI from human-AI to group-AI interaction. These encompass the complexities of group dynamics, the potential amplification of cognitive biases, issues surrounding trust, as well as the critical facets of XAI evaluation in the context of group-AI. While acknowledging the possibility of additional challenges, our discussion provides an initial framework for contrasting the nuanced design requirements of XAI in facilitating AI-assisted decision-making within group settings.

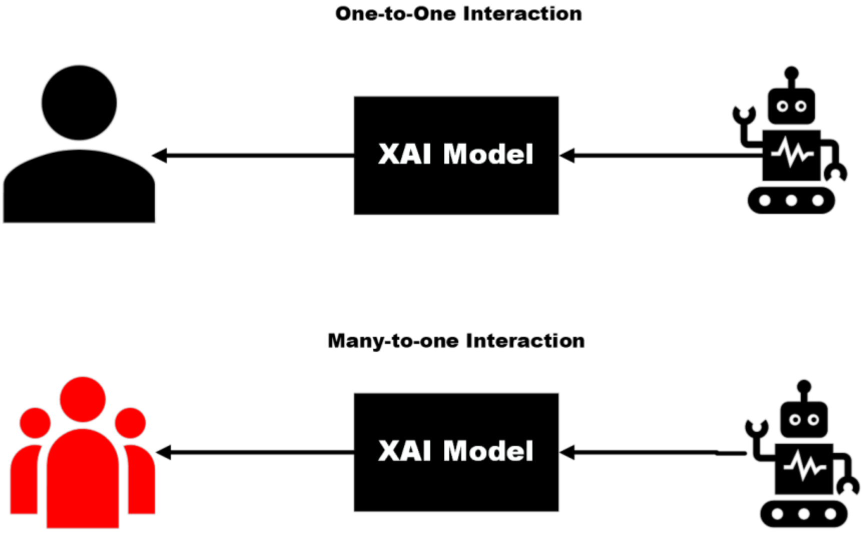

Figure 1: Comparison of One-to-One and Many-to-One Interactions with an XAI Interface. In the One-to-One interaction mode, a single user interacts directly with the XAI interface, receiving explanations tailored to their specific queries or actions. In contrast, the Many-to-One interaction mode involves multiple users interacting with the XAI interface simultaneously, with explanations generated to accommodate diverse user inputs and preferences.

2. The Complexity of Group Diversity

In group decision-making, individuals often wield varying degrees of influence and expertise, regardless of the context—be it a professional team, a board of directors, or a community organisation [19]. This diversity encompasses a nuanced interplay of individual personalities, social structures, power dynamics, and communication patterns within the group [21]. Such complexity in group diversity further impacts the interaction between groups and AI, particularly in XAI contexts. This diversity underscores the necessity for XAI systems to accommodate inclusive explanations tailored to the diverse needs and backgrounds of group members [20]. For example, individuals with varying levels of familiarity with AI and machine learning concepts may necessitate explanations that are lucid, accessible, and devoid of technical jargon. Moreover, the diversity within groups may prompt XAI developers to accommodate diverse learning styles and linguistic preferences, thereby enhancing the accessibility of XAI explanations [5]. This entails providing explanations in multiple formats or languages and integrating interactive features to facilitate engagement and comprehension among all group members [11]. Additionally, the array of group interactions may be further complicated by individual attitudes and perceptions toward AI technology [22]. Cultural values and norms, for instance, have been demonstrated to influence attitudes toward AI [1]. While some individuals may embrace XAI as a valuable tool for enhancing decision-making capabilities, others may harbour scepticism or resistance due to concerns about job displacement, loss of autonomy, or ethical implications. Consequently, XAI for group-AI interaction must address these diverse perspectives and cultivate a culture of trust, transparency, and open communication to mitigate resistance to XAI adoption and foster constructive collaboration within the group.

3. Bias Amplification in Group-XAI

Biases in group-AI interaction can be more pronounced than human-AI interaction, presenting significant challenges for the design and development of XAI systems [4] [17]. One prevalent bias is groupthink, where group members prioritise consensus and overlook dissenting viewpoints to maintain harmony [23]. In this context, if explanations do not encourage critical thinking and challenge groupthink, individuals within the group may unquestioningly accept AI-generated insights without thorough examination for explanations [13]. XAI systems should not only explain recommendations but also encourage scrutiny and diverse viewpoints [26]. This could involve presenting multiple explanations, highlighting uncertainties, and actively soliciting feedback from users with varying perspectives. XAI systems may also implement mechanisms for independent review and validation. For instance, introducing a Devil's Advocacy role within the team can challenge group consensus and encourage critical evaluation of XAI explanations [24]. This individual identifies potential flaws or biases, fostering a more balanced consideration of decision options.

Another bias that could impact XAI design in group-AI interaction scenarios is the equality bias. It refers to the tendency for individuals to downplay their expertise or to weigh everyone's opinion equally, regardless of competence or expertise [27]. This bias can have detrimental effects on decision-making processes, particularly when there is a genuine disparity in knowledge or experience within the group. In the context of XAI, the equality bias could be amplified when group members defer too readily to the AI recommendations, regardless of their domain expertise or experience in the subject matter [28]. For example, suppose a group of healthcare professionals is using an XAI system to diagnose patients. In that case, individuals with specialised medical knowledge may inadvertently downplay their expertise and defer to the AI's recommendations and explanations, even when they have valid insights or concerns that should be taken into account. XAI design shall account for such bias by designing explanations that encourage individuals to recognise and value their expertise and insights, as well as those of others within the group. Additionally, fostering a culture of collaboration and open communication within the group can help ensure that diverse perspectives and expertise are taken into account when making decisions with the assistance of AI systems. This may involve providing XAI explanations that highlight the knowledge and contributions of individual group members, as well as mechanisms for facilitating constructive dialogue and debate within the group.

4. Trust within the Group

Trust has been a crucial element in human-AI interaction, influencing the dynamics and effectiveness of human-AI teams [1]. When considering group-AI interaction, trust dynamics become more complex, involving not only trust between group members and AI but also among group members themselves. It has been discussed that incorporating XAI into group decision-making processes can impact trust dynamics among group members [25]. The integration of XAI into group decision-making processes introduces new dimensions to these dynamics, with potential implications for team cohesion and effectiveness. Suppose an explanation contradicts the opinions or recommendations of certain group members, it could create tensions or conflicts within the group, undermining trust and cohesion. In addition, some scenarios could involve individuals within the group perceiving XAI explanations as more reliable or objective than human judgments, which may lead to a shift in trust dynamics within the group [9]. To navigate these complexities and foster trust among group members, XAI development needs to consider the social dynamics of group interaction. Open communication about XAI's role in decision-making and clear explanations of its outputs are crucial for trust calibration. Additionally, establishing protocols for interpreting XAI explanations in context, along with mechanisms for addressing conflicts arising from AI-influenced decisions, can safeguard against trust erosion. This ensures that XAI's benefits are harnessed without jeopardizing ethical principles or human values.

5. Evaluating XAI for Group Interaction

Evaluating XAI for group interaction presents distinct challenges compared to traditional XAI evaluations focused on individual users. Traditionally, XAI examines how individuals interact with AI systems, understand explanations, and make decisions based on them [1] [6]. Here, evaluation metrics assess explanation clarity, relevance, and user satisfaction [9] [10]. Trust and interpersonal dynamics are also crucial factors, with conflicts arising from discrepancies between user expectations and AI behaviour, requiring strategies for resolution [1] [8]. However, in group-AI interaction, evaluation extends beyond individual users. We need to consider the entire group ecosystem. This includes how AI-generated explanations are communicated within the group, how they impact group cohesion and communication patterns, and how conflicts are resolved among members, considering factors like power dynamics and individual expertise [18]. Additionally, group-XAI evaluation involves understanding the social influence of explanations on the group's decision-making dynamics. This encompasses considerations of scalability (how well does XAI adapt to groups of varying sizes?) and consensus-building (how can XAI support groups in achieving agreement despite diverse perspectives?) [18].

Therefore, evaluating XAI for group interaction demands methodologies that account for the complexities of social interactions and group dynamics. This might involve incorporating social network analysis to understand how information flows within the group and identify potential bottlenecks or communication silos. Longitudinal studies could be conducted to assess the impact of XAI on group performance and decision-making quality over time. Ultimately, understanding these differences is crucial for designing effective XAI systems that empower both individual users and collaborative decision-making processes, while mitigating potential pitfalls and fostering a healthy group environment.

6. Conclusion and Future Directions

In conclusion, the integration of Machine Learning (ML) into decision-making processes across various sectors has prompted the development of XAI to address concerns regarding ethics, legality, explainability, and safety. In this paper, we showed that much of the existing research focuses on one-to-one interactions between XAI interfaces and individual decision-makers. However, many real-world scenarios require the interaction between a group of humans and the XAI interface. Building on the current research on Human-Centered XAI, this paper has discussed four key considerations when transitioning from human-AI to group-AI interaction in the context of XAI. These challenges include complexities in group dynamics, cognitive bias amplification, trust issues within the group, and group-centric evaluation. By drawing upon current work in Human-Centered XAI research, we contribute to advancing the field and facilitate discussions on group-XAI interaction. This paper calls for further research in this area to enhance the effectiveness and usability of XAI in collaborative decision-making settings, ultimately leading to more informed and successful outcomes in various domains.

References

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

[2] Anna B. Chouldechova, Emily Putnam-Hornstein, Andrew Tobin, and Rhema Vaithianathan. 2019. Toward algorithmic accountability in public services: A qualitative study of affected community perspectives on algorithmic decisionmaking in child welfare services. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 2019, 1–12. https://doi.org/10.1145/3290605.3300271

[3] Goodell, J.W., Kumar, S., Lim, W.M. and Pattnaik, D., 2021. Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis. Journal of Behavioral and Experimental Finance, 32, p.100577.

[4] Hall, J. and Williams, M.S., 1970. Group dynamics training and improved decision-making. The Journal of Applied Behavioral Science, pp.39-68.

[5] Naiseh, M., Jiang, N., Ma, J. and Ali, R., 2020. Personalising explainable recommendations: literature and conceptualisation. In Trends and Innovations in Information Systems and Technologies: Volume 2 8 (pp. 518-533). Springer International Publishing.

[6] Saeed, W. and Omlin, C., 2023. Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. \textit{Knowledge-Based Systems}, 263, p.110273.

[7] Naiseh, M., 2020. Explainability design patterns in clinical decision support systems. In Research Challenges in Information Science: 14th International Conference, RCIS 2020, Limassol, Cyprus, September 23–25, 2020, Proceedings 14 (pp. 613-620). Springer International Publishing.

[8] Mohammad Naiseh, Reem S. Al-Mansoori, Dena Al-Thani, Nan Jiang, and Raian Ali. 2021. Nudging through Friction: An approach for Calibrating Trust in Explainable AI. In 2021 8th International Conference on Behavioral and Social Computing (BESC), 2021, 1-5.

[9] Naiseh, M., Cemiloglu, D., Al Thani, D., Jiang, N. and Ali, R., 2021. Explainable recommendations and calibrated trust: two systematic user errors. \textit{Computer}, \textit{54}(10), pp.28-37.

[10] Clara Bove, Jonathan Aigrain, Marie-Jeanne Lesot, Charles Tijus, and Marcin Detyniecki. 2022. Contextualization and Exploration of Local Feature Importance Explanations to Improve Understanding and Satisfaction of Non-Expert Users. In 27th International Conference on Intelligent User Interfaces, 807-819.

[11] Chromik, M. and Butz, A., 2021. Human-XAI interaction: a review and design principles for explanation user interfaces. In Human-Computer Interaction–INTERACT 2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30–September 3, 2021, Proceedings, Part II 18 (pp. 619-640). Springer International Publishing.

[12] Naiseh, M., Simkute, A., Zieni, B., Jiang, N. and Ali, R., 2024. C-XAI: A Conceptual Framework for Designing XAI tools that Support Trust Calibration. Journal of Responsible Technology, p.100076.

[13] Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M.T. and Weld, D., 2021, May. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1-16).

[14] Dodge, J., Liao, Q.V., Zhang, Y., Bellamy, R.K. and Dugan, C., 2019, March. Explaining models: an empirical study of how explanations impact fairness judgment. In Proceedings of the 24th international conference on intelligent user interfaces (pp. 275-285).

[15] Choo, C.W., 1996. The knowing organization: How organizations use information to construct meaning, create knowledge and make decisions. International journal of information management, 16(5), pp.329-340.

[16] Thomas, J.P. and McFadyen, R.G., 1995. The confidence heuristic: A game-theoretic analysis. Journal of Economic Psychology, 16(1), pp.97-113.

[17] Hausken, K., 1995. The dynamics of within-group and between-group interaction. Journal of Mathematical Economics, 24(7), pp.655-687.

[18] Askarisichani, O., Bullo, F., Friedkin, N.E. and Singh, A.K., 2022. Predictive models for human–AI nexus in group decision making. Annals of the New York Academy of Sciences, 1514(1), pp.70-81.

[19] Black, D., 1948. On the rationale of group decision-making. Journal of political economy, 56(1), pp.23-34.

[20] Miller, T., 2019. Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence, 267, pp.1-38.

[21] Liao, Q.V., Gruen, D. and Miller, S., 2020, April. Questioning the AI: informing design practices for explainable AI user experiences. In Proceedings of the 2020 CHI conference on human factors in computing systems (pp. 1-15).

[22] Qi, R., Zheng, Y., Yang, Y., Zhang, J. and Hsiao, J., 2023. Individual differences in explanation strategies for image classification and implications for explainable AI. In Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 45, No. 45).

[23] Janis, I.L., 1972. Victims of groupthink: A psychological study of foreign-policy decisions and fiascoes.

[24] Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P.N., Inkpen, K. and Teevan, J., 2019, May. Guidelines for human-AI interaction. In Proceedings of the 2019 chi conference on human factors in computing systems (pp. 1-13).

[25] Schmid, U. and Wrede, B., 2022. What is missing in xai so far? an interdisciplinary perspective. KI-Künstliche Intelligenz, 36(3), pp.303-315.

[26] Simkute, A., Surana, A., Luger, E., Evans, M. and Jones, R., 2022, October. XAI for learning: Narrowing down the digital divide between “new” and “old” experts. In Adjunct Proceedings of the 2022 Nordic Human-Computer Interaction Conference (pp. 1-6).

[27] Mahmoodi, A., Bang, D., Olsen, K., Zhao, Y.A., Shi, Z., Broberg, K., Safavi, S., Han, S., Nili Ahmadabadi, M., Frith, C.D. and Roepstorff, A., 2015. Equality bias impairs collective decision-making across cultures. Proceedings of the National Academy of Sciences, 112(12), pp.3835-3840.

[28] El-Assady, M. and Moruzzi, C., 2022. Which biases and reasoning pitfalls do explanations trigger? Decomposing communication processes in human–AI interaction. IEEE Computer Graphics and Applications, 42(6), pp.11-23.