

Social eXplainable AI (Social XAI): Towards expanding the social benefits of XAI

Mohammad Naiseh

Faculty of Science and Technology, Bournemouth University, United Kingdom

Abstract .

The eXplainable Artificial Intelligence (XAI) research agenda is constantly expanding. From a mainly technical topic, it has now developed into a multidisciplinary area of research, supporting domain experts, policymakers, and lay users. However, humans working in domains, such as healthcare, media and policing still find themselves questioning the value of XAI and whether it is worth their time and effort. This hampers effective human-machine collaboration and results in various trust-related biases. An emerging field of social XAI introduces the need to operationalise XAI and overcome its potential user issues. For example, XAI has been shown to lead to over-reliance on AI, for some in contrast to algorithmic aversion or is simply ignored as a redundant feature. Based on current work in Human-Centered XAI, I outline a set of design features that could improve human-XAI interaction and ultimately expand the social benefits of XAI. I present a framework of three dimensions that shows how the current work in Human-Centered XAI is increasing the effectiveness of the human-XAI interaction.

Keywords:

Explainable AI, Social XAI, Human-Centered AI.

1 Introduction

Artificial Intelligence (AI) is becoming an important part of many jobs, supporting decision-making in public services [1], healthcare [2], child welfare services [3], and journalism [4]. In some cases, AI is intended to automate mundane, repetitive tasks and to allow humans to focus on higher-level and creative tasks [5]. However, humans show resentment toward AI-driven decision-making systems as these systems can be inscrutable and nonintuitive [6]. Humans report an inability to understand the logic behind AI outputs and have problems establishing meaningful trust in such an AI system [7]. Without being able to understand and accurately judge AI outputs, decision-makers are prevented from harnessing the benefits algorithmic systems can offer [8].

In recent years, the field of eXplainable Artificial Intelligence (XAI) has seen a surge in research aimed at enhancing AI transparency and interpretability [6]. Human-Computer Interaction (HCI) research has also looked into various aspects of user-centred explanations [4], encompassing presentation, information content, framing, and communication methods. XAI is increasingly recognised as a crucial component of human-AI systems, serving to calibrate user reliance and mitigate potential errors stemming from blind adherence to AI decisions [5]. Despite the emphasis on providing explanations, studies have shown that the performance and reliance of human-AI teams have not necessarily improved [5,7,8]. Users have reported that explanations sometimes feel redundant and fail to meet task requirements [5], leading to misuse or misinterpretation [9]. This is partly because XAI approaches often offer only short-term solutions, losing relevance over time [10]. When explanations are perceived as redundant, they risk being

ignored or superficially examined [9]. Moreover, XAI often lacks the ability to spark curiosity or motivate experts to delve deeper into problem-solving [11]. There is thus a pressing need to translate XAI into tangible measures and functions within human-AI environments, expanding its social benefits. Additionally, several researchers noted that XAI to be socially effective must address not only technical aspects but also human factors, such as cognitive biases, and raise awareness about XAI [12,13,14].

Several studies already demonstrated how XAI can be used beyond just explaining AI models and expanding its social benefits. In this paper, I refer to social XAI as approaches, tools and design techniques that are to be implemented with XAI systems with the aim that not just explaining AI but also providing benefits to users and the human-AI team. The IMPACT framework structuring this book also makes the point that XAI is highly relevant to foster transparency in the realm of AI likely resulting in more positive views on the AI technology (see Chapter 1 and Montag et al., 2024a/b). Social XAI has shown promising results to upskill current XAI users. For example, interactive features with XAI helped users discover unknown rules describing certain phenomena and upskilled users in AI terms [18]. Similarly, receiving feedback regarding the reasons for the success or failure of the system has been shown to help users learn how to avoid potential AI mistakes [20]. Social XAI helped users train for difficult social situations or high-risk tasks in a safe and controlled way [19] or foster expertise development [21], by using techniques such as cognitive forcing [22]. Cognitive forcing functions are HCI interventions that disrupt heuristic reasoning and encourage users to engage in more deliberate and thoughtful decision-making processes by making the process more mentally demanding [22]. In another example, social XAI has helped in teaching new surgeons, by first learning different aspects of expert surgeon performance and then using this information to enhance novices' understanding of medical data [23]. Further, social XAI has shown its potential to be effective in teaching AI to children. Hiltron et al [24] showed that children can be taught ML concepts when they can experiment with explainable models. The ability to interact with an interpretable model significantly improved 10-14-year-olds' understanding of the concept of bias in terms of gender discrimination [25].

Social XAI also allowed domain experts to make data-supported decisions to improve the systems they are using and even make them fairer. DeVos et al [26] showed that decision-makers have a strong ability to detect and audit potentially harmful algorithmic behaviours when given explanations that allow them to apply their sensemaking strategies. Social explanations are part of user-driven algorithm audits and help leverage users' prior experience of harm and biases [26]. For example, when XAI enables decision-makers to make holistic risk assessments and adjust for algorithmic limitations, they can reduce the racial disparity in child welfare screening in comparison to if AI was used alone [27]. Moreover, social

XAI can help users reflect on their own decisions and potentially reduce the potential for biases in the human AI team [31]. Hun Lee and colleagues [30] showed that social XAI can be used to create collective, hybrid intelligence on a complex decision-making task with improved accuracy, and improved consistency of decision-making. These examples illustrate that the importance of XAI extends beyond simply explaining AI decisions to humans. Instead, it is crucial to design XAI that is socially beneficial and minimises biases, enabling more effective and equitable human-AI collaboration.

The significance of this research lies in its ability to bridge the gap between technical advancements in XAI and the human dynamics essential for the successful adoption and utilisation of XAI systems. The author builds upon insights from human-centred XAI research to present a robust framework of design features aimed at expanding the social benefits of XAI within human-AI ecosystems. The framework not only aims to optimise user engagement with XAI interfaces but also to facilitate profound social benefits to XAI users

2 Social XAI framework

Based on the current literature on Human-Centered eXplainable Artificial Intelligence (XAI), I present a social XAI framework that emerges as a synthesis of research insights and design principles aimed at enhancing the *usability, human-like, and educational* value of XAI.

2.1 Usability Enhancement

The concept of usability enhancement within XAI has been discussed by scholars to underscore the role of user interface design in expanding the societal benefits of AI technologies. This dimension embodies approaches to interface design that prioritise accessibility and engagement of AI explanations to ultimately encourage users to interact with AI explanations. It has been shown that encouraging interaction with XAI explanations can expand the social benefits of XAI by fostering greater transparency, trust, and collaboration between users and AI systems [4, 8, 16]. When users actively engage with explanations, they develop a deeper understanding of AI outputs, enabling them to make more informed decisions and contribute meaningfully to decision-making processes. Additionally, increased interaction with XAI explanations promotes user empowerment and accountability, facilitating the responsible and ethical deployment of AI technologies in various domains [18,19].

One related concept to usability enhancement is *explanation abstraction*. Researchers have used several techniques to break down complex explanations into digestible chunks e.g., [8,18], as advocated by the reduction principle of the Persuasive Systems Design (PSD) model [25]. By aligning with human

cognitive processing capabilities, this approach ensures that explanations are easily understood and retained by users across different knowledge levels and backgrounds [26].

Furthermore, *easy navigation* through XAI interfaces has been identified as a key determinant for usability enhancement [14,26]. Designers in this case aim to make the explanations easy to find and navigate through the design. Therefore, users are more likely to interact and engage with explanations. Similarly, *tunnelling* design has been used in XAI to guide users through the process of reading and interacting with the explanation [23,31]. In one related research, authors have used visual cues to highlight explanations' chunks which are important to read or using navigational cues to switch from one section of the interface to another [23]. This aligns with the phenomena related to learners' readiness to learn and read. Learners usually tend to follow measures and instructions in which the instructor guides the learning process [27].

Moreover, *interactive XAI* features have been used to allow users to interact with the explanation and ask questions [25]. Interactive XAI interfaces proved to be an effective way to encourage users to interact with XAI and enable them to construct their own experiences [22]. For instance, in one relevant study, researchers have used interactive Local Interpretable Model-agnostic Explanations (LIME) [30] to explain cancer prediction. The interactive XAI allowed users to group and split cancer image segments and experiment with different explainability scenarios. Such interactive and experimental interaction between the XAI system and the users was helpful in improving users' learning experience and serving various learning styles [29].

Finally, the literature discussed the role of *personalised presentation* of explanations in increasing engagement with XAI features across diverse contexts and user preferences [33]. Personalisation in XAI involves tailoring the presentation or the framing of explanations to individual users based on their characteristics, such as demographics, behaviour, and preferences. One related study discussed personalising XAI explanations based on user demographics could involve tailoring the content and presentation of explanations to match the user's age [20]. For younger or less technically inclined users, the XAI system may provide more visual, simplified explanations using plain language. In contrast, for older or more technically-savvy users, the system may offer more detailed, technical explanations with advanced visualizations and data.

2.2 Human-Like Enhancement

In this discussion, I delve into the attributes of explanations that are significant from the perspective of the explainee. These attributes, drawn from research in social science, psychology, and

cognitive science, aim to ensure that algorithmic explanations resonate naturally with end users, resembling human-like and everyday social explanations.

One prominent feature of the human-like enhancement of XAI is *novelty*, which has been widely discussed as a means to capture users' attention and encourage interaction with AI explanations [6]. Novelty involves providing in-depth knowledge and surprising facts while avoiding circularity in the information presented. Repetitive information and obvious assertions can render explanations meaningless and deter users from engaging with them [9, 12]. For instance, consider an AI tool generating diet plan recommendations; a circular explanation such as "*This diet plan is effective because it helps lose weight*" may fail to captivate users' interest. *Coherence* is another critical aspect of human-like enhancement, wherein explanation chunks and components must work in harmony to encourage user interaction. Studies have demonstrated that users may disregard explanations that lack coherence or fail to "hang together" [6]. For example, an AI-based system issuing security device recommendations with an explanation like "*This is the most secure device because it has the cheapest delivery cost*" presents an incoherent rationale, likely to be ignored by users. In other words, the explanations provided must be logical.

Generalisable explanations that encompass a broader range of cases and facts have been found to elicit higher levels of acceptance from users [17,21]. This principle is important, as it suggests that explanations should aim to capture the underlying patterns and principles governing the AI's decision-making, rather than focusing solely on the details of a particular instance. This has been shown to improve human-AI interaction as users are better able to understand the system's reasoning and apply that understanding to new situations [26]. While this principle may seem contradictory to the novelty feature, general explanations tend to be more engaging in the absence of novelty, as novel explanations are contextually limited. The key is to strike a balance, where the explanations are informative and insightful enough to maintain user engagement, while still being sufficiently generalised to be applicable across a range of scenarios.

Lastly, the concept of *Contractiveness* has emerged as a significant feature in the literature on human-centered XAI [15,16,32]. Contractive explanations not only elucidate the current AI prediction but also illustrate counterfactual cases, addressing the human tendency to question why a particular action was taken instead of another. By highlighting the alternative actions the AI could have taken and the reasons for the chosen course of action, contractive explanations help users better comprehend the underlying logic and reasoning behind the AI's decisions [7].

2.3 Learning Enhancement

In this section, I explore how several researchers have endeavoured to develop a *learning-focused XAI* to effectively upskill users and foster enduring mental models of AI operations. The interest in such learning-focused XAI has arisen from empirical evidence indicating that individuals working with AI may lack the algorithmic literacy needed to interpret AI outputs [12]. Additionally, they may harbour fundamental misconceptions about the workings of algorithmic systems or oversimplify their operations [34].

Several studies have developed XAI systems with *guidance features* to help users in the process of interpreting AI explanations [30,31], e.g., information such as when, how, and why to use explanations was used for a better learning experience. This included video tutorials or FAQ tutorials accessible to users [35]. Another approach was to promote *Self-learning* in the XAI interface to enhance the learning experience of XAI [16,30]. Researchers have argued that people have different learning styles and limited cognitive resources which might require them to follow different learning strategies [23]. In this context, the XAI developers have designed their interfaces to help users self-learn from the explanation by providing tools that facilitate that process, e.g., taking notes or archiving explanations for future interactions [14]. In this way, users were able to personalise their learning strategies, aims and learning sources without help from the system.

In addition, *suggestion features* have been used to provide users with suggestions for additional reading materials or resources related to AI concepts and explanations [23]. It has been shown that XAI can deepen user's understanding of AI principles and applications by recommending relevant articles, papers, or educational content. These suggestions can be tailored to users' interests, knowledge levels, and learning objectives, enhancing the effectiveness of the learning experience [34]. Finally, *collaborative learning* functionalities have been shown to enable users to engage in collaborative activities, discussions, and knowledge sharing with peers, experts, or AI systems themselves [35]. XAI was designed to leverage collective intelligence, diverse perspectives, and shared experiences to enhance learning outcomes by facilitating collaboration among users [36]. For example, in a collaborative learning scenario within an XAI system, users were able to participate in discussion forums or online communities where they could share their experiences, insights, and questions related to AI explanations [36]. Users were also able to collaborate with peers to analyse and interpret AI outputs, discuss best practices, and troubleshoot challenges together [37].

Social XAI Framework

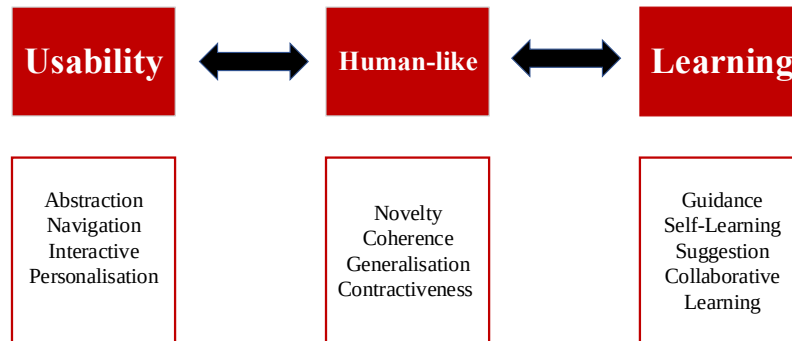


Figure 1 The Social XAI Framework encompasses three dimensions: Usability, Human-Line, and Learning Enhancement.

3 Research Gap and Future Directions

The Social XAI framework presented herein synthesises insights from Human-Centered eXplainable Artificial Intelligence (XAI) literature, offering a structured approach to enhance the usability, social, and educational aspects of AI systems. Through an examination of usability, human-like, and learning dimensions, this framework underscores the importance of user interaction, engagement, and empowerment in realising the full potential of XAI. While the framework provides valuable guidelines for designing XAI systems, it is not comprehensive. The main aim of this chapter is to open the discussion for further research and refinement of the framework to encompass a broader range of factors and contexts relevant to social XAI. Future research will focus on validating and expanding the Social XAI framework through empirical studies, user feedback, and real-world applications. Additionally, efforts should be directed towards addressing the ethical, legal, and societal implications of XAI, ensuring responsible and equitable deployment of XAI. Collaborative initiatives involving interdisciplinary collaboration, industry partnerships, and community engagement will be essential in advancing the field of Social XAI and addressing its limitations.

4 References

- [1] Davenport, T. and Kalakota, R., 2019. The potential for artificial intelligence in healthcare. *Future healthcare journal*, 6(2), p.94.
- [2] Rasch, R., Kott, A. and Forbus, K.D., 2003. Incorporating AI into military decision making: an experiment. *IEEE Intelligent Systems*, 18(4), pp.18-26.
- [3] 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).
- [4] Naiseh, M., Jiang, N., Ma, J. and Ali, R., 2020, September. Explainable recommendations in intelligent systems: delivery methods, modalities and risks. In *International Conference on Research Challenges in Information Science* (pp. 212-228). Springer, Cham.

- [5] 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.
- [6] Sokol, K. and Flach, P., 2020, January. Explainability fact sheets: a framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 56-67).
- [7] Sokol, K. and Flach, P., 2020. One explanation does not fit all. *KI-Künstliche Intelligenz*, pp.1-16.
- [8] Bansal, G., Wu, T., Zhu, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M.T. and Weld, D.S., 2020. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. *arXiv preprint arXiv:2006.14779*.
- [9] Naiseh, M., Cemiloglu, D., Al Thani, D., Jiang, N. and Ali, R., 2021. Explainable recommendations and calibrated trust: two systematic user errors. *Computer*, 54(10), pp.28-37.
- [10] Rozenblit, L. and Keil, F., 2002. The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive science*, 26(5), pp.521-562.
- [11] Min K. Lee, Ji T. Kim, and Leah Lizarondo. 2017. A human-centred approach to algorithmic services: Considerations for fair and motivating smart community service management that allocates donations to non-profit organizations. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 3365–3376.
- [12] Schneeberger T, Gebhard P, Baur T, et al. PARLEY: A transparent virtual social agent training interface. In: *Proceedings of the 24th International Conference on Intelligent User Interfaces*, Marina del Ray, CA, 2019, pp.35–36.
- [13] Sheh, Raymond Ka-Man. 2017. "Why Did You Do That?" Explainable Intelligent Robots. In *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*.
- [14] Naiseh, M., Soorati, M.D. and Ramchurn, S., 2022, November. Outlining the Design Space of eXplainable Swarm (xSwarm): Experts' Perspective. In *International Symposium on Distributed Autonomous Robotic Systems* (pp. 28-41). Cham: Springer Nature Switzerland.
- [15] Mirchi, Nykan, Vincent Bissonnette, Recai Yilmaz, Nicole Ledwos, Alexander Winkler-Schwartz, and Rolando F. Del Maestro. 2020. The Virtual Operative Assistant: An explainable artificial intelligence tool for simulation-based training in surgery and medicine. *PloS one* 15, no. 2 (2020): e0229596.
- [16] Simkute, Auste, Aditi Surana, Ewa Luger, Michael Evans, and Rhianne Jones. XAI for learning: Narrowing down the digital divide between “new” and “old” experts. In *Adjunct Proceedings of the 2022 Nordic Human-Computer Interaction Conference*, 1-6.
- [17] Naiseh, Mohammad, Reem S. Al-Mansoori, Dena Al-Thani, Nan Jiang, and Raian Ali. 2021b. Nudging through Friction: an Approach for Calibrating Trust in Explainable AI. In *2021b 8th International Conference on Behavioral and Social Computing (BESC)*, pp. 1-5. IEEE, 2021.
- [18] Narayanan, M., Chen, E., He, J., Kim, B., Gershman, S., Doshi-Velez, F.: How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1802.00682* (2018)
- [19] Oinas-Kukkonen, H. and Harjumaa, M., 2008, June. A systematic framework for designing and evaluating persuasive systems. In *International conference on persuasive technology* (pp. 164-176). Springer, Berlin, Heidelberg.
- [20] 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.
- [21] Calvaresi, D., Ciatto, G., Najjar, A., Aydoğan, R., Van der Torre, L., Omicini, A. and Schumacher, M., 2021, May. Expectation: personalized explainable artificial intelligence for decentralized agents with heterogeneous knowledge. In *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems* (pp. 331-343). Cham: Springer International Publishing.
- [22] 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.
- [23] Gao, Y., Gu, S., Jiang, J., Hong, S.R., Yu, D. and Zhao, L., 2022. Going beyond xai: A systematic survey for explanation-guided learning. *ACM Computing Surveys*.
- [24] Harjumaa, M. and Oinas-Kukkonen, H., 2007. An analysis of the persuasiveness of smoking cessation websites. In *Proceedings of the Second International Symposium on Medical Information and Communication Technology*. CWC, University of Oulu Oulu.
- [25] Sokol, K. and Flach, P., 2020. One explanation does not fit all. *KI-Künstliche Intelligenz*, pp.1-16.
- [26] Naiseh, M., Al-Thani, D., Jiang, N. and Ali, R., 2021. Explainable recommendation: when design meets trust calibration. *World Wide Web*, 24(5), pp.1857-1884.
- [27] Te'eni, D. and Sani-Kuperberg, Z., 2005. Levels of abstraction in designs of human–computer interaction: The case of e-mail. *Computers in human behavior*, 21(5), pp.817-830.
- [28] Kolb, D.A., 2014. *Experiential learning: Experience as the source of learning and development*. FT press.
- [29] Walster, E., Aronson, E. and Abrahams, D., 1966. On increasing the persuasiveness of a low prestige communicator. *Journal of Experimental Social Psychology*, 2(4), pp.325-342.
- [30] Wijekoon, A., Corsar, D., Wiratunga, N., Martin, K. and Salimi, P., 2024. Tell me more: Intent Fulfilment Framework for Enhancing User Experiences in Conversational XAI. *arXiv preprint arXiv:2405.10446*.
- [31] Cowan, D.D. and Lucena, C.J.P.D., 1995. Abstract data views: An interface specification concept to enhance design for reuse. *IEEE Transactions on software engineering*, 21(3), pp.229-243.

- [32] Naiseh, M., 2021. C-XAI: Design Method for Explainable AI Interfaces to Enhance Trust Calibration (Doctoral dissertation, Bournemouth University).
- [33] Tim Miller. 2018. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* (2018).
- [34] 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.
- [35] Bussone, Adrian, Simone Stumpf, and Dymrna O'Sullivan. "The role of explanations on trust and reliance in clinical decision support systems." In *2015 international conference on healthcare informatics*, pp. 160-169. IEEE, 2015.
- [36] De-Arteaga, Maria, Riccardo Fogliato, and Alexandra Chouldechova. "A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores." In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1-12. 2020.
- [37] Mirchi, N., Bissonnette, V., Yilmaz, R., Ledwos, N., Winkler-Schwartz, A., & Del Maestro, R. F. (2020). The Virtual Operative Assistant: An explainable artificial intelligence tool for simulation-based training in surgery and medicine. *PloS one*, 15(2), e0229596.
- [38] Montag, C., Nakov, P., & Ali, R. (2024). Considering the IMPACT framework to understand the AI-well-being-complex from an interdisciplinary perspective. *Telematics and Informatics Reports*, 13, 100112.
- [39] Montag, C., Ali, R., Al-Thani, D., & Hall, B. J. (2024). On artificial intelligence and global mental health. *Asian Journal of Psychiatry*, 91, 103855.