HHAI 2024: Hybrid Human AI Systems for the Social Good
F. Lorig et al. (Eds.)

© 2024 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA240211

Developing Meaningful Explanations for Machine Learning Models in the Telecom Domain

Henry Maathuis
University of Applied Sciences Utrecht
ORCiD ID: Henry Maathuis https://orcid.org/0009-0002-5542-0478.

Abstract. This study aims to develop and assess Explainable AI (XAI) tools tailored for internal telecom end-users. It focuses on delivering meaningful explanations informed by design principles, cognitive biases, and human decision-making theories. The research explores customizing XAI for telecom use-cases to support internal decision-making, while considering user preferences captured through elicitation studies. As part of this PhD study, a prescriptive framework will integrate cognitive biases, design principles, and human decision-making theory to effectively communicate AI explanations to end-users. User studies will be conducted to evaluate the effectiveness of the prototypes following from the framework.

Keywords. Explainable AI (XAI), Telecom, Meaningful Explanations, Cognitive Biases, User Interface Design, Decision-making support

1. Context

Over the years, powerful Machine Learning models, particularly non-linear Deep Learning (DL) models have emerged and have proven to be highly effective in various real-world applications [1]. One of the key reasons for their success is that they can capture complex relationships and patterns that simpler, linear models such as Logistic or Linear regression cannot [2,3,4].

Unfortunately, the increased complexity of DL models comes at the expense of explainability and interpretability [5] due to the many parameters and the non-linear nature of these models. This is particularly problematic in fields such as healthcare, or finance, where incorrect decisions, potentially due to the lack of transparency and interpretability of such models, can have significant impacts on individuals. Similarly, in the telecom domain, inaccurate decisions can lead to widespread service disruptions, imposing costs on service providers while simultaneously impacting customers. Hence, the ability to explain a model's decision is crucial in such domains [6].

Despite this limitation, the performance gains of DL models in many real-world applications cannot be denied, and they continue to outperform the simpler, more interpretable models in many cases. In problematic domains where performance is the primary concern, deploying such high-performance models should be approached with caution [7].

As such, a rapidly evolving field of research has emerged coined Explainable AI (XAI). One of the aims of XAI is to enable users of DL models to understand the reasoning behind the model's predictions. This can potentially lead to more trust in the model [8,9,10]. There are two main types of explainability tools for machine learning models: *model-agnostic* and *model-specific*. Model-agnostic tools like SHAP [11] and LIME [5] attempt to interpret the output of arbitrary models by identifying the importance of each feature in generating the output. They are also the most popular explainability tools in practice given their model-agnostic nature. On the other hand, model-specific methods offer a deeper understanding of how a particular model operates and the rationale behind its predictions. These explanations potentially leverage the internal workings of the model, such as its architecture, parameters, and decision-making processes, to provide insights into its behavior. Model-specific explanations can offer more detailed and precise explanations but are typically limited to understanding only the specific model (or class of models) being analyzed.

Unfortunately, most model-specific and model-agnostic explainability research is motivated by technical considerations, catering specifically to AI engineers. To ensure however that explanations generated by these tools are understandable and meaningful to a broader range of stakeholders, there is a pressing need to transition towards a more comprehensive Human-Computer Interaction (HCI) approach [13,14]. This shift acknowledges the importance of incorporating human-centered design principles into explainability efforts [15,16], thereby enhancing the usability and effectiveness of XAI tools. Meaningfulness is closely tied to language and communication, suggesting that the conveyed information should hold significance or purpose for the recipient. In HCI, the concept of meaningfulness varies among individuals based on factors like education, expertise, and contextual elements such as mood and task. Overall, the notion of what constitutes a meaningful explanation is subjective and contingent upon various individual and contextual factors [17].

Tailoring explanations is critical for generating meaningful explanations in XAI [18,19]. Research in this domain has found that tailored explanations lead to a higher degree of satisfaction and trust in the system than those obtained with generic explanations [20,21] and that providing personalized explanations to users leads to better performance on decision-making tasks [22]. Another study involves developing explanations that are specific to the user's domain knowledge and expertise. For example, a system may provide a detailed explanation to a domain expert but a more simplified explanation to a novice user. An argument is made that explanations that are non-tailored can result in algorithmic aversion and various biases [23].

Prior research discusses the influence of cognitive biases on XAI-assisted decision-making [24]. While this study outlines principles to mitigate certain biases and links them to technical explanation types like feature importances or counterfactuals, it does not address the connection to tailored explanations. Additionally, in [25], user-centric explanations are generated, which connect explanations with the dual process model of human reasoning [26]. However, the authors did not provide guidelines for designing user interfaces to assist users in better understanding the inner workings of AI models.

As part of my PhD studies, I aim to design concrete guidelines for designing meaningful human-computer interactions. To this end, we develop a framework that integrates theories on cognitive biases, design principles and decision-making to offer specific guidelines on how to effectively visualize and communicate explanations of AI

systems to end-users. The framework will be evaluated by designing interaction prototypes for several telecommunication use cases. User studies will be conducted to evaluate the effectiveness of the prototypes. Requirements elicitation interviews will be conducted to further steer the development of the prototypes and to tailor the explanations to the wishes and needs of the users of the system.

2. Research Questions

My research goal is to develop and evaluate XAI tools that are meaningful for internal users of AI systems within the telecom industry. A meaningful explanation considers the wishes and requirements of the users through requirements elicitation. Additionally, a systematic literature review will be conducted to unify design principles, cognitive biases and human decision-making yielding a framework for designing explanations for users. The framework is used as input for the development of the XAI solutions.

The main research question are as follows: *"How can Explainable AI be tailored for telecom use-cases to support internal decision-making while meeting user needs, and how can insights from design principles, cognitive biases, and human decision-making inform the generation of such explanations for stakeholders?"*

The following six research sub-questions are addressed in different research phases.

1. *Which cognitive biases hamper AI-supported human decision-making?*
2. *Which interface design opportunities enhance understandability of XAI explanations in the context of the selected cognitive biases?*
3. *How can theory from cognitive biases and user interface design be combined into a prescriptive framework for developing human-AI interactions for domain-expert decision-making?*
4. *How can human-AI interactions be optimized through requirements elicitation to ensure alignment with user needs?*
5. *What insights can be gained by applying XAI tools to specific use-cases within the telecom industry?*
6. *How can the insights gained from applying the XAI tools to specific use-cases in the telecom industry be used to further optimize and improve XAI tools?*

To the best of our current knowledge, there is no prescriptive framework available that integrates theories on cognitive biases, design principles and human-decision making to offer specific guidelines on how to effectively visualize and communicate explanations of AI systems to end-users.

3. Research Challenges

A significant portion of the PhD research is dedicated to developing a prescriptive framework for Explainable AI (XAI). One of the primary challenges lies in conducting a comprehensive systematic review to identify common patterns among various papers detailing XAI system designs. This entails navigating through a vast array of literature to distill key insights that can inform the development of the framework. Once the

framework is established, another challenge emerges in its application for developing multiple XAI solutions tailored to the telecom industry.

Furthermore, validating both the framework and the XAI solutions presents another hurdle, requiring a series of empirical studies to assess their impact on improving the understandability and usability of XAI explanations in real-world telecom scenarios. Extending the validation process to include potential use-cases from other sectors, such as finance or healthcare, adds complexity but also offers opportunities to demonstrate the framework's versatility and applicability across diverse domains. Overcoming these challenges will be essential to advancing the field of XAI and facilitating its practical implementation in various industries.

4. Method/Approach and Evaluation

Several research papers will be produced to address the research questions at hand. The first research paper which yields a prescriptive framework for developing meaningful human-AI interactions for domain-experts aims to answer the first three research sub-questions

To validate the framework, we assess its efficacy in real-world telecommunication use-cases. This assessment involves comparing prototypes developed from the framework with various baseline prototypes. Baseline prototypes may encompass screens displaying solely the model's outcome alongside a confidence score, as well as the presentation of raw, non-tailored technical explanations. The efficacy is measures specifically using both qualitative and quantitative measures. Amongst others, these could include performance on task, satisfaction of the explanation and trust in the explanation or system has a whole.

The first use-case focusses on a machine learning model designed to identify cable breakages and uncover their root causes by analyzing a vast array of alerts within a telecom network. This task places a significant emphasis on the operator's role, as they are tasked with interpreting these alerts to distinguish potential cable breakages from benign activities like power outages or false positive alerts. Ultimately, it falls upon the operator to determine whether a technician should be dispatched for maintenance. Historically, this process has burdened operators with the tedious task of sifting through numerous alarms to differentiate genuine faults from transient or non-essential alerts. While rule-based methods have offered some respite, their efficacy is limited by the intricate and variable nature of alarm patterns, necessitating continual updates to expert knowledge and manual intervention. To facilitate the operators, an XAI-system will be developed to assist users in their decision-making with the goal of minimizing falsely sending out mechanics based on incorrect model predictions.

Internal users of the AI-systems will be interviewed to gain a better understanding of their practice and what their requirements and wishes are from an automated decision support system.

Requirements elicitation studies will be conducted to answer the 4th research sub-question.

Together with the framework and the elicitation studies, prototypes will be co-created with the various stakeholders of the use-cases. The insights gained will be published to answer the 5th research sub-question.

The insights gained will be used as input for another case study and allow for fine-tuning the existing framework, answering the 6th research sub-question.

5. Discussion and Future Work

Table 1 outlines the project's overall timeline. The initial use-case has been selected, and the setup of the literature study is clearly outlined.

In 2024, the objective is to publish several papers, including a structured literature review presenting the framework. Additionally, two papers detailing technical XAI solutions for a specific use-case will be submitted. A requirements elicitation study will be conducted for the first use-case

Moving into 2025, the plan is to conduct a user evaluation study for the first use-case and continue work on the second use-case.

By 2026, we anticipate discussing the results with stakeholders and evaluating the performance of the prototype for the second use-case. A technical XAI solution paper will also be submitted.

In 2027, the focus shifts to finalizing work on the second prototype and publishing a user evaluation study for the second use-case. Additionally, the PhD thesis will be completed and defended.

Table 1. Timeline of the project

Year	Tasks	Results
2023	Literature Study Identify Use-Case 1	Select Use-Case (DONE) Setup Literature Study Design (DONE)
2024	Requirements Elicitation Study Use-Case 1 Low-Barrier Conference Paper Literature Study Develop Prototype Use-Case 1 Evaluate Prototype Use-Case 1 Discuss results with stakeholders and obtain insights	Overview requirements and needs users Use-Case 1 (March 2024) Submit Low-Barrier Conference Paper (April 2024) Submit Literature Study (July 2024) Submit technical paper regarding Use-Case 1 (December 2024)
2025	Develop Prototype Use-Case 2	Submit user evaluation study paper regarding Use-Case 1 (TBD)
2026	Evaluate Prototype Use-Case 2 Discuss results with stakeholders and obtain insights	Submit technical paper regarding Use-Case 2 (TBD)
2027	Round up work on second prototype Work on PhD Thesis Prepare PhD Defense	Submit user evaluation study paper regarding Use-Case 2 PhD Thesis (book)

Acknowledgements

My supervisory team comprises Dr. E. Postma (Tilburg University) as the doctoral advisor. Dr. D. Sent (Jheronimus Academy of Data Science: JADS) serves as my daily supervisor, and Dr. D. Kolkman (Utrecht University) is another member of my supervisory team. Additionally, for technical support in the development of machine learning models, I receive assistance from Dr. C. van Gemeren (University of Applied Sciences Utrecht), who is not formally part of the supervisory team.

My PhD position is research collaboration (ICAI Lab) between KPN which is a telecom provider in the Netherlands, JADS (Jheronimus Academy of Data Science) and Applied University of Utrecht, all of which play a funding role.

References

[1] LeCun Y, Bengio Y, Hinton G. Deep learning. nature. 2015 May 28;521(7553):436-44. doi:10.1038/nature14539.

[2] Goodfellow I, Bengio Y, Courville A. Deep learning. MIT press; 2016 Nov 10. doi: 0.1007/s10710-017-9314-z.

[3] Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE access. 2018 Sep 16;6:52138-60. doi: 10.1109/ACCESS.2018.2870052.

[4] Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, García S, Gil-López S, Molina D, Benjamins R, Chatila R. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information fusion. 2020 Jun 1;58:82-115. doi: 0.1016/j.inffus.2019.12.012.

[5] Ribeiro MT, Singh S, Guestrin C. " Why should i trust you?" Explaining the predictions of any classifier. InProceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining 2016 Aug 13 (pp. 1135-1144). doi: 10.48550/arXiv.1602.04938.

[6] Lipton ZC. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue. 2018 Jun 1;16(3):31-57. doi: 10.1145/3236386.3241340.

[7] Angelov PP, Soares EA, Jiang R, Arnold NI, Atkinson PM. Explainable artificial intelligence: an analytical review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2021 Sep;11(5):e1424. doi: 10.1002/widm.1424.

[8] Gade K, Geyik SC, Kenthapadi K, Mithal V, Taly A. Explainable AI in industry. InProceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining 2019 Jul 25 (pp. 3203-3204). doi: 10.1145/3366424.3383110.

[9] Shin D. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. International Journal of Human-Computer Studies. 2021 Feb 1;146:102551. doi: 10.1016/j.ijhcs.2020.102551.

[10] Došilović FK, Brčić M, Hlupić N. Explainable artificial intelligence: A survey. In2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO) 2018 May 21 (pp. 0210-0215). IEEE. doi: 10.23919/MIPRO.2018.8400040.

[11] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Advances in neural information processing systems. 2017;30. doi: 10.48550/arXiv.1705.07874.

[12] Ali S, Abuhmed T, El-Sappagh S, Muhammad K, Alonso-Moral JM, Confalonieri R, Guidotti R, Del Ser J, Díaz-Rodríguez N, Herrera F. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. Information fusion. 2023 Nov 1;99:101805. Doi: 10.1016/j.inffus.2023.101805.

[13] Abdul A, Vermeulen J, Wang D, Lim BY, Kankanhalli M. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. InProceedings of the 2018 CHI conference on human factors in computing systems 2018 Apr 21 (pp. 1-18). doi: 10.1145/3173574.3174156.

[14] Chromik M, Butz A. Human-XAI interaction: a review and design principles for explanation user interfaces. InHuman-Computer Interaction–INTERACT 2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30–September 3, 2021, Proceedings, Part II 18 2021 (pp. 619-640). Springer International Publishing. doi: 10.1007/978-3-030-85616-8_36.

[15] Nakao Y, Strappelli L, Stumpf S, Naseer A, Regoli D, Gamba GD. Towards responsible AI: A design space exploration of human-centered artificial intelligence user interfaces to investigate fairness. International Journal of Human–Computer Interaction. 2023 May 28;39(9):1762-88. doi: 10.1080/10447318.2022.2067936

[16] Schoonderwoerd TA, Jorritsma W, Neerincx MA, Van Den Bosch K. Human-centered XAI: Developing design patterns for explanations of clinical decision support systems. International Journal of Human-Computer Studies. 2021 Oct 1;154:102684. doi: 10.1016/j.ijhcs.2021.102684.

[17] Maxwell W, Dumas B. Meaningful XAI based on user-centric design methodology: Combining legal and human-computer interaction (HCI) approaches to achieve meaningful algorithmic explainability. Available at SSRN 4520754. 2023 Jul 1. doi: 10.2139/ssrn.4520754.

[18] Pedreschi D, Giannotti F, Guidotti R, Monreale A, Ruggieri S, Turini F. Meaningful explanations of black box AI decision systems. InProceedings of the AAAI conference on artificial intelligence 2019 Jul 17 (Vol. 33, No. 01, pp. 9780-9784). doi: 10.1609/aaai.v33i01.33019780.

[19] BERG MV, Kuiper O, HAAS YV, Gerlings J, Sent D, Leijnen S. A Conceptual Model for Implementing Explainable AI by Design: Results of an Empirical Study. InHHAI 2023: Augmenting Human Intellect: Proceedings of the Second International Conference on Hybrid Human-Artificial Intelligence 2023 Jul 7 (Vol. 368, p. 60). IOS Press. doi: 10.3233/FAIA230075.

[20] Dzindolet MT, Peterson SA, Pomranky RA, Pierce LG, Beck HP. The role of trust in automation reliance. International journal of human-computer studies. 2003 Jun 1;58(6):697-718. doi: 10.1016/S1071-5819(03)00038-7.

[21] Markus AF, Kors JA, Rijnbeek PR. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. Journal of biomedical informatics. 2021 Jan 1;113:103655. doi: 10.1016/j.jbi.2020.103655.

[22] Hoffman RR, Mueller ST, Klein G, Litman J. Metrics for explainable AI: Challenges and prospects. arXiv preprint arXiv:1812.04608. 2018 Dec 11. doi: 10.48550/arXiv.1812.04608.

[23] Simkute A, Surana A, Luger E, Evans M, Jones R. XAI for learning: Narrowing down the digital divide between “new” and “old” experts. InAdjunct Proceedings of the 2022 Nordic Human-Computer Interaction Conference 2022 Oct 8 (pp. 1-6). doi: 10.1145/3547522.3547678.

[24] Bertrand A, Belloum R, Eagan JR, Maxwell W. How cognitive biases affect XAI-assisted decision-making: A systematic review. InProceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society 2022 Jul 26 (pp. 78-91). doi: 10.1145/3514094.3534164.

[25] Wang D, Yang Q, Abdul A, Lim BY. Designing theory-driven user-centric explainable AI. InProceedings of the 2019 CHI conference on human factors in computing systems 2019 May 2 (pp. 1-15). doi: 10.1145/3290605.3300831.

[26] Kahneman D. Thinking, fast and slow. New York: Macmillan 2011. doi: 10.1007/s00362-013-0533-y.