The Role of Explanations of AI Systems: Beyond Trust and Helping to Form Mental Models

Milda Norkute
milda.norkute@thomsonreuters.com
Thomson Reuters
Zug, Zug, Switzerland

ABSTRACT

This paper discusses research that explored different roles for explanations of AI systems. A lot of the research focuses on investigating the role of explanations in mediating the level of users' trust in the AI system and helping them form correct mental models about it. This paper argues that more research should be dedicated to investigate the alternative roles that explanations could play in supporting the user's interactions with AI systems such as helping them enrich the AI suggestions they are presented with or correct them, help users do tasks more efficiently and effectively.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**.

KEYWORDS

Explainable artificial intelligence, interpretable machine learning

1 INTRODUCTION

Because of recent advances in artificial intelligence (AI) and machine learning (ML), AI solutions are being created and increasingly integrated into various technology solutions across many different sectors. AI methods are solving increasingly complex computational tasks, making them more important to the future of our society than ever before [6]. However, when decisions made by AI systems have an effect on lives of humans, for instance in housing, law or medicine, we have a need to understand how the decisions by AI systems were made [14]. It is often stated that the decisions made by AI systems should be explainable in the AI principles of the organizations that create AI products [9]. Furthermore, the European Union approved a data protection law known as the General Data Protection Regulation or "GDPR" [8] which also includes a "right to explanation" in 2016. Therefore, AI practitioners have to look for concrete ways and methods to explain the decisions made by their AI models.

To make this easier, efforts are being made to systematize different explanation methods and strategies. For example, some attempts have been made to categorize explanations by model types (logistic/linear regression, decision tree, etc.) or explainability categories (explanation by simplification, feature relevance explanation, visual explanation etc) [3]. Generally, it can be said that there are two main approaches to developing interpretable models. One approach is to create simple, clear models instead of black-box systems such as point systems [16] or generalized additive models that help with visualizing the impact of each feature on the predictions of the model [5] [11]. The second approach is to provide post-hoc explanations for potentially complex, black-box models [10]. The position of this paper is that regardless of which explainability approach is pursued, the core focus when choosing what explanations to use should be what role these explanations are supposed to play in supporting the AI system's interactions with users. Recent research looking into the role and value of explanations will be discussed and additional direction will be proposed. This paper is focusing on the explanations for the audience of users of the model.

2 DIFFERENT ROLES OF EXPLANATIONS

2.1 Trust and Mental Models

One of the key roles for the explanations is to help users trust the AI suggestions. For example, Bansal et al. 2021 [2] observed complementary improvements from AI augmentation in their experiments, however, they were not increased by explanations. Rather, explanations increased the chance that humans will accept the AI's recommendation, regardless of its correctness. This points towards the important role explanations have in increasing user's trust in the model. The authors also suggest that explanations should be informative, instead of convincing to follow the recommendation.

In addition to mediating trust, explanations have a role to play in helping users form mental models about the system. Lu and Yin (2021) [12] found that the level of agreement between people and the model on decision-making tasks that people have high confidence in, significantly affects reliance on the model if people receive no information about the model's performance. They also found that people have a tendency to over-generalize the performance of a model which is either observed or estimated by themselves. Therefore, this would suggest that having a mental model of the AI system is not only important for trust, but that without explanations it is difficult to understand what the AI system can really do.

Anik and Bunt (2021) [1] explored the concept of data-centric explanations where the explanations describe the training data to end-users. They investigated the potential utility of such approach, including the information about training data that participants find most compelling. They also investigated reactions to explanations across four different system scenarios. They found that participants' trust in AI system was impacted positively when the training data seemed balanced and negatively when the explanations revealed problems. Like prior work, they found that participants cared most

Mensch und Computer 2021, Workshopband, Workshop on User-Centered Artificial Intelligence (UCAI '21)

https://doi.org/10.18420/muc2021-mci-ws02-387

about the explanations for high-stakes system scenarios. According to authors, data-centric explanations have the potential to impact not only how users judge the trustworthiness of a system - when to trust it and rely on it, but also to assist users in assessing its fairness.

2.2 Beyond Trust and Mental Models

The role of explainability mechanisms can indeed go beyond increasing users' trust in the model or helping him understand how the model works. Additional use cases for explanations have been explored in the context of single user recommender systems. Potential goals of explanations in addition to trust (increasing user's confidence in the recommender system) have been identified as efficiency (reducing the time needed to complete a task), persuasiveness (using explanations to change a user's choice), effectiveness (helping the user to make higher-quality decisions), transparency (why an item has been recommended), scrutability (providing ways to help make the profile of the user possible to manage), satisfaction (explanations focused on aspects of enjoyment and usability), and credibility (assessed likelihood that a recommendation is accurate) [7] [17] [4].

Such efforts would also be welcome in other areas of AI such as Natural Language Processing (NLP). Natural Language processing is considered a difficult problem in computer science - although humans can easily master a language, the ambiguity and imprecise characteristics of the natural languages are what make NLP difficult for machines to implement. As a result, a lot of research focuses on increasing the performance of various language models for different NLP tasks and explainability is rarely considered when developing them [15]. Norkute et al. (2021) [13] made an attempt to explore the effects and impact of different explainability methods for abstractive summarization. The goal was to show the summary reviewers where the summary originated from by highlighting portions of the source text document. The first explainability method created text highlights based on attention weights from the Deep Learning (DL) model, a Pointer Generator network, built as a legal text summarization solution. The second explainability method, named source attribution was a model-agnostic formula that compares the source text and summary to identify overlapping language. The study found that participants were significantly faster in reviewing the summaries generated by the model when working with highlights based on attention scores from the DL model, but not with highlights based on a source attribution method. The participants did report an increased trust in the DL model because of the highlights based on attention scores. The participants also expressed a specific preference for the attention highlights. This was because the attention highlights had more use cases. The highlights based on the source attribution approach were only useful in pointing the participants towards the area of the document were the details relevant to the summary might be. This also was possible with attention highlights. In addition to this, the participants said they were able to use the highlights based on attention scores to enrich the machine-generated summary. They even helped the participants realize the summary was wrong in some cases. These findings further support the suggestion that in addition to helping users trust the model, explanations can offer additional support to the users. These learnings also indicate that researching the different roles explainability mechanisms can play in NLP as well as other AI areas and tasks is worthwhile.

3 CONCLUSION

The studies discussed explored different types and roles of explanations. While most research typically looks at the the role of explanations in mediating the level of trust in the AI system and helping them form correct mental models about the AI system, there may be a bigger role that the explanations can play. This includes helping users correct the mistakes of AI system or enriching its decision as well as helping users do tasks more efficiently and effectively. Future research should investigate what additional roles explanations can play in supporting users interactions with the AI systems for different AI tasks and areas, possibly even as features that have standalone value to the user. This could be done by designing studies where users are presented with explainability features only and not the AI suggestions - although such experiments of course would only make sense for specific contexts and use cases such as as summarization where users could be exposed to the text highlights showing where the summary came from and not the summary. Furthermore, it could be worthwhile to consider explainability aspects while developing the AI models instead of looking for ways to explain them using post-hoc methods after they have been developed - expanding the availability of different explainability methods would also help to widen the spectrum of how explanations can be used to support users interacting with the AI models.

REFERENCES

[1] Ariful Islam Anik and Andrea Bunt. 2021. Data-Centric Explanations: Explaining Training Data of Machine Learning Systems to Promote Transparency. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 75, 13 pages. https://doi.org/10.1145/3411764.3445736

[2] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. *Does the Whole Exceed Its Parts? The Effect of AI Explanations on Complementary Team Performance.* Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411764.3445717

[3] Vaishak Belle and Ioannis Papantonis. 2020. Principles and Practice of Explainable Machine Learning. *CoRR* abs/2009.11698 (2020). arXiv:2009.11698 https://arxiv.org/abs/2009.11698

[4] Mustafa Bilgic and Raymond Mooney. 2005. Explaining Recommendations: Satisfaction vs. Promotion.

[5] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. *Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-Day Readmission.* Association for Computing Machinery, New York, NY, USA, 1721–1730. https://doi.org/10.1145/2783258.2788613

[6] Dominik Dellermann, Philipp Ebel, Matthias Söllner, and Jan Marco Leimeister. 2019. Hybrid Intelligence. *Business Information Systems Engineering* 61, 5 (Mar 2019), 637–643. https://doi.org/10.1007/s12599-019-00595-2

[7] A. Felfernig, N. Tintarev, T. N. T. Trang, and M. Stettinger. 2021. Designing Explanations for Group Recommender Systems. arXiv:2102.12413 [cs.IR]

[8] Bryce Goodman and Seth Flaxman. 2017. European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation". *AI Magazine* 38, 3 (Oct. 2017), 50–57. https://doi.org/10.1609/aimag.v38i3.2741

[9] Lambert Hogenhout. 2021. A Framework for Ethical AI at the United Nations. *CoRR* abs/2104.12547 (2021). arXiv:2104.12547 https://arxiv.org/abs/2104.12547

[10] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. 2019. Faithful and Customizable Explanations of Black Box Models. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (Honolulu, HI, USA) *(AIES '19)*. Association for Computing Machinery, New York, NY, USA, 131–138. https://doi.org/10.1145/3306618.3314229

[11] Yin Lou, Rich Caruana, and Johannes Gehrke. 2012. Intelligible Models for Classification and Regression. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Beijing, China) *(KDD '12)*. Association for Computing Machinery, New York, NY, USA, 150–158. https://doi.org/10.1145/2339530.2339556

[12] Zhuoran Lu and Ming Yin. 2021. Human Reliance on Machine Learning Models When Performance Feedback is Limited: Heuristics and Risks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 78, 16 pages. https://doi.org/10.1145/3411764.3445562

[13] Milda Norkute, Nadja Herger, Leszek Michalak, Andrew Mulder, and Sally Gao. 2021. Towards Explainable AI: Assessing the Usefulness and Impact of Added Explainability Features in Legal Document Summarization. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 53, 7 pages. https://doi.org/10.1145/3411763.3443441

[14] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as Mechanisms for Supporting Algorithmic Transparency. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173677

[15] Hua Shen and Ting-Hao 'Kenneth' Huang. 2021. Explaining the Road Not Taken. arXiv:2103.14973 [cs.CL]

[16] Berk Ustun and Cynthia Rudin. 2015. Supersparse linear integer models for optimized medical scoring systems. *Machine Learning* 102, 3 (Nov 2015), 349–391. https://doi.org/10.1007/s10994-015-5528-6

[17] K. Verbert, D. Parra, P. Brusilovsky, and E. Duval. 2013. Visualizing recommendations to support exploration, transparency and controllability. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces (IUI'13, Santa Monica CA, USA, March 19-22, 2013))*. Association for Computing Machinery, Inc, United States, 351–362. https://doi.org/10.1145/2449396.2449442 18th International Conference on Intelligent User Interfaces (IUI 2013), IUI 2013 ; Conference date: 19-03-2013 Through 22-03-2013.