AI Explainability: Why One Explanation Cannot Fit All

Milda Norkute
Thomson Reuters Labs
Zug, 6300, Switzerland
milda.norkute@thomsonreuters.com

CHI 2021 Extended Abstracts, May 8-13, 2021, Yokohoma, Japan.

ACM ISBN 978-1-4503-6819-3/20/04.
DOI: https://doi.org/10.1145/3334480.

Abstract
This paper describes the challenges by faced by practitioners in selecting explainability methods for Artificial Intelligence (AI) models. Research into explainability of Natural Language Processing (NLP) models with attention mechanisms is discussed to illustrate how the value of the same explanation method depends on not only the audience but also the context and the use case for the explanation. It is proposed that when designing explanations, we should research what users of the model will use them for, so that they can be designed to task.

Author Keywords
Explainable artificial intelligence, interpretable machine learning

CSS Concepts
• Human-centered computing~Human computer interaction (HCI); **• Computing methodologies~Artificial Intelligence**

Introduction
Thanks to recent advances in artificial intelligence (AI) and machine learning (ML), AI solutions are being built and integrated into many technology solutions across various sectors. AI methods are achieving unprecedented performance levels when solving

increasingly complex computational tasks, making them more important to the future of our society than ever before [10]. However, when decisions made by AI systems have an effect on lives of humans (for instance in law, finance or medicine), there is a need to understand how the decisions by AI systems were made [8]. It is often stated that the decisions made by AI systems should be explainable in the AI principles of the organizations that build AI products [6, 9]. Furthermore, the European Union introduced a data protection law known as the General Data Protection Regulation or "GDPR" [4] which also includes a "right to explanation" in 2016. This means that AI practitioners need to find 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.), explainability categories (explanation by simplification, feature relevance explanation, visual explanation etc.) [2]. There are also some established techniques such as LIME, SHAP, ICE, etc. [2].

This paper argues that these explanation methods must be studied from various audience perspectives. Furthermore, the explanation methods must be investigated within the context of the specific task where they are used, because the same explanation method used in one context may have different use cases and role in the user's workflow in another context. Thus, when evaluating the explanations, we must begin by researching what users will use them for. This is illustrated by discussing research where, the same explainability technique was found to be quite valuable in one use case but of little value in another.

Overview of XAI audiences

The purpose of explainability in AI models can vary greatly based on the audience. Barredo Arrieta [1] identifies five main audience types: domain experts and users of the model, interacting with its outputs directly, users affected by model's decisions, regulatory entities, creators of the model – data scientists, product owners and others, managers and executive board members. For example, the purpose of having explainability for the users of the model is to trust the model, while users affected by model decisions could benefit from explainability by understanding their situation better, verify whether the decisions were fair. Since these audiences have different goals, this means that an explanation that may be considered to be good by one type of audience but not another. However, further to this, we will see that one explanation for the same type of audience (users of the model) may also be of different value if it is used in a different context and has another purpose.

Is attention an explanation?

I take a look at Natural Language Processing (NLP) models with attention mechanisms to illustrate how the same explanation may be of different value.

Jain and Wallace [5] explored whether a relationship exists between attention weights and model outputs. In their work, they performed extensive experiments across a variety of NLP tasks including binary classification, question answering and natural language inference assessing the degree to which attention weights provide meaningful explanations for predictions

and found that they did not. They found that correlation between intuitive feature importance measures including gradient and feature erasure approaches and learned attention weights was weak for recurrent encoders. According to the authors, although attention modules yield improved performance on NLP tasks, their ability to provide explanations for predictions of the model is questionable. According to them, this is especially true when a complex encoder is used, as it may entangle inputs in the hidden space. Presenting heatmaps of attention weights could seem to suggest a story about how a model arrived at a particular decision, but the relationship between this and attention is not always obvious. The authors, therefore concluded that standard attention modules do not provide meaningful explanations and should not be treated as though they do. From the perspective of the data scientist audience, attention mechanisms cannot be used reliably as explainability mechanism.

Norkute [7] tested an explainability method based on attention weights from the Deep Learning (DL) model, a Pointer Generator network, built as a legal text summarization solution. It was tested from the perspective of users of the model audience. Highlights which aimed to show the reviewers where the summary originated from by highlighting portions of the source text document were created based on attention scores from the DL model. Another explainability method, named source attribution, which is a model-agnostic formula that compares the source text and summary to identify overlapping language, was tested as well. 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 also reported increased trust in the DL model and expressed a preference for the attention highlights over the other type of 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. However, in addition to this, the participants were able to use the highlights based on attention scores to enrich the machine-generated summary as well as help realize the summary was wrong in some cases. Therefore, attention mechanisms were proven to be a useful explainability mechanism for the users of the model audience in this specific abstractive summarization use case. Furthermore, one of the reasons why highlights based on attention scores were preferred over the other explainability method was that they had more use cases and therefore were more useful to users.

Meanwhile Branting [3] also tested two approaches to a form of legal decision support one of which used an attention network for prediction and attention weights to highlight salient case text. Participants were randomly assigned to four conditions: case text only, case text plus highlights, case text plus negative and positive precedents, case text plus negative and positive precedents and highlighting. This approach was shown to be capable of predicting decisions, but attention-weight-based text highlighting did not demonstrably improve human decision speed or accuracy. The experiment participants had commented that they struggled to understand the connection between the highlighting text and the issue they were

supposed to decide. The presence of precedents did help to decide the case correctly, even if it took them slightly more time. The paper states that highlighting produced by another predictive model might be more useful for decision support in such case. Since the explainablity methods used by Branting and Norkute were essentially the same but the perceived usefulness by the same type of audience – users of the model, was different, this suggests that the evaluation results of specific explainability methods are influenced by the specific use case for the method. The two studies had different roles for the explanations - although both methods were used legal contexts, one method was intended to help review verify the summaries, while the other was intended to facilitate decision making. Therefore, when selecting suitable explainability methods, we should research what users will use explanations for, as they then can be designed to task.

Discussion and Conclusion

The research studies discussed illustrate that the evaluation of the usefulness of the explainability method cannot be detached from the specific context and its specific intended use cases in that context. The explainability method that is perceived as useful by the same audience - users of the model in one context for one task might be perceived as not useful in another, even if the AI model and the explanation used are essentially the same. Therefore, future research studying the usefulness of various explainability methods should be careful when trying to generalize the findings. Further to this, the evaluation and selection the process of explanations should always begin with an exploration of what users will be using the explanations for specifically.

References

[1] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier del Ser, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion 58: 82–115. arXiv:1910.10045. Retrieved from: https://arxiv.org/abs/1910.10045

[2] Vaishak Belle and Ioannis Papantonis. 2020. Principles and Practice of Explainable Machine Learning. arXiv:2009.11698. Retrieved from: https://arxiv.org/abs/2009.11698

[3] K. Branting, B. Weiss, B. Brown, et al. 2019. Semi-supervised methods for explainable Legal prediction. *Proceedings of the 17th International Conference on Artificial Intelligence and Law, ICAIL 2019*, Association for Computing Machinery, Inc, 22–31.

[4] Bryce Goodman and Seth Flaxman. 2016. European Union regulations on algorithmic decision-making and a "right to explanation." DOI:https://doi.org/10.1609/aimag.v38i3.2741

[5] Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. arXiv:1902.10186. Retrieved from: https://arxiv.org/abs/1902.10186

[6] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. Nature Machine Intelligence 1, 9: 389–399. DOI:https://doi.org/10.1038/s42256-019-0088-2

[7] Milda Norkute, Nadja Herger, Leszek Michalak, Andrew Mulder, Sally Gao. 2021. Towards Explainable AI: Assessing the Usefulness and Impact of Added Explainability Features in Legal Document Summarization. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts* (CHI'21 Extended Abstracts), May 8–13, 2021, Yokohama, Japan. ACM, New York, NY, USA, 10 pages. DOI: https://doi.org/10.1145/3411763.3443441

[8] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as mechanisms for supporting algorithmic transparency. *Conference on Human Factors in Computing Systems - Proceedings*, Association for Computing Machinery. DOI: https://doi.org/10.1145/3173574.3173677

[9] Thomson Reuters. Artificial Intelligence at Thomson Reuters. Retrieved September 20, 2020 from https://www.thomsonreuters.com/en/artificial-intelligence/introduction-to-artificial-intelligence-at-thomson-reuters.html

[10] 2016. *The AI Now Report The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term*. Retrieved from: http://acikistihbarat.com/Dosyalar/AINowSummaryReport-artificial-intelligence-effects-in-near-future.pdf