Argument Attribution Explanations in Quantitative Bipolar Argumentation Frameworks

Xiang Yin[1] , **Nico Potyka**[2,1] , **Francesca Toni**[1]
[1]Department of Computing, Imperial College London, UK
[2]School of Computer Science and Informatics, Cardiff University, UK
{xy620, ft}@imperial.ac.uk, potykan@cardiff.ac.uk

1 Introduction

Explainable AI (XAI) is playing an increasingly important role in AI towards safety, reliability and trustworthiness (Adadi and Berrada 2018). Various methods have been proposed in this field for providing explanations for several AI algorithms, models, and systems (e.g. see recent overviews (Minh et al. 2022; Adadi and Berrada 2018)).

A popular category of explanation methods is *feature attribution*, aiming at assigning a "feature importance score" to each input feature fed to the AI of interest (notably machine learning models), denoting its contribution to the output decision by the AI. Explanations returned by feature attribution methods are intuitive in that they focus on explaining the outputs in terms of the inputs alone, making it unnecessary to go into the details of the inner mechanism of the underlying AI. Furthermore, feature attribution explanations are easy for people to understand by just checking the positive or negative influence of the input features towards the outputs and the ranking of the magnitude of the scores.

Alongside feature attribution methods, in recent years *argumentative XAI* is increasingly showing benefits for various forms of AI (see (Čyras et al. 2021) for an overview). Basically, argumentative XAI applies computational argumentation (Atkinson et al. 2017) to extract *argumentation frameworks (AFs)* as skeletons for explanations. For example, (Čyras et al. 2019) uses abstract AFs (Dung 1995) to explain the outputs of schedulers, (Potyka 2021) proposes to use weighted bipolar AFs to explain multi-layer perceptrons and (Cocarascu, Rago, and Toni 2019) propose to use *Quantitative Bipolar AFs (QBAFs)* (Baroni et al. 2015) to explain movie review aggregations. Whereas feature attribution methods focus on the input-output behaviour of the underlying AI, AFs as explanations point to the dialectical relationships among arguments, abstractly representing interactions among the inner components of the underlying AI encoded by the AFs. These AFs provide a natural mechanism for users to interact with the AI (Rago et al. 2020) and may help find irrationalities in the underlying AI to aid debugging and improving the AI (Garcia, Rotstein, and Simari 2007).

Existing forms of argumentative XAI are predominantly *qualitative* in that they focus on explaining the reasoning outcomes of AFs with debates/disputes/dialogues in the spirit of *extension-based semantics* (e.g. as in (Dung 1995)). These qualitative explanations mirror interactions within the inner mechanism of the underpinning AI as dialectical exchanges between arguments. For example, (Čyras et al. 2019) use 'explanation via (non-)attacks' and (Cocarascu, Rago, and Toni 2019) define explanations as template-driven dialogues using attacks and supports in the AFs. Instead, explaining the *quantitative* reasoning outcomes of AFs under *gradual semantics* (e.g. those proposed in (Baroni et al. 2015; Potyka 2021)) has not received much attention, in spite of the widespread use of this form of semantics in several applications (e.g. fake news detection (Kotonya and Toni 2019), movie recommendations (Cocarascu, Rago, and Toni 2019) and fraud detection (Chi et al. 2021)). However, in many application settings, it is important to see how arguments in AFs topically influence one another, and how much positive/negative influence is transmitted from one argument to another. This is especially the case when explanations are needed for a *topic argument* of interest (e.g. an argument corresponding to the output of a classifier as in (Albini et al. 2020; Potyka 2021)) and it is essential to assess which arguments have more importance towards the topic argument.

In this paper, we contribute to filling this gap by proposing a novel theory of *Argument Attribution Explanations (AAEs)* by incorporating the spirit of feature attribution from machine learning in the context of QBAFs. With respect to qualitative explanations alone, AAEs allow to measure and compare the contribution of different arguments towards topic arguments in QBAFs under the Discontinuity Free Quantitative Argumentation Debate (*DF-QuAD*) gradual semantics (Rago et al. 2016), even when the comparison is difficult with qualitative explanations alone. Additionally, AAEs take the *base scores* of arguments in QBAFs into account. Different base scores should give rise to different explanations, but qualitative explanations disregard base scores, as they only consider the QBAFs' structure regardless of quantitative information.

2 Contribution

Overall, the contribution of this paper is threefold.

- We introduce a novel theory of AAEs by adapting the concept of feature attribution explanation from machine learning to QBAFs. This novel approach enables us to quantitatively determine the influence of individual argu-

ments on the overall outcome of another (topic) argument, thus providing a more interpretable explanation of the reasoning processes within QBAFs.

- Our work not only proposes the theory but also rigorously studies several desirable properties of AAEs. These properties include explainability, missingness, completeness, and counterfactuality, among others. Some of these properties are newly introduced, while others are adapted from existing literature to fit our argumentative context. This comprehensive analysis ensures that the AAEs are not only theoretically sound but also practically useful and reliable.
- To illustrate the practical utility of our proposed AAEs, we conduct two detailed case studies. The first case study applies AAEs to the scenario of fake news detection, demonstrating how our method can enhance the transparency and trustworthiness of AI systems in evaluating the credibility of news sources. The second case study explores the application of AAEs in movie recommender systems, showing how our approach can improve the interpretability of recommendations.

3 Relevance to KR

Our research focuses on the explainability of formal QBAFs, an area of great importance to both argumentation theory researchers and those working in XAI. Specifically, it targets researchers dedicated to improving the transparency of formal models using XAI techniques. By exploring the intricate mechanisms of QBAFs, our work seeks to bridge the gap between theoretical models and practical interpretability, offering a compelling and valuable resource for researchers in these intersecting areas.

4 Significance of Results

The significance of our work lies in its comprehensive enhancement of explainability in QBAFs, bridging theoretical and practical gaps in argumentative XAI. By introducing the novel theory of Argument Attribution Explanations (AAEs), we offer a comprehensive theory for quantitatively assessing argument influences, thereby improving interpretability. Our detailed analysis of AAEs' desirable properties ensures theoretical soundness and practical reliability. Demonstrating practical utility through case studies in fake news detection and movie recommender systems, our work fosters greater transparency, trust, and user confidence in AI-driven decision-making, making it a valuable resource for researchers and practitioners in argumentation theory and XAI.

Acknowledgments

This research was partially funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 101020934, ADIX) and by J.P. Morgan and by the Royal Academy of Engineering under the Research Chairs and Senior Research Fellowships scheme. Any views or opinions expressed herein are solely those of the authors.

References

Adadi, A., and Berrada, M. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access* 6:52138–52160.

Albini, E.; Lertvittayakumjorn, P.; Rago, A.; and Toni, F. 2020. Deep argumentative explanations. *arXiv:2012.05766*.

Atkinson, K.; Baroni, P.; Giacomin, M.; Hunter, A.; Prakken, H.; Reed, C.; Simari, G. R.; Thimm, M.; and Villata, S. 2017. Towards artificial argumentation. *AI Mag.* 38(3):25–36.

Baroni, P.; Romano, M.; Toni, F.; Aurisicchio, M.; and Bertanza, G. 2015. Automatic evaluation of design alternatives with quantitative argumentation. *Argument & Computation* 6(1):24–49.

Chi, H.; Lu, Y.; Liao, B.; Xu, L.; and Liu, Y. 2021. An optimized quantitative argumentation debate model for fraud detection in e-commerce transactions. *IEEE Intelligent Systems* 36(2):52–63.

Cocarascu, O.; Rago, A.; and Toni, F. 2019. Extracting dialogical explanations for review aggregations with argumentative dialogical agents. In *18th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*.

Čyras, K.; Letsios, D.; Misener, R.; and Toni, F. 2019. Argumentation for explainable scheduling. In *33rd AAAI Conference on Artificial Intelligence*.

Čyras, K.; Rago, A.; Albini, E.; Baroni, P.; and Toni, F. 2021. Argumentative XAI: a survey. In *30th International Joint Conference on Artificial Intelligence (IJCAI)*.

Dung, P. M. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence* 77(2):321–358.

Garcia, A. J.; Rotstein, N. D.; and Simari, G. R. 2007. Dialectical explanations in defeasible argumentation. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, 295–307. Springer.

Kotonya, N., and Toni, F. 2019. Gradual argumentation evaluation for stance aggregation in automated fake news detection. In *6th Workshop on Argument Mining*.

Minh, D.; Wang, H. X.; Li, Y. F.; and Nguyen, T. N. 2022. Explainable artificial intelligence: a comprehensive review. *Artificial Intelligence Review* 55(5):3503–3568.

Potyka, N. 2021. Interpreting neural networks as quantitative argumentation frameworks. In *35th AAAI Conference on Artificial Intelligence*.

Rago, A.; Toni, F.; Aurisicchio, M.; and Baroni, P. 2016. Discontinuity-free decision support with quantitative argumentation debates. In *15th International Conference on the Principles of Knowledge Representation and Reasoning (KR)*.

Rago, A.; Cocarascu, O.; Bechlivanidis, C.; and Toni, F. 2020. Argumentation as a framework for interactive explanations for recommendations. In *17th International Conference on Principles of Knowledge Representation and Reasoning (KR)*.