Argumentative XAI: A Survey (Extended Abstract)

Kristijonas Čyras[1][0000-0002-4353-8121], Antonio Rago[2][0000-0001-5323-7739], Emanuele Albini[2][0000-0003-2964-4638], Pietro Baroni[3][0000-0001-5439-9561], and Francesca Toni[2][0000-0001-8194-1459]

[1] Ericsson Research, Sweden
[2] Department of Computing, Imperial College London, UK
[3] Dipartimento di Ingegneria dell'Informazione, Università degli Studi di Brescia, Italy
kristijonas.cyras@ericsson.com, {a.rago, emanuele, ft}@imperial.ac.uk, pietro.baroni@unibs.it

Abstract. Explainable AI (XAI) has been investigated for decades and, together with AI itself, has witnessed unprecedented growth in recent years. Among various approaches to XAI, argumentative models have been advocated in both the AI and social science literature, as their dialectical nature appears to match some basic desirable features of the explanation activity. In this survey we overview XAI approaches built using methods from the field of *computational argumentation*, leveraging its wide array of reasoning abstractions and explanation delivery methods. We overview the literature focusing on different types of explanation (intrinsic and post-hoc), different models with which argumentation-based explanations are deployed, different forms of delivery, and different argumentation frameworks they use. We also lay out a roadmap for future work. The full paper [14] can be found at ijcai.org/proceedings/2021/600.

Keywords: argumentation · explainable AI · survey

1 Introduction

Explainable AI (XAI) has attracted a great amount of attention in recent years, due mostly to its role in bridging applications of AI and humans who develop or use them. Approaches to support XAI have been proposed (see e.g. some recent overviews [1,20]) and the crucial role of XAI in human-machine settings has been emphasised [27]. Whereas some recent efforts are focused on explaining machine learning models [1], XAI has also been a recurrent concern in other AI settings, e.g. expert systems [32], answer set programming [18] and planning [9]. Amongst several approaches, *argumentative* explanations are advocated in the social sciences [3], focusing on the human perspective, and argumentation's potential advantages for XAI have been pointed out [25,7,30]. In [14] we provide a comprehensive survey of literature in XAI viewing explanations as argumentative (independently of the underlying methods to be explained). In this extended abstract we summarise the most salient points of this survey.

Many methods for generating explanations in XAI can be seen as argumentative. Indeed, attribution methods, including model-agnostic [23] and model-specific [29] approaches, link inputs to outputs via (weighted) positive and negative relations, and contrastive explanations identify reasons pro and con outputs [9,24]. In [14] we focus instead on overtly argumentative approaches, with an emphasis on the several existing XAI solutions using forms of *computational argumentation* (see [4] for a recent overview of this field and [14] for background).

The application of computational argumentation to XAI is supported by its strong theoretical and algorithmic foundations, and its flexibility particularly in the wide variety of *argumentation frameworks* (AFs) on offer. These AFs give ways to specify *arguments* and *dialectical relations* between them, as well as *semantics* to evaluate the dialectical *acceptability* or *strength* of arguments, while differing (sometimes substantially) in how they define these components. When AFs are used to obtain explanations, (weighted) arguments and dialectical relations may suitably represent anything from input data, e.g. categorical data or pixels in an image, to knowledge, e.g. rules, to components of the method being explained, e.g. filters in convolutional neural networks, to problem formalisations, e.g. planning, scheduling or decision making models, to outputs, e.g. classifications, recommendations, or logical inference. This flexibility and wide-ranging applicability has led to a multitude of methods for *AF-based explanations*, providing the motivation and need for the survey in [14]. Our contributions are:

- we overview the literature on AF-based explanations, cataloguing representative approaches according to what they explain and how (outlined in §2);
- we overview the prevalent forms which AF-based explanations take after being drawn from AFs (omitted here due to a lack of space);
- we lay out a roadmap for future work, covering: the need for properties of AF-based explanations, computational aspects, and further applications and other potential developments of AF-based explanations (summarised in §3).

We ignore argumentative explanations based on informal notions or models lacking aspects of AFs (notably, semantics) and application domains of argumentative XAI, covered in a recent, orthogonal survey [34].

2 Types of Argumentative Explanations

We review the literature for models explained using argumentative explanations *built from AFs*, referred to as *AF-based explanations*. We divide them into:

- *intrinsic*, i.e. defined for models that natively use argumentative techniques, an example of which is given in Figure 1;
- *post-hoc*, i.e. obtained from non-argumentative models; we further divide these explanations into those providing a *complete* or an *approximate* representation of the explained model, exemplified respectively in Figures 2 and 3.

Briefly, an intrinsic AF-based explanation provides details and reasons behind the working of an AI system that uses argumentation for reasoning. One advantage here are the well-known reasoning mechanisms of argumentation, but arguably not all AI tasks can be immediately addressed using argumentation. In

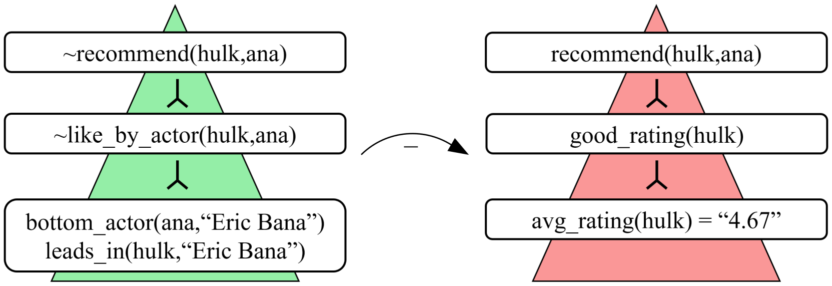

Fig. 1. Intrinsic AF-based explanation for the recommender system of [8] explaining why the movie *Hulk* was not recommended to the user *Ana*. The undefeated argument against this recommendation (left) attacks (labelled −) the argument for the recommendation (right), which is thus defeated. DeLP structured arguments are constructed of statements linked by defeasible rules (indicated by ⅄).

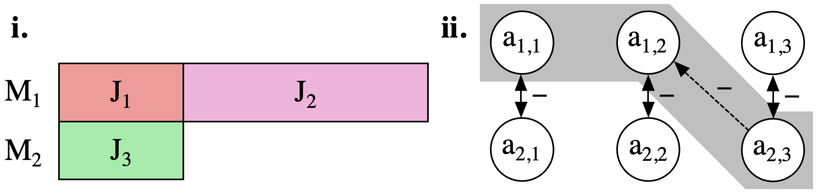

Fig. 2. Complete post-hoc AF-based explanations for makespan scheduling [13]. i. An inefficient schedule with jobs J_1, J_2 assigned to machine M_1 and J_3 to M_2 corresponds 1-1 with ii. the non-conflict-free extension (in grey) in the corresponding abstract argumentation framework where argument $a_{i,j}$ represents assignment of J_j to M_i and attacks capture scheduling constraints. The (dashed) attack underlies an explanation as to why the schedule is inefficient: J_2 and J_3 can be swapped between M_2 and M_1.

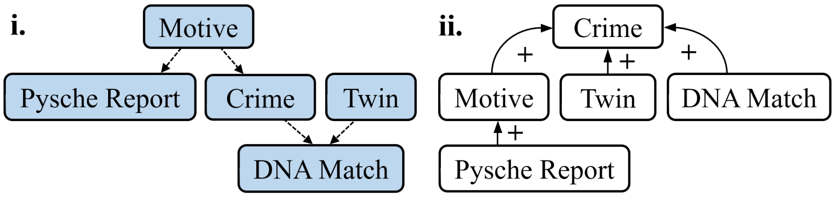

Fig. 3. Approximate post-hoc AF-based explanations for Bayesian networks [33]. i. Bayesian network with conditional dependencies (conditional probabilities ignored). ii. Extracted support argumentation framework, where the support relation is directly derived from each variable's Markov blanket.

lieu, a complete post-hoc AF-based explanation explains a model captured one-to-one argumentatively. This allows explanations to make use of the well-studied properties of AFs, also requiring the AFs to fully capture properties of the explained model. This requirement is relaxed in favour of providing an accessible conceptual representation when an approximate post-hoc AF-based explanation is instead used to explain a model captured only partially using argumentation.

This three-fold distinction among types of AF-based explanation is not crisp, and some of the approaches we survey may be deemed to be hybrids. We use the term ‘model’ in a very general sense, in the spirit of [19], to stand for a variety of systems, here amounting to these categories: recommender systems, classifiers and probabilistic methods, decision-making and knowledge-based systems, planners and schedulers, as well as tools for logic programming. Note also that our focus is on explaining models *other than* the argumentation process itself.

3 A Roadmap for Argumentative XAI

We identify some gaps in the state-of-the-art on argumentation-based XAI and discuss opportunities for further research, focusing on three avenues: studying theoretical *properties* and *computational aspects* of AF-based explanations, as well as broadening both applications and the scope of AF-based explanations.

Properties. AFs have been well studied regarding their properties, e.g. [16,5], but AF-based explanations less so. Notable exceptions include forms of *fidelity*, amounting to sound and complete mappings from systems being explained and the generated AF-based explanations [13,17], and properties of extension-based *explanation semantics* [22]. Other desirable properties from the broader XAI landscape [31] have been mostly neglected, though some user-acceptance aspects such as *cognitive tractability* [13] as well as *transparency* and *trust* [26] have been considered for AF-based explanations. For some properties, experiments with human users may be needed, as in much of the XAI literature [1], and creativity in the actual format of AF-based explanations shown to humans required.

Computational aspects. To effectively support XAI solutions, AF-based explanations need to be efficiently computable. This necessitates research on computational complexity and effective implementations. In the case of *intrinsic AF-based explanations*, this entails both good understanding of the complexity of, and building systems for, the relevant reasoning tasks: e.g. [12,10] rely upon the tractable membership reasoning task for the grounded extension for AA. In the case of *post-hoc* (*complete* or *approximate)* AF-based explanations, a further hurdle is the extraction of AFs from the models in need of explanation, prior to the extraction of the AF-based explanations themselves: the (complete) approach of [13] exemplifies a tractable such extraction. This line of research can benefit from various translation approaches developed in different areas of KR.

For all types of AF-based explanations, further consideration must be given to representational aspects of both reasoning and learning, and consequently to the extraction task of explanations of various formats from AFs. For illustration, the AF-based explanations for the approaches of [12,10] rely upon DTs that can be extracted efficiently from AFs, given the grounded extension. Further, [28] give complexity results for extracting certain sets of arguments as explanations in AA.

In general, however, computational issues in AF-based explanations require a more systematic investigation in terms of underpinning representation, reasoning tasks and their interplay with explanation, as well as explanation extraction.

Extending applications and the scope of AF-based explanations. While already having a variety of instantiations and covering a wide range of application contexts, AF-based explanations have a wide potential of further development.

Concerning applications, arguably the strongest demand for XAI solutions is currently driven by applications of machine learning (ML). In this context, it is interesting to note that in a loose sense some forms of ML have dialectical roots: supervised ML uses positive and negative examples of concepts to be learnt, and reinforcement learning uses positive and negative rewards. Further, several of the existing XAI solutions for ML, albeit not explicitly argumentative in the sense of this survey, are argumentative in spirit, as discussed in more detail in [14] (e.g. SHAP [23] can be seen as identifying reasons for and against outputs). However, AF-based explanations have been only sparingly deployed in ML-driven (classification and probabilistic) settings. Specifically, the analysis of dialectics is a crucial, yet often ignored, underpinning of XAI for ML. We envisage a fruitful interplay, where the explanation needs of ML, while benefiting from the potential of argumentation techniques, also stimulate further research in computational argumentation. Likewise for explainability in machine reasoning [11] areas such as planning, constraint and logic programming, which have already benefited from argumentative reasoning.

As a first step, it would be interesting to see whether existing approaches on logic-based explanations, either model-agnostic [21,15] or model-specific [29], could be understood as AF-based explanations, potentially relying upon existing logic-based AFs such as [6], or ADFs/AFs with structured arguments. Connections with the widely used *counterfactual explanations (CFs)* (e.g. see [31]) represent another stimulating investigation topic. CFs identify, as explanations for models' outputs, hypothetical changes in the inputs that would change these outputs. They thus show some dialectical flavour and call for the study of forms of AF-based explanations able to provide CF functionalities. For instance, *relation-based CFs* [2] may be interpreted as AF-based explanations for suitable AFs (with different types of support to match the underpinning relations). Given that CFs are based on 'changes', the corresponding form of AF-based explanations (discussed in detail in [14]) could also support this kind of development.

4 Conclusion

Argumentative XAI is an active research area, focusing on explanations built using computational argumentation. We have set out a roadmap for future developments of AF-based explanations and their use, which we hope will be beneficial to the KR research community at large.

Acknowledgements. This research was funded in part by the Royal Academy of Engineering, UK, and by J.P. Morgan.

References

1. Adadi, A., Berrada, M.: Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access **6**, 52138–52160 (2018). https://doi.org/10.1109/ACCESS.2018.2870052
2. Albini, E., Rago, A., Baroni, P., Toni, F.: Relation-based counterfactual explanations for bayesian network classifiers. In: Proceedings of the Twenty-Ninth Int. Joint Conf. on Artificial Intelligence, IJCAI 2020. pp. 451–457 (2020)
3. Antaki, C., Leudar, I.: Explaining in conversation: Towards an argument model. Europ. J. of Social Psychology **22**, 181—-194 (1992)
4. Baroni, P., Rago, A., Toni, F.: How many properties do we need for gradual argumentation? In: Proceedings of the Thirty-Second AAAI Conf. on Artificial Intelligence (2018)
5. Baroni, P., Rago, A., Toni, F.: From fine-grained properties to broad principles for gradual argumentation: A principled spectrum. Int. J. Approx. Reason. **105**, 252–286 (2019). https://doi.org/10.1016/j.ijar.2018.11.019
6. Besnard, P., Hunter, A.: A logic-based theory of deductive arguments. Artificial Intelligence **128**(1-2), 203–235 (2001)
7. Bex, F., Walton, D.: Combining Explanation and Argumentation in Dialogue. Argument & Computation **7**(1), 55–68 (2016). https://doi.org/10.3233/AAC-160001
8. Briguez, C.E., Budán, M.C., Deagustini, C.A.D., Maguitman, A.G., Capobianco, M., Simari, G.R.: Argument-based mixed recommenders and their application to movie suggestion. Exp. Sys. with Applications **41**(14), 6467–6482 (2014)
9. Chakraborti, T., Sreedharan, S., Kambhampati, S.: The Emerging Landscape of Explainable Automated Planning & Decision Making. In: Bessiere, C. (ed.) 29th International Joint Conference on Artificial Intelligence. pp. 4803–4811. IJCAI, Yokohama (2020). https://doi.org/10.24963/ijcai.2020/669
10. Cocarascu, O., Stylianou, A., Čyras, K., Toni, F.: Data-empowered argumentation for dialectically explainable predictions. In: ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020). pp. 2449–2456 (2020). https://doi.org/10.3233/FAIA200377
11. Čyras, K., Badrinath, R., Mohalik, S.K., Mujumdar, A., Nikou, A., Previti, A., Sundararajan, V., Feljan, A.V.: Machine Reasoning Explainability (2020), `http://arxiv.org/abs/2009.00418`
12. Čyras, K., Birch, D., Guo, Y., Toni, F., Dulay, R., Turvey, S., Greenberg, D., Hapuarachchi, T.: Explanations by arbitrated argumentative dispute. Expert Syst. Appl. **127**, 141–156 (2019). https://doi.org/10.1016/j.eswa.2019.03.012
13. Čyras, K., Letsios, D., Misener, R., Toni, F.: Argumentation for explainable scheduling. In: The Thirty-Third AAAI Conf. on Artificial Intelligence, AAAI 2019. pp. 2752–2759 (2019)
14. Čyras, K., Rago, A., Albini, E., Baroni, P., Toni, F.: Argumentative XAI: A Survey. In: Zhou, Z.H. (ed.) 30th International Joint Conference on Artificial Intelligence. pp. 4392–4399. IJCAI, Montreal (2021). https://doi.org/10.24963/ijcai.2021/600
15. Darwiche, A., Hirth, A.: On the reasons behind decisions. In: ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020). pp. 712–720 (2020). https://doi.org/10.3233/FAIA200158

16. Dung, P.M., Mancarella, P., Toni, F.: Computing Ideal Sceptical Argumentation. Artificial Intelligence **171**(10-15), 642–674 (2007). https://doi.org/10.1016/j.artint.2007.05.003
17. Fan, X.: On generating explainable plans with assumption-based argumentation. In: PRIMA 2018: Principles and Practice of Multi-Agent Systems - 21st International Conference, Tokyo, Japan, October 29 - November 2, 2018, Proceedings. pp. 344–361 (2018). https://doi.org/10.1007/978-3-030-03098-8_21
18. Fandinno, J., Schulz, C.: Answering the "Why" in Answer Set Programming - A Survey of Explanation Approaches. Theory and Practice of Logic Programming **19**(2), 114–203 (2019). https://doi.org/10.1017/S1471068418000534
19. Geffner, H.: Model-free, Model-based, and General Intelligence. In: Lang, J. (ed.) 27th International Joint Conference on Artificial Intelligence. pp. 10–17. IJCAI, Stockholm (2018), `http://arxiv.org/abs/1806.02308`
20. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM Comput. Surv. **51**(5), 93:1–93:42 (2019)
21. Ignatiev, A., Narodytska, N., Marques-Silva, J.: On relating explanations and adversarial examples. In: Advances in Neural Information Processing Systems 32: Annual Conf. on Neural Information Processing Systems 2019, NeurIPS 2019. pp. 15857–15867 (2019)
22. Liao, B., van der Torre, L.: Explanation semantics for abstract argumentation. In: Computational Models of Argument - Proceedings of COMMA 2020, Perugia, Italy, September 4-11, 2020. pp. 271–282 (2020). https://doi.org/10.3233/FAIA200511
23. Lundberg, S.M., Lee, S.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems 30: Annual Conf. on Neural Information Processing Systems 2017. pp. 4765–4774 (2017)
24. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence **267**, 1–38 (2019)
25. Moulin, B., Irandoust, H., Bélanger, M., Desbordes, G.: Explanation and Argumentation Capabilities: Towards the Creation of More Persuasive Agents. Artificial Intelligence Review **17**(3), 169–222 (2002). https://doi.org/10.1023/A:1015023512975
26. Rago, A., Cocarascu, O., Bechlivanidis, C., Toni, F.: Argumentation as a framework for interactive explanations for recommendations. In: Proceedings of the 17th International Conference on Principles of Knowledge Representation and Reasoning, KR 2020, Rhodes, Greece, September 12-18, 2020. pp. 805–815 (2020). https://doi.org/10.24963/kr.2020/83
27. Rosenfeld, A., Richardson, A.: Explainability in Human-Agent Systems. Autonomous Agents and Multi-Agent Systems pp. 1–33 (may 2019). https://doi.org/10.1007/s10458-019-09408-y
28. Saribatur, Z.G., Wallner, J.P., Woltran, S.: Explaining Non-Acceptability in Abstract Argumentation. In: Giacomo, G.D., Catalá, A., Dilkina, B., Milano, M., Barro, S., Bugarín, A., Lang, J. (eds.) 24th European Conference on Artificial Intelligence. pp. 881–888. IOS Press, Santiago de Compostela (2020). https://doi.org/10.3233/FAIA200179
29. Shih, A., Choi, A., Darwiche, A.: A symbolic approach to explaining bayesian network classifiers. In: Proceedings of the Twenty-Seventh Int. Joint Conf. on Artificial Intelligence, IJCAI 2018. pp. 5103–5111 (2018)
30. Sklar, E.I., Azhar, M.Q.: Explanation through argumentation. In: Proceedings of the 6th International Conference on Human-Agent Interaction, HAI 2018,

Southampton, United Kingdom, December 15-18, 2018. pp. 277–285 (2018). https://doi.org/10.1145/3284432.3284470

31. Sokol, K., Flach, P.A.: Explainability fact sheets: a framework for systematic assessment of explainable approaches. In: FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020. pp. 56–67 (2020). https://doi.org/10.1145/3351095.3372870
32. Swartout, W.R., Paris, C., Moore, J.D.: Explanations in Knowledge Systems: Design for Explainable Expert Systems. IEEE Expert **6**(3), 58–64 (jun 1991). https://doi.org/10.1109/64.87686
33. Timmer, S.T., Meyer, J.C., Prakken, H., Renooij, S., Verheij, B.: A two-phase method for extracting explanatory arguments from bayesian networks. Int. J. Approx. Reason. **80**, 475–494 (2017). https://doi.org/10.1016/j.ijar.2016.09.002
34. Vassiliades, A., Bassiliades, N., Patkos, T.: Argumentation and Explainable Artificial Intelligence: A Survey. Knowledge Eng. Rev. **36**(2) (2021)