Good Explanations for Formal Argumentation

AnneMarie Borg[1] and Floris Bex[1,2]

[1]Department of Information and Computing Sciences, Utrecht University
[2]Department of Law, Technology, Markets and Society, Tilburg University

July 14, 2020

In recent years, *explainable AI* (XAI) has received much attention, mostly directed at new techniques for explaining decisions of (subsymbolic) machine learning algorithms [11]. However, explanations traditionally also play an important role in (symbolic) knowledge-based systems [8]. Computational argumentation is one research area in symbolic AI that is frequently mentioned in relation to XAI. For example, arguments can be used to provide reasons for or against decisions [1, 8, 9]. The focus can also be on the argumentation itself, where it is explained whether and why a certain argument or claim can be accepted under certain semantics for computational argumentation [5, 6, 7]. It is the latter type of explanations that we are interested in.

Two central, related concepts in computational argumentation are *abstract argumentation frameworks* [3] – sets of arguments (abstract entities) and the attack relations between them – and *structured or logical argumentation frameworks* [2] – where arguments are constructed from a knowledge base and a set of rules and the attack relation is based on the individual elements in the arguments. Common for argumentation frameworks, whether abstract or structured, is that we can determine their extensions, sets of arguments that can collectively be considered as acceptable, under different semantics [3]. In XAI terms, this is very much a *global* explanation - what can we conclude from the model as a whole? However, as formal argumentation is being applied in real-life AI systems with lay-users, we would rather have a simpler, more compact explanation for the acceptability of individual arguments or claims – this is a *local* explanation for a particular decision or conclusion [4].

An important aspect of a *good* explanation, is that the receiver of the explanation understands the decision and trusts the system. To this end it is important to incorporate findings from the social sciences on how humans request, generate, interpret and evaluate explanations such as discussed in e.g., [9, 10, 11]. Although the discussion and implementation of these findings and their implications for artificial intelligence systems are usually aimed at machine learning applications, many are applicable to explanations for argumentation-based decisions as well. We will therefore discuss the most important of these findings for explanations in the context of formal argumentation. Along the way, further advantages of argumentation in an XAI context will become clear as well.

References

[1] Katie Atkinson, Pietro Baroni, Massimiliano Giacomin, Anthony Hunter, Henry Prakken, Chris Reed, Guillermo Simari, Matthias Thimm, and Serena Villata. Towards Artificial Argumentation. *AI magazine*, 38(3):25–36, 2017.

[2] Philippe Besnard, Alejandro Garcia, Anthony Hunter, Sanjay Modgil, Henry Prakken, Guillermo Simari, and Francesca Toni. Introduction to structured argumentation. *Argument & Computation*, 5(1):1–4, 2014.

[3] Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–357, 1995.

[4] Lilian Edwards and Michael Veale. Slave to the algorithm: Why a 'right to an explanation' is probably not the remedy you are looking for. *Duke Law & Technology Review*, 16(1):18–84, 2017.

[5] Xiuyi Fan and Francesca Toni. On computing explanations in argumentation. In Blai Bonet and Sven Koenig, editors, *Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI'15)*, pages 1496–1502. AAAI Press, 2015.

[6] Xiuyi Fan and Francesca Toni. On explanations for non-acceptable arguments. In Elizabeth Black, Sanjay Modgil, and Nir Oren, editors, *Proceedings of the 3rd International Workshop on Theory and Applications of Formal Argumentation, (TAFA'15)*, LNCS 9524, pages 112–127. Springer, 2015.

[7] Alejandro García, Carlos Chesñevar, Nicolás Rotstein, and Guillermo Simari. Formalizing dialectical explanation support for argument-based reasoning in knowledge-based systems. *Expert Systems with Applications*, 40(8):3233 – 3247, 2013.

[8] Carmen Lacave and Francisco J Diez. A review of explanation methods for heuristic expert systems. *The Knowledge Engineering Review*, 19(2):133–146, 2004.

[9] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1 – 38, 2019.

[10] Wojciech Samek and Klaus-Robert Müller. Towards explainable artificial intelligence. In Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller, editors, *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 5–22. Springer, 2019.

[11] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*, 2017.