Sequence Explanations for Acceptance in Abstract Argumentation (Extended Abstract)

Lars Bengel and Matthias Thimm

Artificial Intelligence Group, University of Hagen, Germany
{lars.bengel, matthias.thimm}@fernuni-hagen.de

1 Introduction

In recent years, explainability has been a major focus in the field of artificial intelligence (AI). One of the more promising approaches to explainable artificial intelligence is *formal argumentation* [4,6,22], which is well-suited to provide human-understandable explanations [3,25,26]. Various recent works are concerned with computing post-hoc argumentative explanations for black-box AI models [19,27]. On the other hand, the problem of explaining the reasoning within formal argumentation methods has also received lots of attention in the literature over the years [16,30,32]. In this work, we consider the latter scenario, in particular, we are concerned with providing explanations for the acceptance [21] of an argument within an *abstract argumentation framework* (AF) [20]. Formal argumentation is inherently linked with dialectics [23,28] and two fundamental aspects of dialectical argumentation are the *procedurality* and the *exchange of arguments*, i. e., the fact that arguments and counterarguments are brought forward one after another in alternating fashion [23]. The aim of this work is to define an explanation method that takes both of these aspects into account and incorporates them properly within the explanations themselves, which has so far not been considered in the literature.

To achieve this goal, we introduce *sequence explanations* for argument acceptance in argumentation frameworks. We base our work on the notion of *serialisability* [31], which provides a procedural form of representation for argumentation semantics [10,11]. A sequence explanation is then essentially a series of minimally acceptable (atomic) sets of arguments that leads to the acceptance of the argument in question. We define minimal sequence explanations that ensure that every decision and argument in the sequence is actually relevant to explain the acceptance of the target argument. Moreover, we expand sequence explanations to also incorporate counterarguments in order to obtain full *dialectical sequence explanations*. These then also allow us to distinguish between two different levels of strength of arguments that challenge the acceptance of the argument within the dialectical explanation.

The full version of the work presented here has been published in [12], which also includes a principle-based analysis based on existing and novel principles for acceptance explanation methods, further variants of sequence explanations and formal results on the relation to other explanation approaches.

2 Preliminaries

We consider the *abstract argumentation framework*, which is a directed graph $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ where $\mathcal{A}$ is a finite set of argument nodes and $\mathcal{R}$ is a relation of attack $\mathcal{R} \subseteq \mathcal{A} \times \mathcal{A}$ [20]. For some set $S \subseteq \mathcal{A}$ we define the set of arguments attacked by (attacking) S as follows

$$S^+_\mathcal{F} = \{a \in \mathcal{A} \mid \exists b \in S : b\mathcal{R}a\}, \qquad S^-_\mathcal{F} = \{a \in \mathcal{A} \mid \exists b \in S : a\mathcal{R}b\}.$$

We denote $\mathsf{Relevant}_\mathcal{F}(\boldsymbol{a}) = \{b \in \mathcal{A} \mid \text{there is a directed path from } b \text{ to } a\}$ as the set of arguments relevant for $\boldsymbol{a}$ in $\mathcal{F}$ [16]. We say that a set $S \subseteq \mathcal{A}$ is *conflict-free* iff for all $\boldsymbol{a}, b \in S$ it is not the case that $\boldsymbol{a}\mathcal{R}b$. A set S *defends* an argument $b \in \mathcal{A}$ iff for all $\boldsymbol{a}$ with $\boldsymbol{a}\mathcal{R}b$ there is $c \in S$ with $c\mathcal{R}a$. Furthermore, a set S is called *admissible* (ad) iff it is conflict-free and S defends all $\boldsymbol{a} \in S$. Let $\mathsf{ad}(\mathcal{F})$ denote the set of admissible sets of $\mathcal{F}$.

Non-empty minimal admissible sets have been coined *initial sets* [33,34].

Definition 1. *For $\mathcal{F} = (\mathcal{A}, \mathcal{R})$, a set $S \subseteq \mathcal{A}$ with $S \neq \emptyset$ is called an* initial set *(is) if S is admissible and there is no admissible $S' \subsetneq S$ with $S' \neq \emptyset$.*

We denote with $\mathsf{is}(\mathcal{F})$ the set of initial sets of the AF $\mathcal{F}$. Initial sets can essentially be understood as the atomic semantic units for admissibility-based semantics. Furthermore, we recall the definition of the *reduct* [8].

Definition 2. *For $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ and $S \subseteq \mathcal{A}$, the S-reduct $\mathcal{F}^S$ is defined as the AF $\mathcal{F}^S = (\mathcal{A}', \mathcal{R} \cap \mathcal{A}' \times \mathcal{A}')$ with $\mathcal{A}' = \mathcal{A} \setminus (S \cup S^+)$.*

The S-reduct of an AF is the AF that remains after removing S and all arguments attacked by S. In other words, we basically resolve S in the AF.

We recall the concept of *serialisability* [31], which is a notion that allows to characterise admissible sets in a constructive and procedural manner. For that, we define the *serialisation sequence* which is a decomposition of an admissible set into a series of initial sets of the respective reducts [14].

Definition 3. *For $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ $\mathcal{S} = (S_1, \dots S_n)$ is a* serialisation sequence *if $S_1 \in \mathsf{is}(\mathcal{F})$ and for each $2 \leq i \leq n$ we have that $S_i \in \mathsf{is}(\mathcal{F}^{S_1 \cup \dots \cup S_{i-1}})$.*

For a serialisation sequence $\mathcal{S} = (S_1, \dots S_n)$, we denote with $\hat{\mathcal{S}} = S_1 \cup \dots \cup S_n$ the admissible set induced by $\mathcal{S}$. Any admissible set is induced by at least one such sequence [31]. We denote with $\mathfrak{S}(\mathcal{F})$ the serialisation sequences of $\mathcal{F}$.

Example 1. We consider the AF $\mathcal{F}_1$ in Figure 1 with $\mathsf{is}(\mathcal{F}_1) = \{\{\boldsymbol{d}\}, \{\boldsymbol{e}\}, \{\boldsymbol{f}\}\}$. Consider the sequence $(\{\boldsymbol{f}\}, \{\boldsymbol{b}\}, \{\boldsymbol{g}\}, \{\boldsymbol{e}\}) \in \mathfrak{S}(\mathcal{F}_1)$. We have $\{\boldsymbol{f}\} \in \mathsf{is}(\mathcal{F}_1)$ and in the reduct ${\mathcal{F}_1}^{\{\boldsymbol{f}\}}$, we remove $\boldsymbol{f}, \boldsymbol{c}$ and $\boldsymbol{h}$. Thus, $\boldsymbol{b}$ is now unattacked and $\{\boldsymbol{b}\} \in \mathsf{is}({\mathcal{F}_1}^{\{\boldsymbol{f}\}})$. Is is then easy to see that $\{\boldsymbol{h}\} \in \mathsf{is}({\mathcal{F}_1}^{\{\boldsymbol{b},\boldsymbol{f}\}})$ and clearly $\{\boldsymbol{e}\}$ is an initial set of ${\mathcal{F}_1}^{\{\boldsymbol{b},\boldsymbol{f},\boldsymbol{g}\}}$ since it defends itself against the only other remaining argument $\boldsymbol{d}$. The sequence then induces the admissible set $\{\boldsymbol{b}, \boldsymbol{e}, \boldsymbol{f}, \boldsymbol{g}\}$. You may verify that there are multiple other sequences that induce this set, for instance $(\{\boldsymbol{e}\}, \{\boldsymbol{f}\}, \{\boldsymbol{g}\}, \{\boldsymbol{b}\})$. We may also include the initial set $\{\boldsymbol{d}\}$ at some point in the sequence to induce a different admissible set. Naturally, every sub-sequence of the above sequences is also a serialisation sequence of $\mathcal{F}_1$.

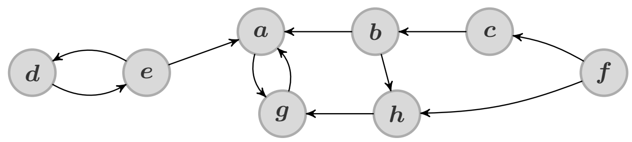

Fig. 1. The AF $\mathcal{F}_1$ from Examples 1 - 5.

3 Sequence Explanations for Argument Acceptance

We introduce a novel approach for explanations of argument acceptance built on the notion of serialisation sequences. Serialisation sequences provide construction schemes for admissible sets. Meaning, the *sequence explanations* (and the variants that we introduce in the following) are only built on the notion of admissibility and are independent of semantics. Instead of constructing arbitrary admissible sets, we will use this procedure to accept atomic semantic building blocks, i. e., initial sets, until we reach the argument whose acceptance we want to explain. Intuitively, an explanation for the acceptance of an argument $\boldsymbol{a}$ then represents a process of decisions ultimately leading to the acceptance of $\boldsymbol{a}$.

Definition 4. *Let $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ be an AF and $\boldsymbol{a} \in \mathcal{A}$. We define the set of* sequence explanations $\textit{SeqEx}(\mathcal{F}, \boldsymbol{a})$ *for the acceptance of $\boldsymbol{a}$ given $\mathcal{F}$ as:*

$$\textit{SeqEx}(\mathcal{F}, \boldsymbol{a}) = \{(S_1, \dots S_n) \in \mathfrak{S}(\mathcal{F}) \mid \boldsymbol{a} \in S_n\}$$

Example 2. Consider again the AF $\mathcal{F}_1$ depicted in Figure 1. $(\{\boldsymbol{f}\}, \{\boldsymbol{e}\}, \{\boldsymbol{b}\})$ is a sequence explanation for the acceptance of $\boldsymbol{b}$. We also have the sequence explanations $(\{\boldsymbol{d}\}, \{\boldsymbol{f}\}, \{\boldsymbol{b}\})$, $(\{\boldsymbol{f}\}, \{\boldsymbol{e}\}, \{\boldsymbol{g}\}, \{\boldsymbol{b}\})$ or $(\{\boldsymbol{f}\}, \{\boldsymbol{b}\})$. Note however, that the first three sequence explanations include the arguments $\boldsymbol{d}, \boldsymbol{e}$ or $\boldsymbol{g}$, even though none of them is relevant for $\boldsymbol{b}$, i. e., we have $\mathsf{Relevant}_{\mathcal{F}_1}(\boldsymbol{b}) = \{\boldsymbol{b}, \boldsymbol{c}, \boldsymbol{f}\}$.

As highlighted by the above example, this definition does not ensure that all arguments that occur in the explanation for the acceptance of an argument $\boldsymbol{a}$ are actually relevant for the argument $\boldsymbol{a}$.

In order to properly incorporate relevance into the explanations, we refine the definition of an explanation to be a minimal serialisation sequence $(S_1, \dots S_n)$ such that $\boldsymbol{a} \in S_n$. In other words, such an explanation for $\boldsymbol{a}$ represents a minimal sequence of conflict resolutions that lead to $\boldsymbol{a}$ being acceptable in $\mathcal{F}$. For that, we define the length of a serialisation sequence $\mathcal{S} = (S_1, \dots S_n)$ simply as the number of initial sets it contains, i. e., $|\mathcal{S}| = n$. For two serialisation sequences $\mathcal{S}, \mathcal{S}'$ we define $\mathcal{S} \sqsubseteq \mathcal{S}'$ iff $|\mathcal{S}| \leq |\mathcal{S}'|$.

Definition 5. *Let $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ be an AF and $\boldsymbol{a} \in \mathcal{A}$. We define the set of* minimal sequence explanations $\textit{MinSeqEx}(\mathcal{F}, \boldsymbol{a})$ *for the acceptance of $\boldsymbol{a}$ given $\mathcal{F}$ as:*

$$\textit{MinSeqEx}(\mathcal{F}, a) = \min_{\sqsubseteq} \textit{SeqEx}(\mathcal{F}, a)$$

Example 3. We continue Example 2 with the AF $\mathcal{F}_1$ in Figure 1. There is only one minimal sequence explanation for $\boldsymbol{b}$, namely $(\{\boldsymbol{f}\}, \{\boldsymbol{b}\})$. On the other hand, for the argument $\boldsymbol{g}$, we have the sequence explanations $(\{\boldsymbol{f}\}, \{\boldsymbol{e}\}, \{\boldsymbol{g}\})$, $(\{\boldsymbol{e}\}, \{\boldsymbol{f}\}, \{\boldsymbol{g}\})$, $(\{\boldsymbol{f}\}, \{\boldsymbol{b}\}, \{\boldsymbol{g}\})$ and $(\{\boldsymbol{f}\}, \{\boldsymbol{g}\})$, but only the latter is minimal.

Indeed, including minimality (wrt. the length of the explanation sequence) is enough to ensure that only relevant arguments are included in the explanation as the following result shows (Proofs are available in the extended version [13]).

Proposition 1. *Let $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ be an AF and $\boldsymbol{a} \in \mathcal{A}$. Then, we have that for every $\mathcal{E} \in \textsc{MinSeqEx}(\mathcal{F}, \boldsymbol{a})$ it holds that $\hat{\mathcal{E}} \setminus \{\boldsymbol{a}\} \subseteq \mathsf{Relevant}_{\mathcal{F}}(\boldsymbol{a})$.*

So far, the sequence explanations take into account the procedural aspect of argumentation by providing a sequence of minimally acceptable sets that essentially support the argument in question. We now turn to the second fundamental aspect of dialectical argumentation, namely the exchange of arguments and counterarguments. In order to construct human-understandable argumentative explanations, we also need to incorporate the appropriate counterarguments, so that the explanations provide a clear line of reasoning. To achieve this, we associate with some sequence explanation $\mathcal{E}_s$, containing the *supporting* arguments for the acceptance of the argument $\boldsymbol{a}$, the sequence $\mathcal{E}_d$ of *defeated* arguments.

Definition 6. *Let $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ be an AF and $\boldsymbol{a} \in \mathcal{A}$. We define a* dialectical sequence explanation *for the acceptance of $\boldsymbol{a}$ given $\mathcal{F}$ as a pair of sequences:*

$$\mathcal{E}_s = (S_1, \dots S_n), \qquad \mathcal{E}_d = (T_1, \dots, T_n),$$

such that $\mathcal{E}_s$ is some sequence explanation for $\boldsymbol{a}$ and for each $i = 1, \dots, n$ we have $T_i = (\hat{\mathcal{E}_s} \cup \{\boldsymbol{a}\})^-_{\mathcal{F}} \cap (S_i)^+_{\mathcal{F}^{S_1 \cup \dots \cup S_{i-1}}}$.

Each T_i is defined to contain the attackers of $\boldsymbol{a}$ and its supporting arguments (represented by $(\hat{\mathcal{E}_S} \cup \{\boldsymbol{a}\})^-_{\mathcal{F}}$), assuming that they are rejected by S_i and have not been rejected in a previous step already, i.e., the arguments attacked by S_i in the reduct $\mathcal{F}^{S_1 \cup \dots \cup S_{i-1}}$.

Example 4. We consider again the AF $\mathcal{F}_1$ in Figure 1. We take the explanation $(\{\boldsymbol{f}\}, \{\boldsymbol{e}\}, \{\boldsymbol{g}\})$ for the acceptance of $\boldsymbol{g}$. The corresponding sequence of defeated arguments is $\mathcal{E}_d = (\{\boldsymbol{h}\}, \{\boldsymbol{a}, \boldsymbol{d}\}, \emptyset)$. Notice that, while $\boldsymbol{c}$ is attacked by $\boldsymbol{f}$, it is not included in $\mathcal{E}_d$, because $\boldsymbol{c}$ does not attack any argument of the explanation sequence and thus does not contribute anything to the explanation. Even though $\boldsymbol{g}$ also attacks $\boldsymbol{a}$, $\boldsymbol{a}$ has already been defeated by $\boldsymbol{e}$ in a previous step of the argumentation process and is therefore not included again.

It can then be shown that a dialectical explanation, based on a minimal sequence explanation for the acceptance of some argument $\boldsymbol{a}$, only contains arguments that are relevant for $\boldsymbol{a}$ and no arguments are repeated.

Proposition 2. *Let $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ be an AF and $((S_1, \dots S_n), (T_1, \dots, T_n))$ is a dialectical explanation for $\boldsymbol{a} \in \mathcal{A}$ with $(S_1, \dots S_n) \in \textsc{MinSeqEx}(\mathcal{F}, \boldsymbol{a})$. It holds that $T_i \subseteq \mathsf{Relevant}_{\mathcal{F}}(\boldsymbol{a})$ and $T_i \cap T_j = \emptyset$ for all $i, j = 0, \dots, n$ with $i \neq j$.*

To facilitate the construction of insightful explanations, our approach also allows us to distinguish further between two types of defeated arguments:

(1) *necessarily rejected* arguments, i. e., they attack the corresponding initial set and must be defended against: $\mathsf{NecRej}_{\mathcal{F}}(S) = S^- \cap S^+$
(2) *incidentally rejected* arguments, i. e., rejection simply follows logically, but is not necessary for its acceptance: $\mathsf{IncRej}_{\mathcal{F}}(S) = S^+ \setminus S^-$

This essentially allows us to distinguish between weak and strong counterarguments. Strong counterarguments actively challenge the explanation (within the sequence explanation) while weak counterarguments do not. This can prove useful when presenting such an explanation to a user, or when analysing the strength of the argument or its explanation.

Example 5. We continue Example 4 with the AF $\mathcal{F}_1$ in Figure 1. Consider the dialectical sequence explanation $(\mathcal{E}_s, \mathcal{E}_d)$ for the acceptance of $\boldsymbol{g}$ with $\mathcal{E}_s = (\{\boldsymbol{f}\}, \{\boldsymbol{e}\}, \{\boldsymbol{g}\})$ and $\mathcal{E}_d = (\{\boldsymbol{h}\}, \{\boldsymbol{a}, \boldsymbol{d}\}, \emptyset)$. We examine, step by step, the defeated attackers of the explanation: $\boldsymbol{h}, \boldsymbol{d}, \boldsymbol{a}$. First, we have that $\mathsf{IncRej}_{\mathcal{F}_1}(\{\boldsymbol{f}\}) = \{\boldsymbol{h}\}$. Furthermore, we have $\mathsf{NecRej}_{\mathcal{F}_1\{\boldsymbol{f}\}}(\{\boldsymbol{e}\}) = \{\boldsymbol{d}\}$. On the other hand, we have $\mathsf{IncRej}_{\mathcal{F}_1\{\boldsymbol{f}\}}(\{\boldsymbol{e}\}) = \{\boldsymbol{a}\}$. Meaning essentially, that $\boldsymbol{h}$ and $\boldsymbol{a}$ are merely weak counterarguments and $\boldsymbol{d}$ is a strong counterargument, in the context of this sequence. If we consider instead the sequence explanation $(\{\boldsymbol{f}\}, \{\boldsymbol{g}\})$, with the defeated arguments $(\{\boldsymbol{h}\}, \{\boldsymbol{a}\})$, $\boldsymbol{h}$ is again a weak counterargument, but $\boldsymbol{a}$ is now a strong contender, since it is necessarily rejected by $\boldsymbol{g}$ in this sequence.

4 Summary and Discussion

We have introduced (dialectical) sequence explanations, which provide a new form of explanation for the acceptance of arguments that incorporate both the procedural and dialectical aspect of argumentation directly into the explanation. Moreover, our approach gives a fine-grained view into the strength of counterarguments. In the literature, there are many explanation approaches that are based on admissibility [2,16,21], however these typically do not comprise any structural information. There exists approaches that incorporate some form of structure [1,7,9], but those lack the dialectical aspect and the conciseness provided by utilising initial sets as the atomic semantic units. Other approach are not built on admissibility and instead consider sub-frameworks [29,32] or critical sets [15] to explain (non-)acceptance. Approaches that take into account dialectical aspects are discussion games [17] and dispute trees [18]. In contrast to our work, they consist of individual arguments, instead of initial sets, and allow arguments to be repeated. In other fields of KR explanations also play an important role, for instance in description logic [5] or logic programming [24].

Ultimately, we believe that representing explanations as sequences is the superior choice if one wants to properly construct argumentative explanations. In particular, to properly model an exchange of arguments, utilising a sequence-based representation is inevitable. Assessing whether this proves true for sequence explanations in practice is the next step for future work.

Acknowledgements

The research reported here was partially supported by the Deutsche Forschungsgemeinschaft (grant 550735820).

References

1. Alfano, G., Calautti, M., Greco, S., Parisi, F., Trubitsyna, I.: Explainable acceptance in probabilistic and incomplete abstract argumentation frameworks. Artif. Intell. **323**, 103967 (2023). https://doi.org/10.1016/J.ARTINT.2023.103967, https://doi.org/10.1016/j.artint.2023.103967
2. Amgoud, L.: Post-hoc explanation of extension semantics. In: ECAI 2024 - 27th European Conference on Artificial Intelligence, 2024. Frontiers in Artificial Intelligence and Applications, vol. 392, pp. 3276–3283. IOS Press (2024). https://doi.org/10.3233/FAIA240875, https://doi.org/10.3233/FAIA240875
3. Antaki, C., Leudar, I.: Explaining in conversation: Towards an argument model. European Journal of Social Psychology **22**(2), 181–194 (1992)
4. Atkinson, K., Bench-Capon, T.J.M., Bollegala, D.: Explanation in AI and law: Past, present and future. Artif. Intell. **289**, 103387 (2020). https://doi.org/10.1016/J.ARTINT.2020.103387, https://doi.org/10.1016/j.artint.2020.103387
5. Baader, F., Peñaloza, R.: Automata-based axiom pinpointing. J. Autom. Reason. **45**(2), 91–129 (2010). https://doi.org/10.1007/S10817-010-9181-2, https://doi.org/10.1007/s10817-010-9181-2
6. Baroni, P., Gabbay, D., Giacomin, M., van der Torre, L. (eds.): Handbook of Formal Argumentation, Vol.1. College Publications (2018), https://www.collegepublications.co.uk/downloads/handbooks00003.pdf
7. Baroni, P., Giacomin, M., Guida, G.: SCC-recursiveness: a general schema for argumentation semantics. Artif. Intell. **168**(1-2), 162–210 (2005). https://doi.org/10.1016/J.ARTINT.2005.05.006, https://doi.org/10.1016/j.artint.2005.05.006
8. Baumann, R., Brewka, G., Ulbricht, M.: Revisiting the foundations of abstract argumentation - semantics based on weak admissibility and weak defense. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020. pp. 2742–2749. AAAI Press (2020). https://doi.org/10.1609/AAAI.V34I03.5661, https://doi.org/10.1609/aaai.v34i03.5661
9. Baumann, R., Ulbricht, M.: Choices and their consequences - explaining acceptable sets in abstract argumentation frameworks. In: Proceedings of the 18th International Conference on Principles of Knowledge Representation and Reasoning, KR 2021. pp. 110–119 (2021). https://doi.org/10.24963/KR.2021/11, https://doi.org/10.24963/kr.2021/11
10. Bengel, L., Sander, J., Thimm, M.: Characterising serialisation equivalence for abstract argumentation. In: ECAI 2024 - 27th European Conference on Artificial Intelligence, 2024. Frontiers in Artificial Intelligence and Applications, vol. 392, pp. 3340–3347. IOS Press (2024). https://doi.org/10.3233/FAIA240883, https://doi.org/10.3233/FAIA240883
11. Bengel, L., Thimm, M.: Serialisable semantics for abstract argumentation. In: Toni, F., Polberg, S., Booth, R., Caminada, M., Kido, H. (eds.) Computational Models of Argument - Proceedings of COMMA 2022. pp. 80–91. IOS Press (2022). https://doi.org/10.3233/FAIA220143, https://doi.org/10.3233/FAIA220143

12. Bengel, L., Thimm, M.: Sequence explanations for acceptance in abstract argumentation. In: Proceedings of the 22th International Conference on Principles of Knowledge Representation and Reasoning, KR 2025 (2025)
13. Bengel, L., Thimm, M.: Sequence Explanations for Acceptance in Abstract Argumentation (Extended Version). Zenodo (Jul 2025). https://doi.org/10.5281/zenodo.16024539, https://doi.org/10.5281/zenodo.16024539
14. Blümel, L., Thimm, M.: A ranking semantics for abstract argumentation based on serialisability. In: Toni, F., Polberg, S., Booth, R., Caminada, M., Kido, H. (eds.) Computational Models of Argument - Proceedings of COMMA 2022. pp. 104–115. IOS Press (2022). https://doi.org/10.3233/FAIA220145, https://doi.org/10.3233/FAIA220145
15. Booth, R., Caminada, M., Dunne, P.E., Podlaszewski, M., Rahwan, I.: Complexity properties of critical sets of arguments. In: Parsons, S., Oren, N., Reed, C., Cerutti, F. (eds.) Computational Models of Argument - Proceedings of COMMA 2014. Frontiers in Artificial Intelligence and Applications, vol. 266, pp. 173–184. IOS Press (2014). https://doi.org/10.3233/978-1-61499-436-7-173, https://doi.org/10.3233/978-1-61499-436-7-173
16. Borg, A., Bex, F.: Minimality, necessity and sufficiency for argumentation and explanation. Int. J. Approx. Reason. **168**, 109143 (2024). https://doi.org/10.1016/J.IJAR.2024.109143, https://doi.org/10.1016/j.ijar.2024.109143
17. Caminada, M.: Argumentation semantics as formal discussion. In: Baroni, P., Gabbay, D., Giacomin, M., van der Torre, L. (eds.) Handbook of Formal Argumentation, Vol.1, pp. 487–518. College Publications (2018), https://www.collegepublications.co.uk/downloads/handbooks00003.pdf
18. Cyras, K., Fan, X., Schulz, C., Toni, F.: Assumption-based argumentation: Disputes, explanations, preferences. IFCoLog Journal of Logics and Their Applications **4**(8), 2407 (2017)
19. Cyras, K., Rago, A., Albini, E., Baroni, P., Toni, F.: Argumentative XAI: A survey. In: Zhou, Z. (ed.) Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021. pp. 4392–4399. ijcai.org (2021). https://doi.org/10.24963/IJCAI.2021/600, https://doi.org/10.24963/ijcai.2021/600
20. Dung, P.M.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artificial Intelligence **77**(2), 321–358 (1995). https://doi.org/10.1016/0004-3702(94)00041-X, https://doi.org/10.1016/0004-3702(94)00041-X
21. Fan, X., Toni, F.: On computing explanations in abstract argumentation. In: Schaub, T., Friedrich, G., O'Sullivan, B. (eds.) ECAI 2014 - 21st European Conference on Artificial Intelligence. Frontiers in Artificial Intelligence and Applications, vol. 263, pp. 1005–1006. IOS Press (2014). https://doi.org/10.3233/978-1-61499-419-0-1005, https://doi.org/10.3233/978-1-61499-419-0-1005
22. Gabbay, D., Giacomin, M., Simari, G.R., Thimm, M. (eds.): Handbook of Formal Argumentation, Vol. 2. College Publications (2021), https://www.collegepublications.co.uk/downloads/handbooks00006.pdf
23. Hage, J.: Dialectical models in artificial intelligence and law. Artificial Intelligence and Law **8**(2/3), 137–172 (2000). https://doi.org/10.1023/A:1008348321016, https://doi.org/10.1023/A:1008348321016
24. Kakas, A.C., Kowalski, R.A., Toni, F.: Abductive logic programming. J. Log. Comput. **2**(6), 719–770 (1992). https://doi.org/10.1093/LOGCOM/2.6.719, https://doi.org/10.1093/logcom/2.6.719

25. Leofante, F., Ayoobi, H., Dejl, A., Freedman, G., Gorur, D., Jiang, J., Paulino-Passos, G., Rago, A., Rapberger, A., Russo, F., Yin, X., Zhang, D., Toni, F.: Contestable AI needs computational argumentation. In: Marquis, P., Ortiz, M., Pagnucco, M. (eds.) Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning, KR 2024, 2024 (2024). https://doi.org/10.24963/KR.2024/83, https://doi.org/10.24963/kr.2024/83
26. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artif. Intell. **267**, 1–38 (2019). https://doi.org/10.1016/J.ARTINT.2018.07.007, https://doi.org/10.1016/j.artint.2018.07.007
27. Potyka, N., Yin, X., Toni, F.: Explaining random forests using bipolar argumentation and markov networks. In: Williams, B., Chen, Y., Neville, J. (eds.) Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, 2023. pp. 9453–9460. AAAI Press (2023). https://doi.org/10.1609/AAAI.V37I8.26132, https://doi.org/10.1609/aaai.v37i8.26132
28. Rescher, N.: Dialectics: A controversy-oriented approach to the theory of knowledge. Suny Press (1977)
29. Saribatur, Z.G., Wallner, J.P., Woltran, S.: Explaining non-acceptability in abstract argumentation. In: ECAI 2020 - 24th European Conference on Artificial Intelligence. Frontiers in Artificial Intelligence and Applications, vol. 325, pp. 881–888. IOS Press (2020). https://doi.org/10.3233/FAIA200179, https://doi.org/10.3233/FAIA200179
30. Seselja, D., Straßer, C.: Abstract argumentation and explanation applied to scientific debates. Synth. **190**(12), 2195–2217 (2013). https://doi.org/10.1007/S11229-011-9964-Y, https://doi.org/10.1007/s11229-011-9964-y
31. Thimm, M.: Revisiting initial sets in abstract argumentation. Argument & Computation **13**(3), 325–360 (2022). https://doi.org/10.3233/AAC-210018, https://doi.org/10.3233/AAC-210018
32. Ulbricht, M., Wallner, J.P.: Strong explanations in abstract argumentation. In: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021. pp. 6496–6504. AAAI Press (2021). https://doi.org/10.1609/AAAI.V35I7.16805, https://doi.org/10.1609/aaai.v35i7.16805
33. Xu, Y., Cayrol, C.: Initial sets in abstract argumentation frameworks. In: Ågotnes, T., Liao, B., Wáng, Y.N. (eds.) Proceedings of the 1st Chinese Conference on Logic and Argumentation (CLAR 2016), Hangzhou, China, April 2-3, 2016. CEUR Workshop Proceedings, vol. 1811, pp. 72–85. CEUR-WS.org (2016), https://ceur-ws.org/Vol-1811/paper6.pdf
34. Xu, Y., Cayrol, C.: Initial sets in abstract argumentation frameworks. Journal of Applied Non-Classical Logics **28**(2-3), 260–279 (2018). https://doi.org/10.1080/11663081.2018.1457252, https://doi.org/10.1080/11663081.2018.1457252