Sequence Explanations for Acceptance in Abstract Argumentation

Lars Bengel, **Matthias Thimm**
Artificial Intelligence Group, University of Hagen, Germany
{larsbengel, matthiasthimm}@fernuni-hagen.de

Abstract

We consider abstract argumentation and explanations for the acceptance of arguments. Based on the notion of serialisability, we introduce *sequence explanations* as a procedural form of explanation for the acceptance of some argument. Intuitively, these explanations represent the process of accepting (and rejecting) arguments in order to conclude the acceptance of a certain argument. We define several variants of sequence explanations and examine them in detail. In particular, we also incorporate counterarguments into the explanations to make them dialectical. Finally, we relate our explanations to other approaches from the literature via a principle-based analysis.

1 Introduction

In recent years, explainability has been a major focus in artificial intelligence (AI) research. One of the promising approaches to explainable artificial intelligence is formal argumentation (Atkinson, Bench-Capon, and Bollegala 2020). Many works have already highlighted how argumentative approaches are well-suited to provide human-understandable explanations (Antaki and Leudar 1992; Miller 2019; Leofante et al. 2024). Various recent works are concerned with computing post-hoc argumentative explanations for black-box AI models (Cyras et al. 2021; Potyka, Yin, and Toni 2022). On the other hand, the problem of explaining the reasoning within formal argumentation methods has also received lots of attention in the literature (Seselja and Straßer 2013; Ulbricht and Wallner 2021; Borg and Bex 2024). In this work, we consider the latter scenario, in particular, we are concerned with providing explanations for the acceptance of arguments within abstract argumentation frameworks (Fan and Toni 2014).

Abstract argumentation frameworks (AFs) as introduced by Dung (1995) continue to be the most prominent argumentation formalism in the literature. In an argumentation framework arguments are modelled as abstract entities and directed attacks represent conflicts between them. Reasoning is performed via semantics that select jointly acceptable sets of arguments, called extensions, according to different criteria. Argumentation is inherently linked with dialectics (Rescher 1977). Two key 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 (Hage 2000). 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. This has so far not been considered in the literature.

In regard of this goal, we will consider the notion of *serialisability* for abstract argumentation, which provides a non-deterministic construction scheme for admissible sets (Thimm 2022). Hereby, an admissible set is constructed by iteratively accepting atomic semantical units and subsequently removing their associated resolved conflict from the argumentation framework. These atomic semantical units are minimal, non-empty admissible sets and are called *initial sets* (Xu and Cayrol 2018). This approach allows us to associate with an arbitrary admissible set so called *serialisation sequences*, which are sequences of initial sets of the respective reducts. These serialisation sequences represent essentially the different construction processes of the corresponding admissible set. Intuitively, these serialisation sequences give structure to the admissible set and are well-suited to provide the procedurality we want for our explanation method (Bengel 2022). In particular, it has also been shown that serialisation sequences are more expressive than extension-based semantics (Bengel, Sander, and Thimm 2024).

In this work, we introduce *sequence explanations* for argument acceptance in AFs. A sequence explanation is essentially a serialisation sequence 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 within the dialectical explanation. Finally, we introduce two principles to address the above discussed aspects of dialectical argumentation and also provide a comprehensive analysis of sequence explanation based on principles from the literature. For that, we also consider the sufficient and necessary explanations of Borg and Bex (2024) and the strong σ-explanations of Ulbricht and Wallner (2021).

To summarise, the main contributions of this paper are:

- We introduce and investigate *sequence explanations* and several variants as a novel procedural form of explanations for argument acceptance (Section 4).
- We expand sequence explanations to also include arguments rejected in each step to make them dialectical (Section 5).
- We discuss our approach in the context of related work and provide a principle-based analysis (Section 6).

In Section 2 we recall the necessary background on argumentation, Section 3 describes other explanation strategies and principles for explanations from the literature and Section 7 concludes the paper. Proofs for all technical results can be found in the extended version of this paper (Bengel and Thimm 2025), available online[1].

2 Preliminaries

We consider *abstract argumentation frameworks* (AF) as introduced by Dung (1995).

Definition 1. An *abstract argumentation framework* (AF) is a pair $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ where $\mathcal{A}$ is a finite set of arguments and $\mathcal{R}$ is a relation of attack $\mathcal{R} \subseteq \mathcal{A} \times \mathcal{A}$.

Let $\mathbb{AF}$ denote the set of all abstract argumentation frameworks and $\mathbb{A}$ denotes the universe of all arguments. For two arguments $a, b \in \mathcal{A}$, the relation $a\mathcal{R}b$ means that argument a *attacks* argument b. For a set $S \subseteq \mathcal{A}$, we denote by $\mathcal{F}\downarrow_S = (S, \mathcal{R} \cap (S \times S))$ the *projection* of $\mathcal{F}$ on S. Additionally, for a set $S \subseteq \mathcal{A}$ we define

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

For a singleton set S, we omit brackets for readability, i. e., we write $a^-_{\mathcal{F}}$ ($a^+_{\mathcal{F}}$). We simply write S^+ (respectively S^-) instead of $S^+_{\mathcal{F}}$ (resp. $S^-_{\mathcal{F}}$) if $\mathcal{F}$ is clear from the context. For two sets S and S' we may write $S\mathcal{R}S'$ iff $S' \cap S^+_{\mathcal{F}} \neq \emptyset$. We say that a set $S \subseteq \mathcal{A}$ is *conflict-free* iff for all $a, b \in S$ it is not the case that $a\mathcal{R}b$. A set S *defends* an argument $b \in \mathcal{A}$ iff for all a with $a\mathcal{R}b$ 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 $a \in S$. Let $\mathsf{ad}(\mathcal{F})$ denote the set of admissible sets of $\mathcal{F}$.

We can then define the classical admissibility-based semantics by restricting admissible sets (Dung 1995; Baroni, Caminada, and Giacomin 2018). In particular, given $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ an admissible set $S \subseteq \mathcal{A}$ is called

- *complete* (co) iff S contains all $a \in \mathcal{A}$ that it defends,
- *grounded* (gr) iff S is complete and $\subseteq$-minimal,
- *preferred* (pr) iff S is $\subseteq$-maximal,
- *stable* (st) iff $S \cup S^+_{\mathcal{F}} = \mathcal{A}$.

Such a set is then also called a σ-extension, for $\sigma \in \{\mathsf{co}, \mathsf{gr}, \mathsf{pr}, \mathsf{st}\}$, and we denote with $\sigma(\mathcal{F})$ the set of σ-extensions of $\mathcal{F}$.

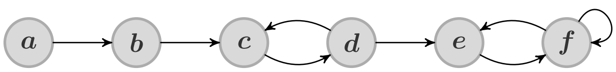

Figure 1: The AF $\mathcal{F}_1$ from Example 1.

Example 1. Consider the AF $\mathcal{F}_1$ depicted in Figure 1. We have that $\{a\}$, $\{a, c\}$, $\{a, d\}$ and $\{a, c, e\}$ are the complete extensions of $\mathcal{F}_1$. The set $\{a\}$ is also the unique grounded extension, while the latter two are the preferred extensions of $\mathcal{F}_1$. Finally, only $\{a, c, e\}$ is a stable extension of $\mathcal{F}_1$.

Non-empty minimal admissible sets have been coined *initial sets* by Xu and Cayrol (2018).

Definition 2. 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 $S' \neq \emptyset$.

$\mathsf{is}(\mathcal{F})$ denotes the set of initial sets of $\mathcal{F}$. We further differentiate between three types of initial sets (Thimm 2022).

Definition 3. For $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ and $S \in \mathsf{is}(\mathcal{F})$, we say that

1. S is *unattacked* iff $S^- = \emptyset$,
2. S is *unchallenged* iff $S^- \neq \emptyset$ and there is no $S' \in \mathsf{is}(\mathcal{F})$ with $S'\mathcal{R}S$,
3. S is *challenged* iff there is $S' \in \mathsf{is}(\mathcal{F})$ with $S'\mathcal{R}S$.

In the following, we denote with $\mathsf{is}^{\not\leftarrow}(\mathcal{F})$, $\mathsf{is}^{\not\leftrightarrow}(\mathcal{F})$, and $\mathsf{is}^{\leftrightarrow}(\mathcal{F})$ the set of unattacked, unchallenged, and challenged initial sets, respectively.

Example 2. Consider again the AF $\mathcal{F}_1$ in Figure 1. The AF has two initial sets: $\{a\}$ and $\{\mathrm{d}\}$. The former being unattacked and the latter unchallenged in $\mathcal{F}_1$.

Furthermore, we recall the definition of the *reduct* (Baumann, Brewka, and Ulbricht 2020).

Definition 4. For $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ and $S \subseteq \mathcal{A}$, the S-reduct $\mathcal{F}^S$ is defined as $\mathcal{F}^S = \mathcal{F}\downarrow_{\mathcal{A} \setminus (S \cup S^+_{\mathcal{F}})}$.

We recall the concept of *serialisability* (Thimm 2022), which is a property of a semantics that allows to characterise extensions in a constructive manner. For that we define the *serialisation sequence* which is a decomposition of an extension into a series of initial sets (Blümel and Thimm 2022).

Definition 5. A *serialisation sequence* for $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ is a sequence $\mathcal{S} = (S_1, \dots S_n)$ with $S_1 \in \mathsf{is}(\mathcal{F})$ 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 $\hat{\mathcal{S}} = S_1 \cup \dots \cup S_n$ and call $\hat{\mathcal{S}}$ the *induced extension* of $\mathcal{S}$. As shown in (Thimm 2022), we can characterise any admissible set by such sequences.

Proposition 1. *Let $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ be an AF and $S \subseteq \mathcal{A}$. It holds that $S \in \mathsf{ad}(\mathcal{F})$ if and only if there exists a serialisation sequence $\mathcal{S}$ with $\hat{\mathcal{S}} = S$.*

We denote with $\mathfrak{S}(\mathcal{F})$ the serialisation sequences of $\mathcal{F}$.

Example 3. We continue Example 2 with the AF $\mathcal{F}_1$ in Figure 1. Consider the sequence $(\{a\}, \{c\}, \{e\}) \in \mathfrak{S}(\mathcal{F}_1)$. We have $\{a\} \in \mathsf{is}(\mathcal{F}_1)$. Subsequently, in the reduct $\mathcal{F}_1^{\{a\}}$

[1] https://doi.org/10.5281/zenodo.16024540

the arguments a and b are removed from $\mathcal{F}_1$ and in the resulting AF $\mathcal{F}_1^{\{a\}}$ we have two (challenged) initial sets: $\{c\}$ and $\{d\}$. Finally, in the reduct $\mathcal{F}_1^{\{a,c\}}$ only the arguments e and f remain, with $\{e\}$ being the only initial set. Thus, $(\{a\}, \{c\}, \{e\})$ is a serialisation sequence, corresponding to the admissible set $\{a, c, e\}$. Notably, for the admissible set $\{a, d\}$ there are two serialisation sequences $(\{a\}, \{d\})$ and $(\{d\}, \{a\})$, due to the fact that both $\{a\}$ and $\{d\}$ are initial sets of $\mathcal{F}_1$ and their respective reducts. Naturally, every sub-sequence of the above sequences is also a serialisation sequence of $\mathcal{F}_1$.

3 Explanations in Abstract Argumentation

In this work, we consider methods of explaining the acceptance of an argument. For that, we define an *argument-explanation strategy* $\textsc{Expl}$ in general to be a function that given an AF $\mathcal{F}$ and some argument a returns sets of arguments E that explain the acceptance of a given $\mathcal{F}$.

Definition 6. Let $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ and $a \in \mathcal{A}$. An *argument-explanation strategy* $\textsc{Expl}$ is a function $\mathbb{AF} \times \mathbb{A} \mapsto 2^{2^{\mathbb{A}}}$. A set $E \in \textsc{Expl}(\mathcal{F}, a)$ for some $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ and $a \in \mathcal{A}$ is called an *explanation* for a in $\mathcal{F}$.

Note that for the sake of comparability, we generally define explanations in the form of sets, because all existing approaches in the literature use set-based representations for explanations.

In the following, we consider, in detail, some approaches for explanations of argument acceptance in abstract argumentation from the literature. First, we consider the sufficient and necessary explanations of Borg and Bex [2021; 2024]. For that, we first recall that for some directed graph $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ and $a, b \in \mathcal{A}$, a sequence $(a_1, \ldots, a_n)$ is called a *directed path* from a to b iff $a_1 = a$ and $a_n = b$, and for every $i = 2, \ldots, n$ it holds that $(a_{i-1}, a_i) \in \mathcal{R}$. Based on that, the notion of *relevance* is defined.

Definition 7. Let $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ be an AF and $a, b \in \mathcal{A}$. We say that b is *relevant* for a iff there exists a directed path from b to a in $\mathcal{F}$ and b does not attack itself. A set $S \subseteq \mathcal{A}$ is relevant for a iff all of its elements are relevant to a.

For some $\mathcal{F} = (\mathcal{A}, \mathcal{R})$, we also denote

$$\mathsf{Relevant}_{\mathcal{F}}(a) = \{b \in \mathcal{A} \mid b \text{ is relevant for } a\}$$

as the set of arguments relevant for a. We then distinguish between *sufficient* and *necessary* for acceptance as follows.

Definition 8. Let $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ be an AF and $a \in \mathcal{A}$. Then

- $S \subseteq \mathcal{A}$ is *sufficient for the acceptance* of a iff S is relevant for a, S is conflict-free and S defends $S \cup \{a\}$ against all attackers,
- $b \in \mathcal{A}$ is *necessary for the acceptance* of a iff b is relevant for a and for every $S \in \mathsf{ad}(\mathcal{F})$ it holds that, if $b \notin S$, then $a \notin S$.

Based on the above notions, we can then define different types of explanation strategies for the acceptance of an argument (Borg and Bex 2024).

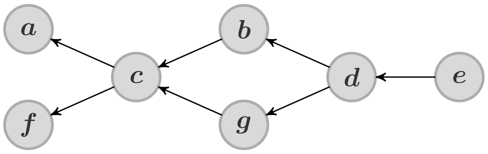

Figure 2: The AF $\mathcal{F}_2$ from Example 4.

Definition 9. Let $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ be an AF and $a \in \mathcal{A}$. Then,

$$\begin{aligned}
\textsc{Suff}(\mathcal{F}, a) &= \{S \cup \{a\} \subseteq \mathcal{A} \mid \\
&\qquad S \text{ is sufficient for the acceptance of } a\}, \\
\textsc{MinSuff}(\mathcal{F}, a) &= \min_{\subseteq} \textsc{Suff}(\mathcal{F}, a), \\
\textsc{Nec}(\mathcal{F}, a) &= \{\{a\} \cup \{b \in \mathcal{A} \mid \\
&\qquad b \text{ is necessary for the acceptance of } a\}\}.
\end{aligned}$$

The following example illustrates the three explanation strategies.

Example 4. Consider the AF $\mathcal{F}_2$ in Figure 2. Every argument except f is relevant for a. We have that $\textsc{Suff}(\mathcal{F}_2, a) = \{\{a, b, e\}, \{a, e, g\}, \{a, b, e, g\}\}$. The former two explanations are also $\subseteq$-minimal sufficient explanations. The unique necessary explanation for the acceptance of a is $\{a, e\}$. Note that the necessary explanation is not admissible and in particular does not defend a.

Secondly, we consider the strong σ-explanations due to Ulbricht and Wallner (2021) which are sets of arguments that ensure the argument is acceptable wrt. σ in every subframework containing the explaining set.

Definition 10. Let $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ be an AF, $a \in \mathcal{A}$ and σ is a semantics. Then,

$$\begin{aligned}
\textsc{Strong}_{\sigma}(\mathcal{F}, a) = \{E \subseteq \mathcal{A} \mid \\
&\forall \mathcal{A}' \subseteq \mathbb{A} \text{ s.t. } E \subseteq \mathcal{A}' \subseteq \mathcal{A} \wedge \\
&\exists S' \in \sigma(\mathcal{F}{\downarrow_{\mathcal{A}'}}) : a \in S'\}
\end{aligned}$$

Example 5. Consider again the AF $\mathcal{F}_2 = (\mathcal{A}_2, \mathcal{R}_2)$ in Figure 2. The set $\{a, e, g\}$ is a strong ad-explanation for the acceptance of a in $\mathcal{F}_2$. Note that every set S' with $\{a, e, g\} \subseteq S' \subseteq \mathcal{A}_2$ is also a strong ad-explanation for a. The same applies to $\{a, b, e\}$.

We only consider the above introduced explanation methods in detail in this work. However, there exist numerous other notable approaches to explanations in abstract argumentation. Fan and Toni (2014) introduce *related admissible* sets as explanations, which are minimal sets of arguments required to show (non-)acceptance of an argument. Similarly, Liao and van der Torre (2020) define *explanation semantics* via a principle-based approach. Likewise, Booth et al. (2014) examine *critical sets* of arguments after whose acceptance the acceptance of the remaining arguments is uniquely determined. The scenario of *abduction*, i. e., determining arguments that explain why some argument is (not) accepted, is considered by Sakama (2018). Notably, they also allow for the introduction of new arguments to the AF for the explanations. To explain non-acceptance, Saribatur, Wallner, and Woltran (2020) establish *strongly rejecting subframeworks* that showcase why some argument is not accepted wrt. different semantics. Complementing the above works, Amgoud (2024) introduces a *post-hoc* approach that discloses relationships between AFs and outputs of a semantics, regardless of its internals and is able to explain more than just (non-)acceptance. Moreover, Doutre, Duchatelle, and Lagasquie-Schiex (2023) study explanations for whole extensions and define multiple classes of explanations to visually explain why a set of arguments is an extension. Similar to our approach, Baumann and Ulbricht (2021) decompose an AF into subframeworks and extensions are constructed via those in order to explain their acceptance. In contrast to our work however, they focus on cycles and their role in the AF for the construction. In a similar fashion, Alfano et al. (2023) introduce a form of structured explanation for probabilistic argumentation frameworks based on the idea of directionality on the strongly connected components. In adjacent fields of knowledge representation, explanations play a similar role. For instance, explaining consequences via minimal subsets of description logic knowledge bases (Baader and Peñaloza 2010), abductive reasoning in logic programming (Kakas, Kowalski, and Toni 1992) or minimal unsatisfiable subsets to explain unsatisfiability of propositional logic formulae (Liffiton and Sakallah 2008). We will however focus on the case of abstract argumentation in this work.

Finally, we will also consider principles for explanation strategies that have been developed in the literature for formally analysing explanation methods (Ulbricht and Wallner 2021). While not technically stated as principles by Borg and Bex (2024), we formulate the notions of relevance, minimality, sufficiency and necessity here as such, since they describe sensible properties of explanations. Note that in the following, while we define the principles on the basis of set-based explanation strategies according to Definition 6, they can of course also be applied to serialisation sequences $\mathcal{S}$ by considering the induced extension $\hat{\mathcal{S}}$. The reason for that is again to ensure comparability of our approach and the above introduced approaches from the literature.

Definition 11. An explanation strategy $\textsc{Expl}$ satisfies the respective principle iff for every AF $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ and $a \in \mathcal{A}$ the respective condition holds:

σ-basic If $E \in \textsc{Expl}(\mathcal{F}, a)$, then there exists $S \in \sigma(\mathcal{F}{\downarrow_E})$ with $a \in S$.

σ-existence If there is some $S \in \sigma(\mathcal{F})$ with $a \in S$, then $\textsc{Expl}(\mathcal{F}, a) \neq \emptyset$.

Monotonicity If $E \in \textsc{Expl}(\mathcal{F}, a)$, then $E' \in \textsc{Expl}(\mathcal{F}, a)$ for any E' with $E \subseteq E' \subseteq \mathcal{A}$.

(Min-)Conflict-Freeness If E is ($\subseteq$-minimal) in $\textsc{Expl}(\mathcal{F}, a)$, then E is conflict-free in $\mathcal{F}$.

Defense If $E \in \textsc{Expl}(\mathcal{F}, a)$, then E defends itself in $\mathcal{F}$.

Independence If $E \in \textsc{Expl}(\mathcal{F}, a)$ and $b \notin E$ with $a \neq b$, then $E \in \textsc{Expl}(\mathcal{F}{\downarrow_{\mathcal{A} \setminus \{b\}}}, a)$.

Relevance If $E \in \textsc{Expl}(\mathcal{F}, a)$, then $E \subseteq \mathsf{Relevant}_{\mathcal{F}}(a)$.

Minimality If $E \in \textsc{Expl}(\mathcal{F}, a)$, then there exists no $E' \in \textsc{Expl}(\mathcal{F}, a)$ with $E' \subsetneq E$.

Sufficiency If $E \in \textsc{Expl}(\mathcal{F}, a)$, then E is relevant to a, conflict-free and E defends $E \cup \{a\}$ in $\mathcal{F}$.

Necessity If $E \in \textsc{Expl}(\mathcal{F}, a)$, then for every $b \in E$ we have that if $b \notin S$ for some $S \in \mathsf{ad}(\mathcal{F})$, then $a \notin S$.

It should be noted that the satisfaction of these principles has so far not been investigated for any of the above mentioned explanation strategies, except partially for the strong σ-explanations of Ulbricht and Wallner (2021). In Section 6, we will complete the principle-based evaluation of the two above introduced approaches and compare them to our sequence explanations. A detailed principle-based analysis of the other approaches is left for future work.

4 Sequence Explanations for Argument Acceptance

We now 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 building blocks (initial sets) until we reach the argument whose acceptance we want to explain. Intuitively, an explanation for the acceptance of an argument a then represents a process of decisions ultimately leading to the acceptance of a.

Definition 12. Let $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ be an AF and $a \in \mathcal{A}$. We define the set of *sequence explanations* $\textsc{SeqEx}(\mathcal{F}, a)$ for the acceptance of a given $\mathcal{F}$ as:

$$\textsc{SeqEx}(\mathcal{F}, a) = \{(S_1, \ldots S_n) \in \mathfrak{S}(\mathcal{F}) \mid a \in S_n\}$$

Example 6. Consider again the AF $\mathcal{F}_2$ depicted in Figure 2. $(\{e\}, \{b\}, \{a\})$ is a sequence explanation for the acceptance of a. Moreover, we have the sequence explanations $(\{e\}, \{b\}, \{g\}, \{a\})$ and $(\{e\}, \{g\}, \{b\}, \{a\})$ both inducing the same admissible set. Note however, that for instance $(\{e\}, \{b\}, \{f\}, \{a\})$ would also be a sequence explanation for a, even though f is not relevant for a.

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 a are actually relevant for the argument a, cf. Definition 7. Nevertheless, as the following result shows, there is still a connection to the necessary explanation of Borg and Bex (2024).

Proposition 2. *Let $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ be an AF and $a \in \mathcal{A}$. Then it holds that $\bigcap_{\mathcal{E} \in \textsc{SeqEx}(\mathcal{F}, a)} \hat{\mathcal{E}} = E$ with $E \in \textsc{Nec}(\mathcal{F}, a)$.*

In order to properly incorporate relevance into the explanations, we refine the definition of an explanation to be a minimal serialisation sequence $(S_1, \ldots S_n)$ such that $a \in S_n$. In other words, such an explanation for a represents a minimal sequence of conflict resolutions that lead to a being acceptable in $\mathcal{F}$.

For that, we define the length of a serialisation sequence $\mathcal{S} = (S_1, \ldots S_n)$ 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 13. Let $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ be an AF and $a \in \mathcal{A}$. We define the set of *minimal sequence explanations* $\textsc{MinSeqEx}(\mathcal{F}, a)$ for the acceptance of a given $\mathcal{F}$ as:

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

Example 7. We consider again the AF $\mathcal{F}_2$ in Figure 2. There are only two minimal sequence explanations for a, namely $(\{e\}, \{b\}, \{a\})$ and $(\{e\}, \{g\}, \{a\})$.

Indeed, including minimality (wrt. the length of the explanation sequence) is already enough to ensure that only relevant arguments are included in the explanation as the following result shows.

Proposition 3. *Let $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ be an AF and $a \in \mathcal{A}$. Then, for every $\mathcal{E} \in \textsc{MinSeqEx}(\mathcal{F}, a)$ we have $\hat{\mathcal{E}} \setminus \{a\} \subseteq \mathsf{Relevant}_{\mathcal{F}}(a)$.*

Beyond that, there is an even closer connection between minimal sequence explanations and the sufficient explanations of Borg and Bex (2024).

Proposition 4. *Let $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ be an AF and $a \in \mathcal{A}$. Then it holds for every $\mathcal{E} \in \textsc{MinSeqEx}(\mathcal{F}, a)$ that $\hat{\mathcal{E}} \in \textsc{Suff}(\mathcal{F}, a)$.*

Conversely, for every (minimal) sufficient explanation there may be multiple sequence explanations, due to the higher expressivity of the sequence-based representation.

We now want to examine under which circumstances an argument should be considered part of an explanation for its own acceptance and in which case it should not. For that, we regard the defense notion. Intuitively, if an argument a is actively involved in its own defense, then it should be part of a proper explanation of its acceptance, otherwise is should not be necessary to include in the explanation. This is also reflected in the notion of relevance by Borg and Bex (cf. Definition 7), i. e., we have that if $a^+_{\mathcal{F}} \cap a^-_{\mathcal{F}} \neq \emptyset$ then $a \in \mathsf{Relevant}_{\mathcal{F}}(a)$. However, this is not incorporated into the sufficient and necessary explanations for the acceptance of a, since those include the argument a explicitly in any case, cf. Definition 9. Moreover, it is not always immediately apparent whether the self-defense of an argument is actually necessary, and thus whether it should be included in an explanation, as shown by the following example.

Example 8. Consider the AF $\mathcal{F}_3$ in Figure 3. While the argument d defends itself against the attacker f, it is also attacked by a. Per definition, the argument c is necessary for the defense of d against a. However, c also defends d against f and thus makes the self-defense of d against f superfluous. Hence, it can be argued that d should not be part of an explanation for its own acceptance.

On the other hand, the argument b is attacked by both a and e. The argument c is again necessary to defend against a, but so is the self-defense of b against e. In this case b should be part of every acceptance explanation for itself, since it actually contributes to its own defense.

Finally, the argument g does not self-defend itself at all and is clearly unnecessary to include in any explanation for its acceptance.

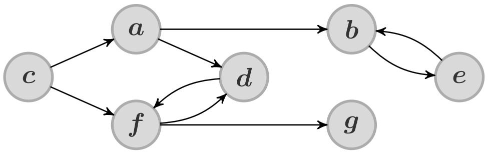

Figure 3: The AF $\mathcal{F}_3$ from Example 8.

To formalise this intuition, we establish the new principle of *Self-Reliance* for acceptance explanations, stating that an explanation for the acceptance of an argument a should only include a, if a actively contributes to its own defense.

Definition 14. Let $\textsc{Expl}$ be an explanation strategy. $\textsc{Expl}$ satisfies *Self-Reliance* iff for every $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ and $a \in \mathcal{A}$ it holds that if $E \in \textsc{Expl}(\mathcal{F}, a)$ and $a \in E$, then $E^-_{\mathcal{F}^{E \setminus \{a\}}} \neq \emptyset$.

In particular, $E^-_{\mathcal{F}^{E \setminus \{a\}}} \neq \emptyset$ states that there must be some attacker of the explanation E after accepting every argument of E except a (represented by the reduct $\mathcal{F}^{E \setminus \{a\}}$). Neither the (minimal) sequence explanations nor any of the other considered approaches satisfy this principle as shown by the following example.

Example 9. Consider again the AF $\mathcal{F}_3$ in Figure 3. We have that $\textsc{Suff}(\mathcal{F}_3, d) = \textsc{MinSuff}(\mathcal{F}_3, d) = \textsc{Nec}(\mathcal{F}_3, d) = \{\{c, d\}\}$. Furthermore, $\{c, d\}$ is also a strong σ-explanation for d and $(\{c\}, \{d\})$ is the only (minimal) sequence explanation for d.

In order to resolve the above described problem, we utilise the procedurality of our explanations together with the distinction of initial sets as described in Definition 3. Meaning, the argument a shall only be included in an acceptance explanation for itself, iff there exists a proper attacker in the final step of the serialisation process where a is accepted, i. e., the initial set S_n with $a \in S_n$ is attacked in $\mathcal{F}^{S_1 \cup \dots \cup S_{n-1}}$. If this is not the case, then $\{a\}$ is an unattacked initial set in the final step and can be detached from the explanation sequence.

Definition 15. Let $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ be an AF and $a \in \mathcal{A}$. We define the set of *strong sequence explanations* $\textsc{StSeqEx}(\mathcal{F}, a)$ for the acceptance of a given $\mathcal{F}$ as follows:

$$\textsc{StSeqEx}(\mathcal{F}, a) = \min_{\sqsubseteq}\{\mathcal{E} = (S_1, \dots S_n) \mid$$
$$\mathcal{E} \in \textsc{MinSeqEx}(\mathcal{F}, a) \wedge S_n \notin \mathsf{is}^{\not\leftarrow}(\mathcal{F}^{S_1 \cup \dots \cup S_{n-1}}) \vee$$
$$(S_1, \dots, S_n, \{a\}) \in \textsc{SeqEx}(\mathcal{F}, a) \text{ with } \{a\} \in \mathsf{is}^{\not\leftarrow}(\mathcal{F}^{\hat{\mathcal{E}}})\}$$

Example 10. We consider again the AF $\mathcal{F}_3$ in Figure 3. The only strong sequence explanation for the acceptance of d is $(\{c\})$, since in the minimal serialisation sequence $(\{c\}, \{d\})$ we have $\{d\} \in \mathsf{is}^{\not\leftarrow}(\mathcal{F}^{\{c\}})$. The same applies to g, where the only strong sequence explanation is also $(\{c\})$. On the other hand, for b the only strong sequence explanation is $(\{c\}, \{b\})$ because $\{b\}$ is challenged initial in the reduct $\mathcal{F}^{\{c\}}$.

Example 11. Consider the AF $\mathcal{F}_4$ in Figure 4. There is only one minimal sequence explanation for the acceptance of g, namely $\mathcal{E}_1 = (\{f\}, \{g\})$. Notably, there are

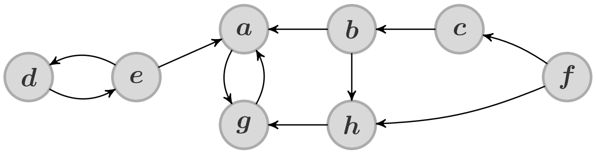

Figure 4: The AF $\mathcal{F}_4$ from Example 11.

also the sequence explanations $\mathcal{E}_2 = (\{f\}, \{b\}, \{g\})$ and $\mathcal{E}_3 = (\{f\}, \{e\}, \{g\})$ as well as $\mathcal{E}_4 = (\{e\}, \{f\}, \{g\})$. While the latter three are not $\sqsubseteq$-minimal, they correspond to the strong sequence explanations, $(\{f\}, \{b\}), (\{f\}, \{e\})$ and $(\{e\}, \{f\})$ respectively, for g, since $\{g\}$ is unattacked initial in the respective reduct in all three sequences. $\mathcal{E}_1$ is also a strong sequence explanation for g, since $\{g\}$ defends itself against a in the reduct $\mathcal{F}_4^{\{f\}}$.

Indeed, this explanation method satisfies the principle of Self-Reliance in general.

Proposition 5. $\textsc{StSeqEx}$ *satisfies* Self-Reliance.

4.1 Restricted Sequence Explanations

Utilising the distinction between initial sets from Definition 3, we refine the sequence explanations to incorporate additional aspects. For instance, we can restrict the sequences to only contain unattacked initial sets, essentially yielding explanations where self-defense is not permitted.

Definition 16. Let $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ be an AF and $a \in \mathcal{A}$. We define the set of *grounded explanations* $\textsc{SeqEx}_{\mathsf{gr}}(\mathcal{F}, a)$ for the acceptance of a given $\mathcal{F}$ as follows:

$$\textsc{SeqEx}_{\mathsf{gr}}(\mathcal{F}, a) = \{\mathcal{E} \in \textsc{SeqEx}(\mathcal{F}, a) \mid \\ \forall S_i \in \mathcal{E} : S_i \in \mathsf{is}^{\nleftarrow}(\mathcal{F}^{S_1 \cup \dots \cup S_{i-1}})\}$$

We can then of course also define $\textsc{MinSeqEx}_{\mathsf{gr}}(\mathcal{F}, a)$ and $\textsc{StSeqEx}_{\mathsf{gr}}(\mathcal{F}, a)$ analogously to Definitions 13 and 15. This restriction is closely related to the grounded and strongly admissible semantics (Baroni, Caminada, and Giacomin 2018; Caminada 2014).

Example 12. Consider again the AF $\mathcal{F}_4$ in Figure 4 and we focus on the argument g. The sequence $(\{f\}, \{e\}, \{g\})$ is *not* a grounded sequence explanation for g, because $\{e\} \in \mathsf{is}^{\leftrightarrow}(\mathcal{F}^{\{f\}})$. Similarly, $(\{f\}, \{g\})$ is not a grounded explanation, because $\{g\}$ requires self-defense in the reduct $\mathcal{F}^{\{f\}}$. The only (minimal) grounded sequence explanation for g is $(\{f\}, \{b\}, \{g\})$, where no self-defense is required for any argument. Accordingly, $(\{f\}, \{b\})$ would be the only strong grounded sequence explanation for g in $\mathcal{F}_4$.

Generally, one can of course also put different restrictions on the explanation sequences. For instance, another reasonable approach would be to weaken the above restriction and allow for both unattacked and unchallenged initial sets. That would be in accordance to the unchallenged semantics (Bengel and Thimm 2022) and essentially represent explanations under which arguments may only self-defend against attackers that are not proper challengers, i. e., arguments that are not admissible themselves.

5 Dialectical Explanations

So far, we have outlined how the sequence explanations implement the procedural aspect of argumentation as a sequence of minimally acceptable sets that essentially support the argument in question. We now turn to the second important element 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. For that, we associate to some sequence explanation $\mathcal{E}_s$, containing the supporting arguments for the acceptance of the argument a, the sequence $\mathcal{E}_d$ of defeated arguments.

Definition 17. Let $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ be an AF and $a \in \mathcal{A}$. We define a *dialectical sequence explanation* for the acceptance of 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 a and for each $i = 1, \dots, n$ we have $T_i = (\hat{\mathcal{E}_s} \cup \{a\})^-_{\mathcal{F}} \cap (S_i)^+_{\mathcal{F}^{S_1 \cup \dots \cup S_{i-1}}}$.

Each T_i is defined to contain the attackers of a and its supporting arguments (represented by $(\hat{\mathcal{E}_S} \cup \{a\})^-_{\mathcal{F}}$) 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 $\mathcal{F}^{S_1 \cup \dots \cup S_{i-1}}$. It can then be shown easily that a dialectical explanation $\mathcal{E}_d$ for the acceptance of a, based on minimal or strong sequence explanations, only contains arguments that are relevant for a.

Proposition 6. *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* $a \in \mathcal{A}$ *with* $(S_1, \dots S_n) \in \textsc{MinSeqEx}(\mathcal{F}, a)$*. It holds that* $T_i \subseteq \mathsf{Relevant}_{\mathcal{F}}(a)$ *and* $T_i \cap T_j = \emptyset$ *for all* $i, j = 0, \dots, n$ *with* $i \neq j$*.*

Example 13. We consider again the AF $\mathcal{F}_4$ in Figure 4. We take the explanation $(\{f\}, \{e\}, \{g\})$ for the acceptance of g. The corresponding sequence of defeated arguments is $\mathcal{E}_d = (\{h\}, \{a, d\}, \emptyset)$. Notice that, while c is attacked by f, it is not included in $\mathcal{E}_d$, because c does not attack any argument of the explanation sequence and thus does not contribute anything to the explanation. Even though g also attacks a, a has already been defeated by e in a previous step of the argumentation process and is therefore not included again. Alternatively, for the strong sequence explanation $(\{f\}, \{b\})$ for g, we have the corresponding defeated sequence $(\{c, h\}, \{a\})$. This time, c is included because it attacks b, which is part of the sequence, but d is no longer relevant to the explanation.

To facilitate the construction of insightful explanations, our approach also allows us to distinguish further for each step 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., their rejection simply follows logically from accepting the initial set, but is not necessary for its acceptance:

$$\mathsf{IncRej}_{\mathcal{F}}(S) = S^+ \setminus S^-$$

Note that $\mathsf{NecRej}_{\mathcal{F}}(S)$ and $\mathsf{IncRej}_{\mathcal{F}}(S)$ are disjunct and we have that the necessarily and incidentally rejected arguments by some admissible $S \subseteq \mathcal{A}$ characterise exactly the arguments attacked by S.

Proposition 7. *Let $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ be an AF and $S \in \mathit{ad}(\mathcal{F})$. Then the following statements hold:*

1. $\mathsf{NecRej}_{\mathcal{F}}(S) \cap \mathsf{IncRej}_{\mathcal{F}}(S) = \emptyset$,
2. *If $S \in \mathit{ad}(\mathcal{F})$, then* $\mathsf{NecRej}_{\mathcal{F}}(S) \cup \mathsf{IncRej}_{\mathcal{F}}(S) = S^{+}_{\mathcal{F}}$.

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

Example 14. We continue Example 13 with the AF $\mathcal{F}_4$ in Figure 4. Consider the dialectical sequence explanation $(\mathcal{E}_s, \mathcal{E}_d)$ for the acceptance of g with $\mathcal{E}_s = (\{f\}, \{e\}, \{g\})$ and $\mathcal{E}_d = (\{h\}, \{a, d\}, \emptyset)$. We examine, step by step, the defeated attackers of the explanation: h, d, a. First, we have that $\mathsf{IncRej}_{\mathcal{F}_4}(\{f\}) = \{h\}$. Furthermore, we have $\mathsf{NecRej}_{\mathcal{F}_4^{\{f\}}}(\{e\}) = \{d\}$. On the other hand, we have $\mathsf{IncRej}_{\mathcal{F}_4^{\{f\}}}(\{e\}) = \{a\}$. Meaning essentially, that h and a are merely weak counterarguments and d is a strong counterargument, in the context of this sequence. If we consider instead the sequence explanation $(\{f\}, \{g\})$, with the defeated arguments $(\{h\}, \{a\})$, h is again a weak counterargument, but a is now a strong contender, since it is necessarily rejected by g in this sequence.

Let $\textsc{DiSeqEx}(\mathcal{F}, a)$ denote the set of dialectical sequence explanations $(\mathcal{E}_s, \mathcal{E}_d)$ for the acceptance of a in $\mathcal{F}$, such that $\mathcal{E}_s \in \textsc{StSeqEx}(\mathcal{F}, a)$. Even though the dialectical sequence explanations can be built on any type of sequence explanation, we will only consider the variant that is based on the strong sequence explanations in the following, since those are the only explanations that satisfy the principle of Self-Reliance. For some dialectical sequence explanation $(\mathcal{E}_s, \mathcal{E}_d)$, we denote with $\hat{\mathcal{E}} = \hat{\mathcal{E}_s} \cup \hat{\mathcal{E}_d}$ the corresponding set representation.

Let us consider again the two key aspects of dialectical argumentation mentioned in the introduction: procedurality and the exchange of arguments. To formally capture the exchange of arguments aspect, we now introduce the principle of *Dialectical Completeness* for acceptance explanations.

Definition 18. Let $\textsc{Expl}$ be an explanation strategy. $\textsc{Expl}$ satisfies *Dialectical Completeness* iff for every $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ and $a \in \mathcal{A}$ it holds that if $E \in \textsc{Expl}(\mathcal{F}, a)$, then there is some admissible $E' \subseteq E \cup \{a\}$ such that $a \in E'$ and $E'^{-}_{\mathcal{F}} \subseteq E$.

This property essentially ensures two things for an acceptance explanations. On the one hand, it ensures that each explanation E contains an admissible "core" set of arguments E' around the query a. On the other hand, it requires that the explanation also takes into account all the counterarguments of at least this core set. In combination, this then ensures that the explanation gives a complete dialectical picture of the acceptance of a.

Example 15. We continue Example 13 with the AF $\mathcal{F}_4$ depicted in Figure 4. Consider the dialectical sequence explanation $(\mathcal{E}_s, \mathcal{E}_d)$ for the acceptance of g with

$$\mathcal{E}_s = (\{f\}, \{e\}, \{g\}) \quad \text{and} \quad \mathcal{E}_d = (\{h\}, \{a, d\}, \emptyset).$$

The corresponding set representation is then $E = \{a, d, e, f, g, h\}$. It is easy to see that we have the admissible subset $E' = \hat{\mathcal{E}_s} = \{e, f, g\}$ which contains the argument g. Furthermore, all attackers of E' in $\mathcal{F}_4$, namely a, d and h, are contained in E. It is also apparent in this example that a set-based representation loses a lot of expressivity compared to our sequence-based representation.

As the following result shows, the dialectical sequence explanations satisfy both Self-Reliance and Dialectical Completeness.

Proposition 8. $\textsc{DiSeqEx}$ *satisfies* Self-Reliance *and* Dialectical Completeness.

6 Discussion

Our sequence explanations provide a new form of explanation for the acceptance of arguments that incorporate both the procedural and dialectical aspect of argumentation. In a similar manner, discussion games provide rule-based dialogues between two players for the acceptance of an argument (Caminada 2018). In contrast to our approach, a discussion game always starts with the argument in question and in every step only individual arguments are played. In particular, they also allow arguments to be repeated. This redundancy may not happen in sequence explanations, as we consider sets of arguments in each step and no argument can occur twice in a dialectical sequence explanation. Discussion games generally also have no mechanism to determine the strength of attackers or to capture the idea of the self-reliance property. The same applies also to dispute trees, where acceptance is determined by considering a tree-like dialogue representation (Cyras et al. 2017). Furthermore, Baroni, Giacomin, and Guida (2005) introduce a scheme for recursively constructing extensions along the strongly connected components (SCCs) of an AF similar to serialisability, but their approach enforces the construction order based on the SCC-structure of the AF. That is very similar to the approach of Alfano et al. (2023), where they construct structural explanations for acceptance in probabilistic AFs by constructing σ-extensions according to the SCC-structure. Notably, the sequences both of these approaches obtain for some σ-extension of an AF consist of smaller σ-extensions of the SCCs of the AF. That is critically different to our approach, where a sequence explanation is comprised of initial sets. These initial sets are the minimal semantical units and they are crucial in ensuring that the explanation is concise and relevant for the argument in question. In addition to that, our dialectical explanations also incorporate counterarguments into the explanation, which neither of these two approaches consider.

Another important advantage of our approach is that SEQEX, MINSEQEX and STSEQEX are generally independent of semantics and built only on the concept of admissibility. Consider, for instance, complete and preferred semantics. They are both built on admissibility, but have additional requirements for extensions: the inclusion of all defended arguments and $\subseteq$-maximality, respectively. When considering the acceptance of some argument a, both of these conditions are completely irrelevant for explaining the acceptance of a. In particular, somehow being dependent on $\subseteq$-maximality for the explanation of the acceptance of a will necessarily lead to the inclusion of non-relevant arguments. Notably, the STSEQEX$_{\mathsf{gr}}$ explanations are closely tied to grounded semantics, where the notion of self-defense is disallowed, which can be a reasonable restriction for acceptance explanations.

We conducted a principle-based analysis to provide a formal and objective comparison between our approach and existing approaches based on the principles of Ulbricht and Wallner (2021) and Borg and Bex (2024) introduced in Section 3, as well as the newly defined Self-Reliance and Dialectical Completeness properties. For that, we considered the explanation strategies introduced in Section 3 as well as the different variants of sequence explanations defined in Sections 4 and 5. Please note again, that the principles are defined on set-based explanations to ensure better comparability between different approaches, since the set representation is the most general.

In Theorem 1 and Table 1, we summarise the results of our principle-based analysis of explanation methods.

Theorem 1. *Let σ be a semantics. The compliance of explanation strategies* SEQEX*,* MINSEQEX*,* STSEQEX*,* STSEQEX$_{gr}$*,* DISEQEX*,* SUFF*,* NEC*,* MINSUFF *and* STRONG$_\sigma$ *wrt. the properties σ-basic, σ-existence, Monotonicity, Min-CF, Defense, Independence, Relevance, Minimality, Sufficiency, Necessity, Self-Reliance and Dialectical Completeness is as shown in Table 1.*

σ-existence is satisfied by all explanation strategies. In contrast to the other variants of sequence explanations, STSEQEX does not satisfy σ-basic for any semantics σ. This is simply due to how the principle is defined: it requires that there exists a σ-extension in the projection onto the explanation that contains the argument. That is obviously not possible if the argument is not contained in the explanation. The same applies to STSEQEX$_{\mathsf{gr}}$ and DISEQEX, which are both based on strong sequence explanations.

Independence is only satisfied by SEQEX but no other variant of sequence explanations. The reason is simply that Relevance and Independence are incompatible, i. e., the set of relevant arguments $\mathsf{Relevant}_{\mathcal{F}}(a)$ may change if we remove some argument from the AF. In a similar way, Monotonicity is trivially violated if Relevance is satisfied.

Due to the fact that STSEQEX$_{\mathsf{gr}}$ explanations are defined to only consist of unattacked initial sets, which rule out self-defense of arguments, they naturally satisfy Minimality in contrast to the basic variants of sequence explanations (see Example 16). As one might expect, their existence is however only guaranteed if the argument is part of the grounded extension of the AF.

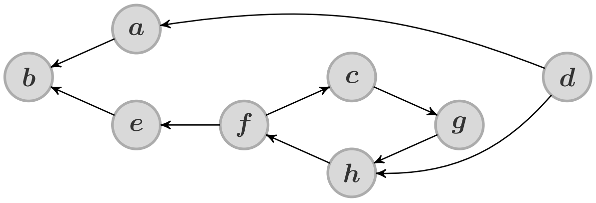

Figure 5: The AF $\mathcal{F}_5$ from Example 16.

Example 16. Consider the argument b in the AF $\mathcal{F}_5$ depicted in Figure 5. A strong sequence explanation for the acceptance of b is the sequence $(\{d\},\{f\})$, because d defends b against a and in the reduct $\mathcal{F}_5^{\{d\}}$ we have the initial set $\{f\}$ which defends b against the remaining attack from e. However, there is the sequence $(\{f,g\},\{d\})$, which is a minimal sequence explanation for the acceptance of b. Note that for the corresponding admissible sets we have $\{d,f\} \subsetneq \{d,f,g\}$, which shows that the STSEQEX explanations are not $\subseteq$-minimal. This stems from the fact that these explanations are rather built on the *minimum number of decisions* to take for the acceptance of the argument. In the above case, it is possible to accept the defender f via the set $\{f,g\}$ without accepting d first, because the defense of f by d against h is not necessary. This is also reflected in the strong sequence explanations for the acceptance of f: $(\{d\})$ and $(\{f,g\})$.

The dialectical sequence explanations (DISEQEX) are the only explanations that satisfy both Self-Reliance and Dialectical Completeness. Compared to STSEQEX, the satisfaction of Dialectical Completeness comes at the price of no longer satisfying Min-CF, Defense and Sufficiency. This is to be expected, since the inclusion of counterarguments into the explanation will introduce conflicts. It should however be noted that the proper sequence-based representation of dialectical sequence explanations, as described in Section 5, gives a clear separation between arguments that support the target and the counterarguments.

As the previous discussion already hinted at, the goal of this principle-based comparison is not to satisfy as many principles as possible. Rather, it is supposed to provide objective criteria to compare different approaches in order to find the one that is most suitable for a specific use case. Furthermore, we would like to highlight again that these principles are generally tailored to set-based explanations, meaning the advantages of our process-based approach and the dialectical aspect are not necessarily visible just considering principle satisfaction. Only the Self-Reliance and Dialectical Completeness principles somehow capture the two aspects of dialectical argumentation outlined by Hage (2000). 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.

	SeqEx	MinSeqEx	StSeqEx	StSeqEx$_{\mathsf{gr}}$	DiSeqEx	Suff	Nec	MinSuff	Strong$_\sigma$
σ-basic	✓	✓	✗	✗	✗	✓	✓	✓	✓*
σ-existence	✓	✓	✓	gr	✓	✓	✓	✓	✓*
Monotonicity	✗	✗	✗	✗	✗	✗	✗	✗	✓*
Min-CF	✓	✓	✓	✓	✗	✓	✓	✓	✗*
Defense	✓	✓	✓	✓	✗	✓	✗	✓	✗*
Independence	✓	✗	✗	✗	✗	✗	✗	✗	✓*
Relevance	✗	✓	✓	✓	✓	✓	✓	✓	✗
Minimality	✗	✗	✗	✓	✗	✗	✓	✓	✗
Sufficiency	✗	✓	✓	✓	✗	✓	✗	✓	✗
Necessity	✗	✗	✗	✗	✗	✗	✓	✗	✗
Self-Reliance	✗	✗	✓	✓	✓	✗	✗	✗	✗
Dia. Complet.	✗	✗	✗	✗	✓	✗	✗	✗	✗

Table 1: Overview over the discussed explanation strategies and their satisfaction of principles for explanations from the literature. All results consider $\sigma \in \{\mathsf{ad}, \mathsf{co}, \mathsf{gr}, \mathsf{pr}, \mathsf{st}\}$. All proofs can be found in the appendix. Results marked with * are from (Ulbricht and Wallner 2021). All other results are new.

7 Conclusion

In this work, we introduced the notion of sequence explanations as a procedural form of explanation for the acceptance of arguments. Such an explanation is then essentially a sequence of minimally acceptable sets of arguments that lead to the acceptance of the argument in question. We discussed several variants which ensure, for instance, that only relevant arguments are included or pose restrictions on what kind of defense is permitted. Furthermore, we expanded the sequence explanations to also include the corresponding relevant counterarguments to provide a full dialectical argumentative explanation for the acceptance of arguments. More specifically, our dialectical sequence explanations are the only explanation strategy that capture both the procedurality and exchange of arguments of dialectical argumentation. Moreover, this approach also gives a fine-grained view into the strength of counterarguments. Finally, we evaluated our approach based on principles from the literature and discussed its advantages over existing approaches from the literature.

For future work, we intend to develop a proper representation and visualisation of the (dialectical) sequence explanations. In particular, we want this representation to be understandable for non-experts and we plan to evaluate their effectiveness in an empirical study. Related to that, developing algorithms to efficiently compute the sequence explanations as well as investigating the computational complexity is another important direction for future work. Moreover, we also want to extend our approach to other formal argumentation approaches. Of particular interest is, for instance, assumption-based argumentation (Dung, Kowalski, and Toni 2009), where arguments are formed by sets of assumptions that together imply a conclusion. In this domain, a sequence explanation would essentially represent the process of accepting (and rejecting) assumptions in order to accept a specific assumption. Another interesting target formalism are abstract dialectical frameworks (Brewka et al. 2017), a powerful generalisation of argumentation frameworks, especially since serialisation sequences have recently been introduced to this domain and are

Acknowledgements

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

References

Alfano, G.; Calautti, M.; Greco, S.; Parisi, F.; and Trubitsyna, I. 2023. Explainable acceptance in probabilistic and incomplete abstract argumentation frameworks. *Artif. Intell.* 323:103967.

Amgoud, L. 2024. Post-hoc explanation of extension semantics. In *ECAI 2024 - 27th European Conference on Artificial Intelligence, 2024*, volume 392 of *Frontiers in Artificial Intelligence and Applications*, 3276–3283. IOS Press.

Antaki, C., and Leudar, I. 1992. Explaining in conversation: Towards an argument model. *European Journal of Social Psychology* 22(2):181–194.

Atkinson, K.; Bench-Capon, T. J. M.; and Bollegala, D. 2020. Explanation in AI and law: Past, present and future. *Artif. Intell.* 289:103387.

Baader, F., and Peñaloza, R. 2010. Automata-based axiom pinpointing. *J. Autom. Reason.* 45(2):91–129.

Baroni, P.; Caminada, M.; and Giacomin, M. 2018. Abstract argumentation frameworks and their semantics. In Baroni, P.; Gabbay, D.; Giacomin, M.; and van der Torre, L., eds., *Handbook of Formal Argumentation*. College Publications. 159–236.

Baroni, P.; Giacomin, M.; and Guida, G. 2005. Scc-recursiveness: a general schema for argumentation semantics. *Artif. Intell.* 168(1-2):162–210.

Baumann, R., and Ulbricht, M. 2021. 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*, 110–119.

Baumann, R.; Brewka, G.; and Ulbricht, M. 2020. 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*, 2742–2749. AAAI Press.

Bengel, L., and Thimm, M. 2022. Serialisable semantics for abstract argumentation. In Toni, F.; Polberg, S.; Booth, R.; Caminada, M.; and Kido, H., eds., *Computational Models of Argument - Proceedings of COMMA 2022*, 80–91. IOS Press.

Bengel, L., and Thimm, M. 2025. *Sequence Explanations for Acceptance in Abstract Argumentation (Extended Version)*. Zenodo.

Bengel, L.; Sander, J.; and Thimm, M. 2024. Characterising serialisation equivalence for abstract argumentation. In *ECAI 2024 - 27th European Conference on Artificial Intelligence, 2024*, volume 392 of *Frontiers in Artificial Intelligence and Applications*, 3340–3347. IOS Press.

Bengel, L. 2022. On serialisability for argumentative explanations. *Online Handbook of Argumentation for AI: Volume 3* abs/2212.07996.

Blümel, L., and Thimm, M. 2022. A ranking semantics for abstract argumentation based on serialisability. In Toni, F.; Polberg, S.; Booth, R.; Caminada, M.; and Kido, H., eds., *Computational Models of Argument - Proceedings of COMMA 2022*, 104–115. IOS Press.

Booth, R.; Caminada, M.; Dunne, P. E.; Podlaszewski, M.; and Rahwan, I. 2014. Complexity properties of critical sets of arguments. In Parsons, S.; Oren, N.; Reed, C.; and Cerutti, F., eds., *Computational Models of Argument - Proceedings of COMMA 2014*, volume 266 of *Frontiers in Artificial Intelligence and Applications*, 173–184. IOS Press.

Borg, A., and Bex, F. 2021. Necessary and sufficient explanations for argumentation-based conclusions. In Vejnarová, J., and Wilson, N., eds., *Symbolic and Quantitative Approaches to Reasoning with Uncertainty - 16th European Conference, ECSQARU 2021*, volume 12897 of *Lecture Notes in Computer Science*, 45–58. Springer.

Borg, A., and Bex, F. 2024. Minimality, necessity and sufficiency for argumentation and explanation. *Int. J. Approx. Reason.* 168:109143.

Brewka, G.; Ellmauthaler, S.; Strass, H.; Wallner, J. P.; and Woltran, S. 2017. Abstract dialectical frameworks. an overview. *Handbook of Formal Argumentation* 1.

Caminada, M. 2014. Strong admissibility revisited. In Parsons, S.; Oren, N.; Reed, C.; and Cerutti, F., eds., *Computational Models of Argument - Proceedings of COMMA 2014*, volume 266 of *Frontiers in Artificial Intelligence and Applications*, 197–208. IOS Press.

Caminada, M. 2018. Argumentation semantics as formal discussion. In Baroni, P.; Gabbay, D.; Giacomin, M.; and van der Torre, L., eds., *Handbook of Formal Argumentation, Vol.1*. College Publications. 487–518.

Cyras, K.; Fan, X.; Schulz, C.; and Toni, F. 2017. Assumption-based argumentation: Disputes, explanations, preferences. *IFCoLog Journal of Logics and Their Applications* 4(8):2407.

Cyras, K.; Rago, A.; Albini, E.; Baroni, P.; and Toni, F. 2021. Argumentative XAI: A survey. In Zhou, Z., ed., *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021*, 4392–4399. ijcai.org.

Doutre, S.; Duchatelle, T.; and Lagasquie-Schiex, M. 2023. Classes of explanations for the verification problem in abstract argumentation. In Bouraoui, Z.; Schwarzentruber, F.; and Wilczynski, A., eds., *17èmes Journées d'Intelligence Artificielle Fondamentale, JIAF 2023*, 124–134.

Dung, P. M.; Kowalski, R. A.; and Toni, F. 2009. Assumption-based argumentation. In Simari, G. R., and Rahwan, I., eds., *Argumentation in Artificial Intelligence*. Springer. 199–218.

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.

Fan, X., and Toni, F. 2014. On computing explanations in abstract argumentation. In Schaub, T.; Friedrich, G.; and O'Sullivan, B., eds., *ECAI 2014 - 21st European Conference on Artificial Intelligence*, volume 263 of *Frontiers in Artificial Intelligence and Applications*, 1005–1006. IOS Press.

Hage, J. 2000. Dialectical models in artificial intelligence and law. *Artificial Intelligence and Law* 8(2/3):137–172.

Kakas, A. C.; Kowalski, R. A.; and Toni, F. 1992. Abductive logic programming. *J. Log. Comput.* 2(6):719–770.

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.; and Toni, F. 2024. Contestable AI needs computational argumentation. *CoRR* abs/2405.10729.

Liao, B., and van der Torre, L. 2020. Explanation semantics for abstract argumentation. In *Computational Models of Argument - Proceedings of COMMA 2020*, volume 326 of *Frontiers in Artificial Intelligence and Applications*, 271–282. IOS Press.

Liffiton, M. H., and Sakallah, K. A. 2008. Algorithms for computing minimal unsatisfiable subsets of constraints. *J. Autom. Reason.* 40(1):1–33.

Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* 267:1–38.

Potyka, N.; Yin, X.; and Toni, F. 2022. Explaining random forests using bipolar argumentation and markov networks (technical report). *CoRR* abs/2211.11699.

Rescher, N. 1977. *Dialectics: A controversy-oriented approach to the theory of knowledge*. Suny Press.

Sakama, C. 2018. Abduction in argumentation frameworks. *J. Appl. Non Class. Logics* 28(2-3):218–239.

Saribatur, Z. G.; Wallner, J. P.; and Woltran, S. 2020. Explaining non-acceptability in abstract argumentation. In *ECAI 2020 - 24th European Conference on Artificial Intelligence*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, 881–888. IOS Press.

Seselja, D., and Straßer, C. 2013. Abstract argumentation and explanation applied to scientific debates. *Synth.* 190(12):2195–2217.

Thimm, M. 2022. Revisiting initial sets in abstract argumentation. *Argument & Computation* 13(3):325–360.

Ulbricht, M., and Wallner, J. P. 2021. Strong explanations in abstract argumentation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021*, 6496–6504. AAAI Press.

Xu, Y., and Cayrol, C. 2018. Initial sets in abstract argumentation frameworks. *Journal of Applied Non-Classical Logics* 28(2-3):260–279.