<http://www.diva-portal.org>

Postprint

This is the accepted version of a paper presented at *CLAR 2021 - Fourth International Conference on Logic and Argumentation, Hybrid (virtual and physical), via Hangzhou, China, October 20-22, 2021*.

Citation for the original published paper:

Kampik, T., Čyras, K. (2021)

Explanations of Non-monotonic Inference in Admissibility-Based Abstract Argumentation

In: Pietro Baroni, Christoph Benzmüller and Yi N. Wang (ed.), *Logic and Argumentation: 4th International Conference, CLAR 2021, Hangzhou, China, October 20–22, 2021, Proceedings* (pp. 209-223). Cham: Springer

Lecture Notes in Computer Science (LNCS)

https://doi.org/10.1007/978-3-030-89391-0_12

N.B. When citing this work, cite the original published paper.

Permanent link to this version:

<http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-188626>

Explanations of Non-Monotonic Inference in Admissibility-based Abstract Argumentation

Timotheus Kampik¹ and Kristijonas Čyras²

¹ Umeå University, Sweden

² Ericsson Research, Sweden

tkampik@cs.umu.se, kristijonas.cyras@ericsson.com

Abstract. In this paper, we introduce a formal framework for explaining change of inference in abstract argumentation, in particular in the context of iteratively drawing inferences from a sequence of normal expansions, with a focus on admissible set-based semantics. We then conduct a formal analysis, showing that given an initial argumentation framework and an extension that has been inferred from it, we can guarantee the existence of explanation arguments for the violation of monotony when inferring an extension from a normal expansion of the initial argumentation framework.

Keywords: Formal Argumentation · Explainable Artificial Intelligence · Non-monotonic Reasoning.

1 Introduction

Recently, formal argumentation approaches have received increasing attention in the Artificial Intelligence (AI) community, partly because argumentation is considered an enabler of *explainable* AI. However, claiming that argumentation-based inference is in itself always explainable is naive; indeed, a range of approaches that facilitate explainability in argumentation have been introduced in recent years [26,11]. In this paper, we advance the research direction of explainable argumentation by introducing a formal approach that explains the violation of monotony in *sequential* argumentation, where we draw inferences from an argumentation framework, then update it with new arguments and attacks, draw new inferences and so on. In particular, we introduce the notion of *monotony violation explanations*, *i.e.* sets of arguments that can explain the violation of monotony, given an argumentation semantics that satisfies a newly introduced relaxed monotony principle; we prove that three of Dung’s initial admissible set-based semantics³ satisfy this principle. With this approach, our work extends the body of research on explainable formal argumentation by enabling the explanation of dynamic inference.

Let us provide an example that gives an intuition of the approach this paper introduces.

³ The exception among the four semantics that Dung introduces in his seminal paper is stable semantics, which is not universally defined.

Example 1. Assume we have a machine reasoner $\mathcal{R}$ that draws inferences from a belief base (that may contain conflicting beliefs). The belief base *expands* with time, and $\mathcal{R}$ draws new inferences after every expansion. Several IT systems subscribe to $\mathcal{R}$'s inferences and update their configurations based on the conclusions that $\mathcal{R}$ serves. Updating a configuration may be costly (imply down-time, manual effort, and the re-routing of workflow instances); hence, $\mathcal{R}$ needs to provide an *explanation* every time *monotony is violated*, *i.e.* when a previous recommendation about a system configuration is withdrawn.

For instance, let us assume an abstract argumentation-based model, in which $\mathcal{R}$ starts with the initial argumentation framework $AF = (AR, AT)$ – where AR is a set of statements (“arguments”) and $AT \subseteq AR \times AR$ (“attacks”) – such that $AF = (\{a\}, \{\})$; *i.e.*, we have only one statement and no attacks. From this framework, $\mathcal{R}$ infers⁴ $\{a\}$. Then, AF *normally expands* to $AF' = (AR', AT') = (\{a, b, c, d\}, \{(a, b), (b, a)\})$: new arguments and attacks are added to AF , but the attacks between arguments in AR remain unaffected⁵. From AF' , $\mathcal{R}$'s inference function allows to infer either $\{a, c, d\}$ or $\{b, c, d\}$. $\mathcal{R}$ opts to infer $\{a, c, d\}$ because it is monotonic w.r.t. the previous inference $\{a\}$ and no explanations need to be provided.

However, things get more tricky when AF' is normally expanded to $AF'' = (AR'', AT'') = (\{a, b, c, d, e, f, g\}, \{(a, b), (b, a), (e, c), (e, d), (e, f), (f, a), (f, e), (f, g), (g, f)\})$. $\mathcal{R}$'s inference function allows to infer either $\{a, e, g\}$, or $\{b, e, g\}$, or $\{b, c, d, f\}$. Now, $\mathcal{R}$ is forced to violate monotony in order to comply with the constraints the inference function imposes. Intuitively, a reasonable decision is to infer $\{b, c, d, f\}$ because it minimizes the number of arguments (w.r.t cardinality) that are part of the previous conclusion but not part of the current conclusion, *i.e.* $|\{a, c, d\} \setminus \{b, c, d, f\}| < |\{a, c, d\} \setminus \{a, e, g\}| < |\{a, c, d\} \setminus \{b, e, g\}|$. (While more nuanced variants of monotony-maximizing extension-selection are introduced in [19], in this paper, we focus on cardinality-based monotony maximization with respect to a particular previous inference, which – self-evidently – may but does not necessarily have to be the most recently drawn inference from a temporal perspective.) To *explain* the violation of monotony, $\mathcal{R}$ may proceed by highlighting *monotony violation explanations*, *i.e.* arguments that satisfy the following constraints:

- The argument has been newly added.
- The argument attacks an argument that was in the previous conclusion.
- The argument is not attacked by any subset of arguments in the new conclusion that defends itself against all attackers and is not in conflict with the previous conclusion.

⁴ In this example, the behavior of $\mathcal{R}$'s inference function coincides with preferred semantics [13], to be defined later.

⁵ We focus on normal expansions because we consider it a reasonable assumption that in an argumentation context, dynamic scenarios are modeled by adding arguments to an argumentation framework without deleting arguments (instead, arguments can be defeated) and without changing the attack relations between existing arguments.

Figure 1 depicts the example’s argumentation frameworks. The *monotony violation explanations* of the conclusion $\{b, c, d, f\}$ as inferred from AF'' w.r.t. the previous conclusion $\{a, c, d\}$ that has been inferred from AF' are the arguments comprising the set $\{e, f\}$. Intuitively, the fact that we infer f explains that we can no longer infer a ; in turn, e explains why we have to infer f , *i.e.* to still be able to infer c and d .

Fig. 1: Explaining the violation of monotony. Here and henceforth, the gray arguments in each argumentation framework comprise the inferred extension; arguments with a dashed border are rejected and arguments with a bold border are monotony violation explanations.

The rest of this paper is organized as follows. Section 2 provides the necessary theoretical preliminaries. Then, Section 3 introduces and analyzes the monotony violation explainability framework. An implementation of the framework is described in Section 4. Section 5 discusses the framework in the context of related work, before Section 6 concludes the paper.

2 Preliminaries

The central model that this paper makes use of is the notion of an abstract argumentation framework.

Definition 1 (Argumentation Framework [13]). *An argumentation framework AF is a tuple (AR, AT) , such that AR is a set of elements (called arguments) and $AT \subseteq AR \times AR$ (called attacks).*

We assume that the arguments in an argumentation framework are finite. Given an argumentation framework $AF = (AR, AT)$, we say for two arguments $a, b \in AR$ that a attacks b if and only if $(a, b) \in AT$. Given $S \subseteq AR$, we say that

S attacks a if and only if there exists an argument $c \in S$, such that c attacks a and we say that a attacks S if and only if there exists an argument $d \in S$, such that a attacks d ; we say that S attacks $P \subseteq AR$ if and only if there exists an argument $e \in S$, such that e attacks P . We say that S defends a if and only if for every argument $g \in AR$, such that g attacks a it holds true that S attacks g .

In abstract argumentation, admissible and conflict-free sets are important notions.

Definition 2 (Conflict-free and Admissible Sets [13]). Let $AF = (AR, AT)$ be an argumentation framework. A set $S \subseteq AR$:

- is conflict-free (in AF) iff $\nexists a, b \in S$ such that a attacks b ;
- is admissible (in AF) iff S is conflict-free and $\forall a \in S$, it holds true that S defends a .

Let us now introduce some argumentation semantics, i.e. functions that infer sets of *extensions* from an argumentation framework, where each extension is a subset of the argumentation framework's arguments.

Definition 3 (Dung's Admissibility-based Semantics [13]). Let $AF = (AR, AT)$ be an argumentation framework. An admissible set $S \subseteq AR$ is a:

- stable extension of AF iff S attacks each argument that does not belong to S . Stable semantics $\sigma_{st}(AF)$ denotes all stable extensions of AF ;
- complete extension of AF iff each argument that is defended by S belongs to S . Complete semantics $\sigma_{co}(AF)$ denotes all complete extensions of AF ;
- preferred extension of AF iff S is a maximal (w.r.t. set inclusion) admissible subset of AR . Preferred semantics $\sigma_{pr}(AF)$ denotes all preferred extensions of AF ;
- grounded extension of AF iff S is the minimal (w.r.t. set inclusion) complete extension of AF . Grounded semantics $\sigma_{gr}(AF)$ denotes all grounded extensions of AF .

Given an argumentation semantics σ and an argumentation framework AF , we call every $E \in \sigma(AF)$ a σ -extension of AF . Colloquially speaking, argumentation framework *expansions* constrain how an argumentation framework is manipulated: (simple) expansions specify that no arguments or attacks can be removed, whereas normal expansions demand that in addition, no new attacks can be added between existing arguments.

Definition 4 (Argumentation Framework Expansions [4]). Let $AF = (AR, AT)$ and $AF' = (AR', AT')$ be argumentation frameworks.

- AF' is an expansion of AF (denoted by $AF \preceq AF'$) iff $AR \subseteq AR'$ and $AT \subseteq AT'$.
- AF' is a normal expansion of AF (denoted by $AF \preceq_N AF'$) iff $AF \preceq AF'$ and $(AR \times AR) \cap (AT' \setminus AT) = \emptyset$.

To facilitate the design and analysis of argumentation semantics, formal argumentation *principles* have been introduced in the literature [25]. The universality principle describes whether an argumentation semantics always returns at least one extension [3].

Definition 5 (Universality). *Let σ be an argumentation semantics. σ is universally defined (σ satisfies universality) iff for every argumentation framework AF it holds true that $|\sigma(AF)| \geq 1$.*

A particularly relevant principle in the context of this work is *weak cautious monotony*, which stipulates that normally expanding an argumentation framework must allow us to infer any set of arguments contained in any of the original argumentation framework's extensions, given this extension is not attacked by any of the newly added arguments.

Definition 6 (Weak Cautious Monotony [20]). *Let σ be an argumentation semantics. σ satisfies weak cautious monotony iff for every two argumentation frameworks $AF = (AR, AT)$ and $AF' = (AR', AT')$, such that $AF \preceq_N AF'$, and $\forall E \in \sigma(AF)$, it holds true that if $\{(a, b) \mid (a, b) \in AT', a \in AR' \setminus AR, b \in E\} = \emptyset$ then $\exists E' \in \sigma(AF')$ such that $E \subseteq E'$.*

As a measure of how much an inference result violates monotony w.r.t. a previous inference, we have defined the notion of a *degree of monotony* in a recent work.

Definition 7 (Degree of Monotony [19]). *Let E and E' be two finite sets of arguments. We define the degree of monotony of E' w.r.t. E , denoted by $deg_{mon}(E', E)$, as follows:*

$$deg_{mon}(E', E) = \begin{cases} 1 & \text{if } |E| = 0; \\ \frac{|E' \cap E|}{|E|} & \text{otherwise.} \end{cases}$$

The degree of monotony allows us to guide extension selection when first picking an extension that a semantics infers from an argumentation framework and then determining the extensions a semantics returns from an update of this argumentation framework, such that these extensions maximize the degree of monotony w.r.t. the previously inferred extension.

Definition 8 (Degree of Monotony-Maximizing Extensions [19]).

Let $AF' = (AR', AT')$ be an argumentation framework. Let σ be a universally defined argumentation semantics, and let E be a finite set of arguments. We define the monotony-maximizing σ -extensions of AF' w.r.t. E , denoted by $Ext_{mon}(E, AF', \sigma)$, as $\{E' \mid E' \in \sigma(AF'), \nexists E'' \in \sigma(AF'), \text{ such that } deg_{mon}(E', E) < deg_{mon}(E'', E)\}$.

Let us illustrate the notion of degree of monotony-maximizing extensions by introducing an example.

Example 2. Consider the argumentation frameworks $AF = (AR, AT) = (\{a, b, c\}, \{(a, b), (b, a)\})$ and $AF' = (AR', AT') = (\{a, b, c, d\}, \{(a, b), (b, a), (d, c)\})$, as well as preferred semantics. Note that $AF \preceq_N AF'$. $\sigma_{pr}(AF) = \{\{a, c\}, \{b, c\}\}$ and $\sigma_{pr}(AF') = \{\{a, d\}, \{b, d\}\}$. Let us assume we first select $E = \{a, c\}$ from $\sigma_{pr}(AF)$ and then normally expand AF to AF' . No matter which extension we select from $\sigma_{pr}(AF')$, we will violate monotony. However, intuitively, it does not make sense to “switch” from an inference that entails a to an inference that entails b . The degree of monotony-maximizing extensions support this intuition: $deg_{mon}(\{a, d\}, E) = \frac{1}{2}$ and $deg_{mon}(\{b, d\}, E) = 0$; $Ext_{mon}(E, AF', \sigma_{pr}) = \{\{a, d\}\}$.

3 Explainability Framework

Let us now provide a formal explainability framework for the intuition we have outlined in the introduction. We call the central principle of our framework the Relaxed Monotony Explainability (RME) principle, as it helps us explain the relaxation (or: the constrained violation) of monotony.

Definition 9 (Relaxed Monotony Explainability (RME) Principle). *Let σ be an argumentation semantics. σ satisfies the Relaxed Monotony Explainability (RME) principle iff for every two argumentation frameworks $AF = (AR, AT)$ and $AF' = (AR', AT')$, such that $AF \preceq_N AF'$, the following statement holds true:*

$\forall E \in \sigma(AF), \exists E' \in \sigma(AF'),$ such that if $E \not\subseteq E'$
 Then $\exists a \in AR' \setminus AR,$ such that
 a attacks E and $\nexists S \subseteq E',$ such that
 S is admissible in $AF', S \cup E$ is conflict-free and S attacks a

Intuitively, the principle says that after a normal expansion, an originally inferred extension must be entailed by some extension of the normal expansion unless the original extension is attacked by some new arguments that in turn are not attacked by a set of arguments in the new extension that defends itself and is not in conflict with arguments in the original extension. We can show that the satisfaction of the RME principle implies the satisfaction of weak cautious monotony.

Proposition 1. *Let σ be an argumentation semantics. If σ satisfies the RME principle then σ satisfies weak cautious monotony.*

Proof. 1. By definition of weak cautious monotony (Definition 6), σ satisfies weak cautious monotony iff the following statement holds true for every two argumentation frameworks $AF = (AR, AT)$ and $AF' = (AR', AT')$, such that $AF \preceq_N AF'$:

$\forall E \in \sigma(AF),$ if $\nexists E' \in \sigma(AF'),$ such that $E \subseteq E'$
 Then $\exists a \in AR' \setminus AR,$ such that
 a attacks E

2. By definition of the RME principle (Definition 9) it follows that σ satisfies the RME principle iff the following statement holds true for every two argumentation frameworks $AF = (AR, AT)$ and $AF' = (AR', AT')$, such that $AF \preceq_N AF'$:

$\forall E \in \sigma(AF)$, if $\nexists E' \in \sigma(AF')$, such that $E \subseteq E'$
 Then $\exists E'' \in \sigma(AF')$, $\exists a \in AR' \setminus AR$, such that
 a attacks E and $\nexists S \subseteq E''$, such that
 S is admissible in AF' , $S \cup E$ is conflict-free and S attacks a

3. From 2. and 3. it follows that if σ satisfies the RME principle then σ satisfies weak cautious monotony. □

Also, let us show that complete, preferred, and grounded semantics satisfy the RME principle.

Proposition 2. *Let σ_x be an argumentation semantics such that $x \in \{co, pr, gr\}$. σ_x satisfies the RME principle.*

Proof. For every two argumentation frameworks $AF = (AR, AT)$ and $AF' = (AR', AT')$, such that $AF \preceq_N AF'$, for every $E \in \sigma_x(AF)$, we have two cases.

Case 1: If $\exists E' \in \sigma_x(AF')$, such that $E \subseteq E'$, it follows from the definition of the RME principle (Definition 9) that the proposition holds true.

Case 2: If $\nexists E' \in \sigma_x(AF')$, such that $E \subseteq E'$, we have two sub-cases.

Sub-case 2.1: $x \in \{co, pr\}$. Because $E \in \sigma_x(AF)$ and $\nexists E' \in \sigma_x(AF')$, such that $E \subseteq E'$, by definition of σ_x (Definition 3), it holds true that E attacks all arguments that attack E in AF and $\nexists S' \subseteq AR'$, such that S' is conflict-free, $E \subseteq S'$ and S' attacks all arguments that attack S' in AF' . It follows that $\exists a \in AR' \setminus AR$, such that a attacks E and a is not attacked by any set $S \subseteq AR$, such that $E \subseteq S$ and S is admissible in AF' . Hence, from the definition of an admissible set (Definition 2) and the definition of σ_x (Definition 3), it follows that $\exists a \in AR' \setminus AR$, $\exists E'' \in \sigma_x(AF')$, such that a attacks E and a is not attacked by any set $S \subseteq E''$, such that S is admissible in AF' , and $S \cup E$ is conflict-free; consequently, σ_x satisfies the RME principle as stipulated in Definition 9 and the proposition holds true for this sub-case.

Sub-case 2.2: $x = gr$. Note that σ_{gr} is universally uniquely defined, i.e. $|\sigma_{gr}(AF')| = 1$. Let E' be the grounded extension of AF' . Suppose for a contradiction that $E \not\subseteq E'$ and $\forall a \in AR' \setminus AR$, such that a attacks E it holds true that a is attacked by some set $S \subseteq E'$, such that S is admissible in AF' and $S \cup E$ is conflict-free. It follows that because E is the grounded extension of AF (and hence by definition the $\subseteq$ -minimal complete extension of AF), E is defended by the $\subseteq$ -minimal complete extension (and hence the grounded extension) of AF' , from which it follows that $E \subseteq E'$. This is a contradiction, which means that

if $E \not\subseteq E'$, then there must be $a \in AR' \setminus AR$, such that a attacks E and a is not attacked by $S \subseteq E'$, such that S is admissible in AF' and $S \cup E$ is conflict-free. Hence, from the definition of the RME principle (Definition 9), it follows that the proposition holds true for $x = gr$.

□

In contrast, we can show that stable semantics violates the RME principle.

Proposition 3. *Stable semantics σ_{st} violates the RME principle.*

Proof. We provide a proof by counter-example. Let us consider the argumentation frameworks $AF = (\{b\}, \{\})$ and $AF' = (\{a, b\}, \{(a, a)\})$. Let us observe that $AF \preceq_N AF'$, $\sigma_{st}(AF) = \{\{b\}\}$, and $\sigma_{st}(AF') = \emptyset$. It follows that the following statement does not hold true:

$\forall E \in \sigma(AF), \exists E' \in \sigma(AF'),$ such that if $E \not\subseteq E'$
 Then $\exists a \in AR' \setminus AR,$ such that
 a attacks E and and $\nexists S \subseteq E',$ such that
 S is admissible in $AF', S \cup E$ is conflict-free and S attacks a

By definition (Definition 9), this violates the RME principle, which proves the proposition. □

It makes sense to combine the RME principle with the *degree of monotony*-maximizing extension selection approach. Let us motivate this using an example.

Example 3. We go back to AF and AF' as introduced in Example 2 (also, see Figure 2) and consider preferred semantics. Again, we select $\{a, c\}$ as our preferred extension of AF . No matter which extension we select from $\sigma_{pr}(AF') = \{\{a, d\}, \{b, d\}\}$, we will violate monotony and $\{d\}$ will be our explanation. However, $\{d\}$ can hardly explain why we no longer infer a ; this change in inference is certainly counter-intuitive, and can be prevented by determining $Exts_{mon}(E, AF', \sigma) = \{\{a, d\}\}$ as a filter when selecting the extension we want to infer from AF' .

Let us introduce the RME_{max} principle, a stricter variant of the RME principle that ensures we can apply the approach outlined by Example 3.

Definition 10 (RME_{max} Principle). *Let σ be an argumentation semantics. σ satisfies the RME_{max} principle iff σ is universally defined and for every two argumentation frameworks $AF = (AR, AT)$ and $AF' = (AR', AT')$, such that $AF \preceq_N AF'$, the following statement holds true:*

$\forall E \in \sigma(AF), \forall E' \in Exts_{mon}(E, AF', \sigma),$ if $E \not\subseteq E'$
 Then $\exists a \in AR' \setminus AR,$ such that
 a attacks E and $\nexists S \subseteq E',$ such that
 S is admissible in $AF', S \cup E$ is conflict-free and S attacks a

Fig. 2: $\{d\}$ explains the violation of monotony when first inferring $\{a, c\}$ from AF and then inferring $\{d, c\}$ from AF' . Intuitively, d better explains losing only c when the selected extension is monotony maximizing (*i.e.* $\{a, d\}$) as opposed to losing a and c when the selected extension is not monotony maximizing (*i.e.* $\{b, d\}$).

Let us highlight that the difference to the RME principle (Definition 9) is that if monotony is violated, the *then* condition must hold true not merely for at least one extension of the expanded argumentation framework, but for all maximally monotonic extensions. Note that by definition, if one maximally monotonic extension violates monotony, then all maximally monotonic extensions violate monotony.

Also note that – as a potential alternative to the $\text{RME}_{\max}$ principle – it is not possible to require an attack from a “new” argument to at least one (or every) previously inferred argument that is no longer inferred after the update as a necessary condition for the violation of monotony (at least not if we want to select a newly inferred extension that is maximally monotonic w.r.t. the previous inference result). We can illustrate this using another example.

Example 4. Consider $AF = (AR, AT) = (\{a, b, c, d\}, \{(a, b), (b, a)\})$, $AF' = (\{a, b, c, d, e\}, \{(a, b), (b, a), (b, e), (e, c), (e, d)\})$ – see Figure 3 – and preferred semantics σ_{pr} . Note that $AF \preceq_N AF'$ and $\sigma_{pr}(AF) = \{\{a, c, d\}, \{b, c, d\}\}$. Assume that before expanding AF to AF' , we have selected $\{a, c, d\}$ from $\sigma_{pr}(AF)$. Given $\sigma_{pr}(AF') = \{\{a, e\}, \{b, c, d\}\}$, let us select $\{b, c, d\}$ as our conclusion, assuming that we want to maximize monotony, *i.e.* to reject as few of the previously inferred arguments as possible. However, $AR' \setminus AR = \{e\}$ does not attack $\{a, c, d\} \setminus \{b, c, d\} = \{a\}$; it merely dictates that to maintain a maximal degree of monotony, our conclusion needs to entail b , which in turn attacks a ; *i.e.*, it is *not necessary* to require the “new” argument e to attack some previously, but no longer inferred argument, namely a , to violate monotony.

Now, let us prove that the RME principle can indeed be combined with the degree of monotony-maximizing extension selection approach, *i.e.* that complete, preferred, and grounded semantics satisfy the $\text{RME}_{\max}$ principle.

Proposition 4. *Let σ_x be an argumentation semantics, such that $x \in \{co, pr, gr\}$. σ_x satisfies the $\text{RME}_{\max}$ principle.*

Fig. 3: e , which is the only argument in $AR' \setminus AR$, does not attack any argument accepted in AF but rejected in AF' after selecting maximally monotonic extensions.

Proof. Let us note that σ_x is universally defined [3]. Because for every argumentation framework, there exists exactly one grounded extension, for $x = gr$ we conclude that $|Exts_{mon}(E, AF', \sigma_x)| = |\sigma_x(AF')| = 1$ and the proof follows from Proposition 2. For $x \in \{co, pr\}$, we have two cases $\forall E \in \sigma_x(AF), \forall E' \in Exts_{mon}(E, AF', \sigma_x)$:

Case 1: $E \subseteq E'$. By definition of RME_{max} , the proposition holds true for this case.

Case 2: $E \not\subseteq E'$. By definition of $Exts_{mon}$ (Definition 8), $\forall E'' \in \sigma_x(AF')$ it holds that $E \not\subseteq E''$ (because otherwise, if $E \subseteq E''$, then $deg_{mon}(E'', E) = 1$, i.e. $deg_{mon}(E'', E)$ is maximal, and hence $deg_{mon}(E'', E) > deg_{mon}(E', E)$ (since $E \not\subseteq E'$), which contradicts $E' \in Exts_{mon}(E, AF', \sigma_x)$). Hence, from Proposition 2 it follows that $\exists E^* \in \sigma_x(AF')$ such that $\exists a \in AR' \setminus AR$, such that a is not attacked by any set $S \subseteq E^*$, such that S is admissible in AF' and $S \cup E$ is conflict-free. Suppose for a contradiction that $\forall a' \in AR' \setminus AR$, such that a' attacks E it holds true that a' is attacked by some set $S' \subseteq E^*$, such that S' is admissible in AF' and $S' \cup E$ is conflict-free. It follows that by definition of σ_x (Definition 3), $S' \cup E$ is an admissible set in AF' and hence, (again by definition of σ_x), $\exists E'' \in \sigma_x(AF')$, such that $E \subseteq E''$. This contradicts that $E \not\subseteq E''$ as noted above, after the premise of the case (Case 2.), and proves the proposition. □

Finally, let us define the notion of *monotony violation explanations*, i.e. arguments whose attacks on arguments in a previously inferred extension explain – in a normal expansion scenario – the violation of monotony, given a semantics that satisfies the RME_{max} principle.

Definition 11 (Monotony Violation Explanations). Let $AF = (AR, AT)$ and $AF' = (AR', AT')$ be argumentation frameworks, such that $AF \preceq_N AF'$ and let σ be an argumentation semantics that satisfies the RME_{max} principle. Let $E \in \sigma(AF)$ and $E' \in \sigma(AF')$, such that $E' \in Exts_{mon}(E, AF', \sigma)$. The monotony violation explanations of E' w.r.t. E, σ, AF and AF' (denoted by

$EXPS_{E,E'}(AF, AF, \sigma)$ are $\{a \mid a \in AR' \setminus AR, a \text{ attacks } E, a \text{ is not attacked by any set } S \subseteq E', \text{ such that } S \text{ is admissible in } AF', S \cup E \text{ is conflict-free}\}$.

Let us go back to the examples and identify the monotony violation explanations.

Example 5. Consider the following examples:

- Consider the argumentation frameworks AF' and AF'' as depicted by Figure 1, preferred semantics and $E = \{a, c, d\}, E \in \sigma_{pr}(AF'), E' = \{b, c, d, f\}, E' \in Ext_{mon}(E, AF'', \sigma_{pr})$. $EXPS_{E,E'}(AF', AF'', \sigma_{pr}) = \{e, f\}$.
- Consider the argumentation frameworks AF and AF' as depicted by Figure 2, preferred semantics and $E = \{b, c\}, E \in \sigma_{pr}(AF), E' = \{b, d\}, E' \in Ext_{mon}(E, AF', \sigma_{pr})$. $EXPS_{E,E'}(AF, AF', \sigma_{pr}) = \{d\}$.
- Consider the argumentation frameworks AF and AF' as depicted by Figure 3, preferred semantics and $E = \{a, c, d\}, E \in \sigma_{pr}(AF), E' = \{b, c, d\}, E' \in Ext_{mon}(E, AF', \sigma_{pr})$. $EXPS_{E,E'}(AF, AF', \sigma_{pr}) = \{e\}$.

Finally, let us show that given complete, preferred or grounded semantics, for every argumentation frameworks, one of its extensions, and any of its normal expansions, for all extensions of the normal expansion that entail the originally inferred extension (if there are any), the set of monotony violation explanations of any of these new extensions w.r.t. the originally inferred extension is empty.

Proposition 5. *Let σ_x be an argumentation semantics such that $x \in \{co, pr, gr\}$. For every two argumentation frameworks $AF = (AR, AT), AF' = (AR', AT')$ such that $AF \preceq_N AF', \forall E \in \sigma_x(AF), E' \in \sigma_x(AF')$, such that $E \subseteq E'$, it holds true that $EXPS_{E,E'}(AF, AF', \sigma_x) = \emptyset$.*

Proof. Suppose for a contradiction that there exist two argumentation frameworks $AF = (AR, AT), AF' = (AR', AT')$, such that $AF \preceq_N AF'$ and $\exists E \in \sigma_x(AF), E' \in \sigma_x(AF')$, such that $E \subseteq E'$ and $EXPS_{E,E'}(AF, AF', \sigma_x) \neq \emptyset$. By definition of σ_x (Definition 3), it holds true that E' is conflict-free and $\forall b \in E', E'$ defends b (in AF'). Because $E \subseteq E'$, it follows that E' is admissible in AF' , $E' \cup E$ is conflict-free and E' attacks all arguments that attack E . Consequently, it holds true that $\nexists a \in AR' \setminus AR$, such that a attacks E and a is not attacked by any set $S \subseteq E'$, such that S is admissible in AF' and $S \cup E$ is conflict-free. From the definition of monotony violation explanations (Definition 11), it follows that $EXPS_{E,E'}(AF, AF', \sigma_x) = \emptyset$, which contradicts $EXPS_{E,E'}(AF, AF', \sigma_x) \neq \emptyset$ as noted above and proves the proposition. $\square$

4 Implementation

A Java implementation, including documentation (tutorial and application programming interface documentation) of the explainability framework is available at <https://git.io/J0izF>. The source code is openly available and the software can be installed using Java dependency management and build tools such as

Maven⁶ and Gradle⁷. The implementation is part of the DiArg argumentation-based dialog reasoner, which in turn is based on the abstract argumentation libraries that are provided by the Tweety project [24]. As the implementation can – to a large extent – rely on existing argumentation reasoners, we consider an analysis of algorithmic implementation details (which depend on the implementation of these dependencies) out of the scope of this paper, yet potentially relevant future work.

5 Discussion

Our explainability framework makes use of research results on *argumentation dynamics* (see [12] for a survey), and in particular relies on notions that have been introduced in the context of monotony analysis, *i.e.* on Baumann’s and Brewka’s *normal expansions* [4]⁸ and our recently introduced weak cautious monotony abstract argumentation principle [20]. By extending the weak cautious monotony principle, we make the aforementioned results applicable to the domain for explainable argumentation. From the perspective of explainable automated reasoning (see [10] for a survey), our research can be considered a novel contribution at the intersection of principle-based [25] and explainable [26] argumentation.

Argumentation-based explanations take many forms (see *e.g.* [11] for a recent survey). A common approach to explaining argument acceptability – *i.e.* membership in extensions, or inference – in argumentation frameworks essentially amounts to traversing the argument graph and can be formalized by extracting graphs that satisfy some formal properties [8]. A popular example of explanations as sub-graphs with desirable properties uses the concept of a dispute tree [14,15], which commonly, but not exclusively, gives rise to conversational explanations, often by means of a dialogue (game). Explanations in argumentation frameworks can also come in the form of extensions themselves, *e.g.* [18,21], as well as by means of the structure of the extensions’ arguments and relationships, where applicable in structured argumentation frameworks.

Notions of argumentation-based explanations that are most related to our work are approaches that modify the argument graph to make an argument (non-)acceptable. Commonly, addition or removal of arguments and/or relations that change the acceptability status of some pre-specified argument is a form of explanation, *e.g.* [17,22,23]. Such changes typically presuppose some “universal” space of argument graph modification [28,6,22]. Instead, our Monotony Violation Explanations amount to sets of arguments that best explain the differences between inferences in an abstract argumentation framework before and after a

⁶ <https://maven.apache.org/>

⁷ <https://gradle.org/>

⁸ Let us highlight that we make use of normal expansions and not of the change operations for abstract argumentation frameworks that were introduced by Cayrol *et al.* [7] because the latter do not support the addition of arbitrarily many arguments as part of a single operation, which makes Baumann’s and Brewka’s normal expansions slightly more convenient in our case.

change – a (normal) expansion – of the argument graph. Importantly, our explanations consider the aspect of monotony maximization when selecting extensions for inference after the change. This roughly pertains to indicating changes that led to altered inferences and indirectly explain why some new inferences are preferred over others. Consequently, our approach can be seen as the reverse of the kind of explanations that indicate what changes could be applied to an argumentation framework to achieve desirable inferences, as in the works discussed above.

Future work can extend the formal framework we provide in this paper, for example as follows.

- Analyses of RME and RME_{max} principle satisfaction for naive set-based [1] and weak admissible set-based [5] argumentation semantics can be conducted. Let us claim that some naive set-based semantics like stage [27], CF2 [1], and stage2 [16] semantics do not satisfy the RME and RME_{max} principles: consider the argumentation frameworks $AF = (\{a, b, c\}, \{(a, b), (b, c), (c, a)\})$, $AF' = (\{a, b, c, d\}, \{(a, b), (b, c), (c, a), (d, c)\})$ for a counter-example. Consequently, new principles are required to explain the violation of monotony for these semantics, and potentially, the identification of additional relaxed monotony explainability principles can apply a principle-based analysis instead of proving the satisfaction with/violation of a principle semantics-by-semantics.
- Analogous principles can potentially be introduced to other formal argumentation approaches, such as gradual (bipolar) argumentation (see e.g. [2]). For instance, one could define monotony with respect to the relative order of argument final strengths after either some expansion of the argument graph or a change in the initial strengths of arguments. The degree of monotony could then be based on some of the many similarity measures over ordered sequences. Instead of looking at monotony-maximizing extensions then, one could look at arguments that changed their relative ordering, and define and extract explanatory arguments (or changes in their initial strengths) that best explain those changes, potentially by also referring to other properties studied in [2].
- Human-Computer Interaction (HCI) studies that assess the effectiveness of the explanations can be conducted. Let us highlight that at the current stage, we consider an HCI evaluation premature. The behavior of Dung-style admissible set-based semantics is in many cases counter-intuitive (see [9], and consider, for instance, the argumentation framework $AF = (\{a, b\}, \{(a, a), (a, b)\})$ and preferred semantics); hence, an HCI study is likely to yield rather “noisy” results, considering that technically correct explanation of counter-intuitive semantics behavior may confuse study participants.

6 Conclusion

In this paper, we have introduced explanations for the acceptability dynamics in dynamic systems, where explanations attribute changes to argumentation frame-

works with reference to a desirable relaxed monotony property. Based on this property, we have constructed a theoretical framework for explaining the violation of monotony in the context of *Dung-style* admissible set-based semantics. While a software implementation of this theoretical framework is provided, we consider this work primarily an initial stepping stone towards real-world applicable explainability of monotony violations of non-monotonic machine reasoners.

Acknowledgments

We thank the anonymous reviewers for their thoughtful and useful feedback. This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

References

1. Baroni, P., Caminada, M., Giacomin, M.: Abstract argumentation frameworks and their semantics. In: Baroni, P., Gabbay, D., Massimiliano, G., van der Torre, L. (eds.) *Handbook of Formal Argumentation*. College Publications, chap. 4, pp. 159–236. College Publications (2018)
2. Baroni, P., Rago, A., Toni, F.: From Fine-Grained Properties to Broad Principles for Gradual Argumentation: A Principled Spectrum. *International Journal of Approximate Reasoning* **105**, 252–286 (feb 2019). <https://doi.org/10.1016/j.ijar.2018.11.019>
3. Baumann, R.: On the nature of argumentation semantics: Existence and uniqueness, expressibility, and replaceability. *Journal of Applied Logics* **4**(8), 2779–2886 (2017)
4. Baumann, R., Brewka, G.: Expanding argumentation frameworks: Enforcing and monotonicity results. *COMMA* **10**, 75–86 (2010)
5. Baumann, R., Brewka, G., Ulbricht, M.: Revisiting the foundations of abstract argumentation-semantics based on weak admissibility and weak defense. In: *AAAI*. pp. 2742–2749 (2020)
6. Booth, R., Gabbay, D.M., Kaci, S., Rienstra, T., van der Torre, L.: Abduction and Dialogical Proof in Argumentation and Logic Programming. In: Schaub, T., Friedrich, G., O’Sullivan, B. (eds.) *21st European Conference on Artificial Intelligence*. *Frontiers in Artificial Intelligence and Applications*, vol. 263, pp. 117–122. IOS Press, Prague (2014). <https://doi.org/10.3233/978-1-61499-419-0-117>
7. Cayrol, C., de Saint-Cyr, F.D., Lagasque-Schiex, M.C.: Change in abstract argumentation frameworks: Adding an argument. *Journal of Artificial Intelligence Research* **38**, 49–84 (2010)
8. Cocarascu, O., Čyras, K., Rago, A., Toni, F.: Explaining with Argumentation Frameworks Mined from Data. In: *1st International Workshop on Dialogue, Explanation and Argumentation in Human-Agent Interaction (DEXAHAI)*. Southampton (2018)
9. Cramer, M., Guillaume, M.: Empirical study on human evaluation of complex argumentation frameworks. In: Calimeri, F., Leone, N., Manna, M. (eds.) *Logics in Artificial Intelligence*. pp. 102–115. Springer International Publishing, Cham (2019)

10. Čyras, K., Badrinath, R., Mohalik, S.K., Mujumdar, A., Nikou, A., Previti, A., Sundararajan, V., Feljan, A.V.: Machine Reasoning Explainability. arXiv preprint arXiv:2009.00418 (2020)
11. Čyras, K., Rago, A., Albini, E., Baroni, P., Toni, F.: Argumentative XAI: A Survey. In: Zhou, Z.H. (ed.) 30th International Joint Conference on Artificial Intelligence. pp. 4392–4399. IJCAI, Montreal (2021). <https://doi.org/10.24963/ijcai.2021/600>
12. Doutre, S., Mailly, J.G.: Constraints and changes: A survey of abstract argumentation dynamics. *Argument & Computation* **9**, 223–248 (2018). <https://doi.org/10.3233/AAC-180425>
13. Dung, P.M.: On the acceptability of arguments and its fundamental role in non-monotonic reasoning, logic programming and n-person games. *Artificial intelligence* **77**(2), 321–357 (1995)
14. Dung, P.M., Kowalski, R., Toni, F.: Dialectic Proof Procedures for Assumption-Based, Admissible Argumentation. *Artificial Intelligence* **170**(2), 114–159 (2006). <https://doi.org/10.1016/j.artint.2005.07.002>
15. Dung, P.M., Mancarella, P., Toni, F.: Computing Ideal Sceptical Argumentation. *Artificial Intelligence* **171**(10-15), 642–674 (2007). <https://doi.org/10.1016/j.artint.2007.05.003>
16. Dvořák, W., Gaggl, S.A.: Stage semantics and the SCC-recursive schema for argumentation semantics. *Journal of Logic and Computation* **26**(4), 1149–1202 (02 2014). <https://doi.org/10.1093/logcom/exu006>
17. Fan, X., Toni, F.: On Computing Explanations for Non-Acceptable Arguments. In: Black, E., Modgil, S., Oren, N. (eds.) *Theory and Applications of Formal Argumentation - 3rd International Workshop*. Lecture Notes in Computer Science, vol. 9524, pp. 112–127. Springer, Buenos Aires (2015). https://doi.org/10.1007/978-3-319-28460-6_7
18. Fan, X., Toni, F.: On Computing Explanations in Argumentation. In: Bonet, B., Koenig, S. (eds.) 29th AAAI Conference on Artificial Intelligence. pp. 1496–1502. AAAI Press, Austin, Texas (2015)
19. Kampik, T., Gabbay, D.: The “Degrees of Monotony”-Dilemma in Abstract Argumentation. In: *Symbolic and Quantitative Approaches to Reasoning with Uncertainty 2021*, to appear (2021)
20. Kampik, T., Nieves, J.C.: Abstract argumentation and the rational man. *Journal of Logic and Computation* **31**(2), 654–699 (02 2021). <https://doi.org/10.1093/logcom/exab003>
21. Liao, B., van der Torre, L.: Explanation Semantics for Abstract Argumentation. In: Prakken, H. (ed.) *Computational Models of Argument*. vol. 326, pp. 271–282. IOS Press (2020). <https://doi.org/10.3233/FAIA200511>
22. Sakama, C.: Abduction in Argumentation Frameworks. *Journal of Applied Non-Classical Logics* **28**(2-3), 218–239 (2018). <https://doi.org/10.1080/11663081.2018.1487241>
23. Saribatur, Z.G., Wallner, J.P., Woltran, S.: Explaining Non-Acceptability in Abstract Argumentation. In: Giacomo, G.D., Catalá, A., Dilkina, B., Milano, M., Barro, S., Bugarín, A., Lang, J. (eds.) 24th European Conference on Artificial Intelligence. pp. 881–888. IOS Press, Santiago de Compostela (2020). <https://doi.org/10.3233/FAIA200179>
24. Thimm, M.: Tweety: A comprehensive collection of java libraries for logical aspects of artificial intelligence and knowledge representation. In: *Proceedings of the Fourteenth International Conference on Principles of Knowledge Representation and Reasoning*. p. 528–537. KR’14, AAAI Press (2014)

25. van der Torre, L., Vesic, S.: The principle-based approach to abstract argumentation semantics. *IfCoLog Journal of Logics and Their Applications* **4**(8), 2735–2778 (October 2017)
26. Vassiliades, A., Bassiliades, N., Patkos, T.: Argumentation and explainable artificial intelligence: a survey. *The Knowledge Engineering Review* **36**, e5 (2021). <https://doi.org/10.1017/S0269888921000011>
27. Verheij, B.: Two approaches to dialectical argumentation: admissible sets and argumentation stages. *Proc. NAIC* **96**, 357–368 (1996)
28. Wakaki, T., Nitta, K., Sawamura, H.: Computing Abductive Argumentation in Answer Set Programming. In: McBurney, P., Rahwan, I., Parsons, S., Maudet, N. (eds.) *6th International Workshop on Argumentation in Multi-Agent Systems*. pp. 195–215. Springer, Budapest (2009). https://doi.org/10.1007/978-3-642-12805-9_12