Necessary and Sufficient Explanations for Argumentation-Based Conclusions

AnneMarie Borg[1(✉)] and Floris Bex[1,2]

[1] Department of Information and Computing Sciences, Utrecht University, Utrecht, The Netherlands
{a.borg,f.j.bex}@uu.nl

[2] Tilburg Institute for Law, Technology, and Society, Tilburg University, Tilburg, The Netherlands

Abstract. In this paper, we discuss *necessary* and *sufficient* explanations – the question whether and why a certain argument or claim can be accepted (or not) – for abstract and structured argumentation. Given a framework with which explanations for argumentation-based conclusions can be derived, we study necessity and sufficiency: what (sets of) arguments are necessary or sufficient for the (non-)acceptance of an argument or claim? We will show that necessary and sufficient explanations can be strictly smaller than minimal explanations, while still providing all the reasons for a conclusion and we discuss their usefulness in a real-life application.

Keywords: Computational argumentation · Structured argumentation · Explainable artificial intelligence

1 Introduction

In recent years, *explainable AI* (XAI) has received much attention, mostly directed at new techniques for explaining decisions of (subsymbolic) machine learning algorithms [18]. However, explanations traditionally also play an important role in (symbolic) knowledge-based systems [10]. Computational argumentation is one research area in symbolic AI that is frequently mentioned in relation to XAI. For example, arguments can be used to provide reasons for or against decisions [1,10,15]. The focus can also be on the argumentation itself, where it is explained whether and why a certain argument or claim can be accepted under certain semantics for computational argumentation [7–9,19]. It is the latter type of explanations that is the subject of this paper.

Two central concepts in computational argumentation are *abstract argumentation frameworks* [6] – sets of arguments and the attack relations between them – and *structured argumentation frameworks* [3] – where arguments are constructed from a knowledge base and a set of rules and the attack relation is based

This research has been partly funded by the Dutch Ministry of Justice and the Dutch National Police.

J. Vejnarová and N. Wilson (Eds.): ECSQARU 2021, LNAI 12897, pp. 45–58, 2021.
https://doi.org/10.1007/978-3-030-86772-0_4

on the individual elements in the arguments. The explanations framework that is introduced in [5] is designed to provide explanations for the (non-)acceptance of arguments and the claim of an argument in case of a structured setting. However, like the other existing works on explanations for argumentation-based conclusions, the framework does not account for findings from the social sciences on human explanations [15].

One of the important characteristics of explanations provided by humans is that they select *the* explanation from a possible infinite set of explanations [15]. In this paper we look at how to select minimal,[1] necessary and sufficient explanations for the (non-)acceptance of an argument. To this end we will introduce variations to the basic framework from [5] that will provide necessary or sufficient explanations for both abstract and structured argumentation (i.e., ASPIC$^+$[17]). Intuitively, a necessary explanation contains the arguments that one has to accept in order to accept the considered argument and a sufficient explanation contains the arguments that, when accepted, guarantee the acceptance of the considered argument. We will show that such explanations exist in most cases and how these relate to the basic explanations from [5] as well as to minimal explanations as introduced in [7]. Moreover, we will discuss a real-life application from the Dutch National Police, where the necessary and sufficient explanations will reduce the size of the provided explanations in a meaningful way.

The paper is structured as follows. We start with a short overview of related work and present the preliminaries on abstract and structured argumentation and the basic explanations from [5]. Then, in Sect. 4 we introduce necessary and sufficient explanations and study how these relate to the better known minimal explanations. In Sect. 5 we show how a real-life application benefits from these explanations and we conclude in Sect. 6.

2 Related Work

We are interested in local explanations for computational argumentation: explanations for a specific argument or claim. We work here with the framework from [5] for several reasons. Often, explanations are only defined for a specific semantics [7,8] and can usually only be applied to abstract argumentation [8,11,19],[2] while the framework from [5] can be applied on top of any argumentation setting (structured or abstract) that results in a Dung-style argumentation framework. Furthermore, when this setting is a structured one based on a knowledge base and set of rules (like ASPIC$^+$ or logic-based argumentation [3]), the explanations can be further adjusted (something which is not considered at all in the literature). To the best of our knowledge, this is the first approach to explanations for formal argumentation in which necessary and sufficient explanations are considered and integrated into a real-life application.

[1] Interpreting [15]'s simplicity as minimality.

[2] These explanations do not account for the sub-argument relation in structured argumentation. For example, in structured argumentation one cannot remove specific arguments or attacks without influencing other arguments/attacks.

3 Preliminaries

An *abstract argumentation framework* (AF) [6] is a pair $\mathcal{AF} = \langle \mathsf{Args}, \mathsf{Att} \rangle$, where Args is a set of *arguments* and $\mathsf{Att} \subseteq \mathsf{Args} \times \mathsf{Args}$ is an *attack relation* on these arguments. An AF can be viewed as a directed graph, in which the nodes represent arguments and the arrows represent attacks between arguments.

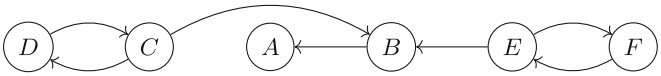

Fig. 1. Graphical representation of the AF $\mathcal{AF}_1$.

Example 1. Figure 1 represents $\mathcal{AF}_1 = \langle \mathsf{Args}_1, \mathsf{Att}_1 \rangle$ where $\mathsf{Args}_1 = \{A, B, C, D, E, F\}$ and $\mathsf{Att}_1 = \{(B, A), (C, B), (C, D), (D, C), (E, B), (E, F), (F, E)\}$.

Given an AF, Dung-style semantics [6] can be applied to it, to determine what combinations of arguments (called *extensions*) can collectively be accepted.

Definition 1. *Let $\mathcal{AF} = \langle \mathit{Args}, \mathit{Att} \rangle$ be an argumentation framework, $\mathsf{S} \subseteq \mathit{Args}$ a set of arguments and let $A \in \mathit{Args}$. Then:* S attacks A *if there is an $A' \in \mathsf{S}$ such that $(A', A) \in \mathit{Att}$, let S^+ denote the set of arguments attacked by S;* S defends A *if* S *attacks every attacker of A;* S *is* conflict-free *if there are no $A_1, A_2 \in \mathsf{S}$ such that $(A_1, A_2) \in \mathit{Att}$; and* S *is an* admissible extension *(*Adm*) if it is conflict-free and it defends all of its elements.*

An admissible extension that contains all the arguments that it defends is a complete extension *(*Cmp*). The* grounded extension *(*Grd*) is the minimal (w.r.t. $\subseteq$) complete extension; A* preferred extension *(*Prf*) is a maximal (w.r.t. $\subseteq$) complete extension; and A* semi-stable extension *(*Sstb*) is a complete extension for which $\mathsf{S} \cup \mathsf{S}^+$ is $\subseteq$-maximal.* $\mathsf{Sem}(\mathcal{AF})$ *denotes the set of all the extensions of $\mathcal{AF}$ under the semantics* $\mathsf{Sem} \in \{\mathsf{Adm}, \mathsf{Cmp}, \mathsf{Grd}, \mathsf{Prf}, \mathsf{Sstb}\}$.

In what follows we will consider an argument accepted if it is part of at least one extension and non-accepted if it is not part of at least one extension.[3]

Definition 2. *Where $\mathcal{AF} = \langle \mathit{Args}, \mathit{Att} \rangle$ is an AF,* Sem *a semantics such that $\mathsf{Sem}(\mathcal{AF}) \neq \emptyset$, it is said that $A \in \mathit{Args}$ is:*

- accepted *if $A \in \bigcup \mathsf{Sem}(\mathcal{AF})$: there is some* Sem*-extension that contains A;*
- non-accepted *if $A \notin \bigcap \mathsf{Sem}(\mathcal{AF})$: A is not part of at least one* Sem*-extension.*

Example 2. For $\mathcal{AF}_1$ we have that $\mathsf{Grd}(\mathcal{AF}_1) = \{\emptyset\}$ and there are four preferred and semi-stable extensions: $\{A, C, E\}$, $\{A, C, F\}$, $\{A, D, E\}$ and $\{B, D, F\}$. Therefore, all arguments from Args_1 are accepted and non-accepted for $\mathsf{Sem} \in \{\mathsf{Prf}, \mathsf{Sstb}\}$.

[3] In [5], four acceptance strategies are considered, the other two are not relevant for our study on necessity and sufficiency and are therefore not introduced here.

3.1 ASPIC$^+$

For our discussion on explanations for structured settings we take ASPIC$^+$ [17] which allows for two types of premises – *axioms* that cannot be questioned and *ordinary premises* that can be questioned – and two types of rules – *strict* rules that cannot be questioned and *defeasible* ones. We choose ASPIC$^+$ since it allows to vary the form of the explanations in many ways (see Sect. 3.2 and [5]).

An ASPIC$^+$ setting starts from an *argumentation system* (AS $= \langle \mathcal{L}, \mathcal{R}, n \rangle$), which contains a logical language $\mathcal{L}$ closed under negation ($\neg$), a set of rules $\mathcal{R} = \mathcal{R}_s \cup \mathcal{R}_d$ consisting of strict ($\mathcal{R}_s$) and defeasible ($\mathcal{R}_d$) rules and a naming convention n for defeasible rules. Arguments are constructed in an argumentation setting from a knowledge base $\mathcal{K} \subseteq \mathcal{L}$ which consists of two disjoint subsets $\mathcal{K} = \mathcal{K}_p \cup \mathcal{K}_n$: the set of axioms ($\mathcal{K}_n$) and the set of ordinary premises ($\mathcal{K}_p$).

Definition 3. *An* argument A *on the basis of a knowledge base* $\mathcal{K}$ *in an argumentation system* $\langle \mathcal{L}, \mathcal{R}, n \rangle$ *is:*

1. ϕ *if* $\phi \in \mathcal{K}$, *where* $\mathsf{Prem}(A) = \mathsf{Sub}(A) = \{\phi\}$, $\mathsf{Conc}(A) = \phi$, $\mathsf{Rules}(A) = \emptyset$ *and* $\mathsf{TopRule}(A) =$ *undefined;*
2. $A_1, \ldots, A_n \rightsquigarrow \psi$, *where* $\rightsquigarrow \in \{\rightarrow, \Rightarrow\}$, *if* $A_1, \ldots, A_n$ *are arguments such that there exists a rule* $\mathsf{Conc}(A_1), \ldots, \mathsf{Conc}(A_n) \rightsquigarrow \psi$ *in* $\mathcal{R}_s$ *if* $\rightsquigarrow = \rightarrow$ *and in* $\mathcal{R}_d$ *if* $\rightsquigarrow = \Rightarrow$.
   $\mathsf{Prem}(A) = \mathsf{Prem}(A_1) \cup \ldots \cup \mathsf{Prem}(A_n)$; $\mathsf{Conc}(A) = \psi$; $\mathsf{Sub}(A) = \mathsf{Sub}(A_1) \cup \ldots \cup \mathsf{Sub}(A_n) \cup \{A\}$; $\mathsf{Rules}(A) = \mathsf{Rules}(A_1) \cup \ldots \cup \mathsf{Rules}(A_n) \cup \{\mathsf{Conc}(A_1), \ldots, \mathsf{Conc}(A_n) \rightsquigarrow \psi\}$; $\mathsf{DefRules}(A) = \{r \in \mathcal{R}_d \mid r \in \mathsf{Rules}(A)\}$; *and* $\mathsf{TopRule}(A) = \mathsf{Conc}(A_1), \ldots, \mathsf{Conc}(A_n) \rightsquigarrow \psi$.

The above notation can be generalized to sets. For example, where S *is a set of arguments* $\mathsf{Prem}(\mathsf{S}) = \bigcup\{\mathsf{Prem}(A) \mid A \in \mathsf{S}\}$ *and* $\mathsf{Conc}(\mathsf{S}) = \{\mathsf{Conc}(A) \mid A \in \mathsf{S}\}$.

Attacks on an argument are based on the rules and premises applied in the construction of that argument.

Definition 4. *Let* A *and* B *be two arguments, we denote* $\psi = -\phi$ *if* $\psi = \neg\phi$ *or* $\phi = \neg\psi$. A attacks *an argument* B *iff* A undercuts, rebuts *or* undermines B*:*

- A undercuts B *(on* B'*) iff* $\mathsf{Conc}(A) = -n(r)$ *for some* $B' \in \mathsf{Sub}(B)$ *such that* B'*'s top rule* r *is defeasible, it denies a rule;*
- A rebuts B *(on* B'*) iff* $\mathsf{Conc}(A) = -\phi$ *for some* $B' \in \mathsf{Sub}(B)$ *of the form* $B'_1, \ldots, B'_n \Rightarrow \phi$, *it denies a conclusion;*
- A undermines B *(on* ϕ*) iff* $\mathsf{Conc}(A) = -\phi$ *for some* $\phi \in \mathsf{Prem}(B) \setminus \mathcal{K}_n$, *it denies a premise.*

Argumentation theories and their corresponding Dung-style argumentation frameworks can now be defined.

Definition 5. *An* argumentation theory *is a pair* $AT = \langle AS, \mathcal{K} \rangle$, *where* AS *is an argumentation system and* $\mathcal{K}$ *is a knowledge base.*

From an argumentation theory AT *the corresponding* AF *can be derived such that* $\mathcal{AF}(AT) = \langle \mathsf{Args}, \mathsf{Att} \rangle$, *where* Args *is the set of arguments constructed from* AT *and* $(A, B) \in \mathsf{Att}$ *iff* $A, B \in \mathsf{Args}$ *and* A *attacks* B *as defined in Definition 4.*

Example 3. Let $\mathrm{AS}_1 = \langle \mathcal{L}_1, \mathcal{R}_1, n \rangle$ where the rules in $\mathcal{R}_1$ are such that, with $\mathcal{K}_1 = \mathcal{K}_n^1 = \{r, s, t, v\}$ the following arguments can be derived:[4]

$$A : s, t \overset{d_1}{\Rightarrow} u \qquad B : p, \neg q \overset{d_2}{\Rightarrow} \neg n(d_1) \qquad C : r, s \overset{d_3}{\Rightarrow} q$$
$$D : v \overset{d_4}{\Rightarrow} \neg q \qquad E : r, t \overset{d_5}{\Rightarrow} \neg p \qquad F : v \overset{d_6}{\Rightarrow} p.$$

The graphical representation of the corresponding argumentation framework $\mathcal{AF}(\mathrm{AT}_1)$ with $\mathrm{AT}_1 = \langle \mathrm{AS}_1, \mathcal{K}_1 \rangle$ is the graph from Fig. 1.

Dung-style semantics (Definition 1) can be applied in the same way as they are applied to abstract argumentation frameworks (recall Example 2). In addition to (non-)acceptance of arguments, in a structured setting we can also consider (non-)acceptance of formulas:

Definition 6. *Let $\mathcal{AF}(AT) = \langle \mathsf{Args}, \mathsf{Att} \rangle$ be an AF, based on AT, let* Sem *be a semantics such that* $\mathsf{Sem}(\mathcal{AF}) \neq \emptyset$ *and let $\phi \in \mathcal{L}$. Then ϕ is:*

- accepted*: if $\phi \in \bigcup \mathsf{Concs}(\mathsf{Sem}(\mathcal{AF}(AT)))$, that is: there is some argument with conclusion ϕ that is accepted;*
- non-accepted*: if $\phi \notin \bigcap \mathsf{Concs}(\mathsf{Sem}(\mathcal{AF}(AT)))$, that is: there is some* Sem*-extension without an argument with conclusion ϕ.*

Example 4. As was the case for arguments (recall Example 2), all formulas in $\{p, \neg p, q, \neg q, r, s, t, u, v\}$ are accepted and $\{p, \neg p, q, \neg q, u\}$ are also non-accepted.

3.2 Basic Explanations

In [5] four types of explanations for abstract and structured argumentation were introduced. These explanations are defined in terms of two functions: $\mathbb{D}$, which determines the arguments that are in the explanation and $\mathbb{F}$, which determines what elements of these arguments the explanation presents. For the basic explanations in this paper, we instantiate $\mathbb{D}$ with the following functions, let $A \in \mathsf{Args}$ and $\mathcal{E} \in \mathsf{Prf}(\mathcal{AF})$ for some AF $\mathcal{AF} = \langle \mathsf{Args}, \mathsf{Att} \rangle$:[5]

- $\mathsf{Defending}(A) = \{B \in \mathsf{Args} \mid B \text{ defends } A\}$ denotes the set of arguments that defend A and $\mathsf{Defending}(A, \mathcal{E}) = \mathsf{Defending}(A) \cap \mathcal{E}$ denotes the set of arguments that defend A in $\mathcal{E}$.
- $\mathsf{NoDefAgainst}(A, \mathcal{E}) = \{B \in \mathsf{Args} \mid B \text{ attacks } A \text{ and } \mathcal{E} \text{ does not defend } A \text{ against } B\}$ denotes the set of all attackers of A that are not defended by $\mathcal{E}$.

The explanations are defined for arguments and formulas.

[4] We ignore the arguments based on the elements from $\mathcal{K}_1$, since these neither attack nor are attacked by any argument.

[5] We write that $B \in \mathsf{Args}$ defends $A \in \mathsf{Args}$ if it attacks an attacker of A or it defends an argument that defends A.

Definition 7. *Let $\mathcal{AF} = \langle \mathit{Args}, \mathit{Att} \rangle$ be an AF and suppose that $A \in \mathit{Args}$ [resp. $\phi \in \mathcal{L}$] is accepted w.r.t.* Sem*. Then:*

$$\mathsf{SemAcc}(A) = \{\mathsf{Defending}(A, \mathcal{E}) \mid \mathcal{E} \in \mathsf{Sem}(\mathcal{AF}) \text{ and } A \in \mathcal{E}\}.$$
$$\mathsf{SemAcc}(\phi) = \{\mathbb{F}(\mathsf{Defending}(A, \mathcal{E})) \mid \mathcal{E} \in \mathsf{Sem}(\mathcal{AF}) \text{ such that } A \in \mathcal{E} \text{ and } \mathsf{Conc}(A) = \phi\}.$$

An acceptance explanation, for an argument or formula, contains all the arguments that defend the argument (for that formula) in an extension. If it is an explanation for a formula, the function $\mathbb{F}$ can be applied to it.

Definition 8. *Let $\mathcal{AF} = \langle \mathit{Args}, \mathit{Att} \rangle$ be an AF and suppose that $A \in \mathit{Args}$ [resp. $\phi \in \mathcal{L}$] is non-accepted w.r.t.* Sem*. Then:*

$$\mathsf{SemNotAcc}(A) = \bigcup_{\mathcal{E} \in \mathsf{Sem}(\mathcal{AF}) \text{ and } A \notin \mathcal{E}} \mathsf{NoDefAgainst}(A, \mathcal{E}).$$
$$\mathsf{SemNotAcc}(\phi) = \bigcup_{A \in \mathit{Args} \text{ and } \mathsf{Conc}(A) = \phi} \; \bigcup_{\mathcal{E} \in \mathsf{Sem}(\mathcal{AF}) \text{ and } A \notin \mathcal{E}} \mathbb{F}(\mathsf{NoDefAgainst}(A, \mathcal{E})).$$

A non-acceptance explanation contains all the arguments that attack the argument [resp. an argument for the formula] and to which no defense exists in some Sem-extension. For a formula $\mathbb{F}$ can be applied again.

The function $\mathbb{F}$ can be instantiated in different ways. We recall here some of the variations introduced in [5].

- $\mathbb{F} = \mathsf{id}$, where $\mathsf{id}(\mathsf{S}) = \mathsf{S}$. Then explanations are sets of arguments.
- $\mathbb{F} = \mathsf{Prem}$. Then explanations only contain the premises of arguments (i.e., knowledge base elements).
- $\mathbb{F} = \mathsf{AntTop}$, where $\mathsf{AntTop}(A) = \langle \mathsf{TopRule}(A), \mathsf{Ant}(\mathsf{TopRule}(A)) \rangle$. Then explanations contain the last applied rule and its antecedents.
- $\mathbb{F} = \mathsf{SubConc}$, where $\mathsf{SubConc}(A) = \{\mathsf{Conc}(B) \mid B \in \mathsf{Sub}(A), \mathsf{Conc}(B) \notin \mathcal{K} \cup \{\mathsf{Conc}(A)\}\}$. Then the explanation contains the sub-conclusions that were derived in the construction of the argument.

Example 5. Consider the AF $\mathcal{AF}(\mathrm{AT}_1)$ from Examples 1 and 3. We have that:

- $\mathsf{PrfAcc}(A) \in \{\{C\}, \{E\}, \{C, E\}\}$ and $\mathsf{PrfAcc}(B) = \{D, F\}$;
- $\mathsf{PrfNotAcc}(A) = \{B, D, F\}$ and $\mathsf{PrfNotAcc}(B) = \{C, E\}$.

If we take $\mathbb{F} = \mathsf{Prem}$, then: $\mathsf{PrfAcc}(u) \in \{\{r, s\}, \{r, t\}, \{r, s, t\}\}$, $\mathsf{PrfAcc}(\neg n(d_1)) = \mathsf{PrfNotAcc}(u) = \{v\}$ and $\mathsf{PrfNotAcc}(\neg n(d_1)) = \{r, s, t\}$.

A conclusion derived from an argumentation system can have many causes and therefore many explanations. When humans derive the same conclusion and are asked to explain that conclusion they are able to select *the* explanation from all the possible explanations. In the social sciences a large amount of possible selection criteria that humans might apply have been investigated, see [15] for an overview. In this paper we focus on necessity and sufficiency.

4 Necessity and Sufficiency

Necessity and sufficiency in the context of philosophy and cognitive science are discussed in, for example, [13,14,20]. Intuitively, an event Γ is sufficient for Δ, if no other causes are required for Δ to happen, while Γ is necessary for Δ, if in order for Δ to happen, Γ has to happen as well. In the context of logical implication (denoted by $\rightarrow$), one could model sufficiency by $\Gamma \rightarrow \Delta$ and necessity by $\Delta \rightarrow \Gamma$ [12].

In the next sections we formulate these logical notions in our argumentation setting. We will assume that the arguments on which the explanation for an argument A is based are relevant for A: $B \in \mathsf{Args}$ [resp. $\mathsf{S} \subseteq \mathsf{Args}$] is *relevant* for A if B (in)directly attacks or defends A (i.e., there is a path from B to A) and does not attack itself [resp. for each $C \in \mathsf{S}$, C is relevant for A].

4.1 Necessity and Sufficiency for Acceptance

In the context of argumentation, a set of accepted arguments is sufficient if it guarantees, independent of the status of other arguments, that the considered argument is accepted, while an accepted argument is necessary if it is impossible to accept the considered argument without it.

Definition 9. *Let $\mathcal{AF} = \langle \mathit{Args}, \mathit{Att} \rangle$ be an AF and let $A \in \mathit{Args}$ be accepted (w.r.t. some Sem). Then:*

- $\mathsf{S} \subseteq \mathit{Args}$ *is* sufficient for the acceptance *of A if S is relevant for A, S is conflict-free and S defends A against all its attackers;*
- $B \in \mathit{Args}$ *is* necessary for the acceptance *of A if B is relevant for A and if $B \notin \mathcal{E}$ for some $\mathcal{E} \in \mathsf{Adm}(\mathcal{AF})$, then $A \notin \mathcal{E}$.*

We denote by $\mathsf{Suff}(A) = \{\mathsf{S} \subseteq \mathit{Args} \mid \mathsf{S}$ is sufficient for the acceptance of $A\}$ the set of all sufficient sets of arguments for the acceptance of A and by $\mathsf{Nec}(A) = \{B \in \mathit{Args} \mid B$ is necessary for the acceptance of $A\}$ the set of all necessary arguments for the acceptance of A.

Example 6. In $\mathcal{AF}_1$ both $\{C\}$ and $\{E\}$ are sufficient for the acceptance of A but neither is necessary, while for B, $\{D, F\}$ is sufficient and D and F are necessary.

Necessary and sufficient explanations are now defined by replacing $\mathsf{Defending}$ in the basic explanations from Sect. 3.2 with Nec resp. Suff.

Definition 10. *Let $\mathcal{AF} = \langle \mathit{Args}, \mathit{Att} \rangle$ be an AF and let $A \in \mathit{Args}$ [resp. $\phi \in \mathcal{L}$] be accepted. Then sufficient explanations are defined by:*

- $\mathsf{Acc}(A) \in \mathsf{Suff}(A)$;
- $\mathsf{Acc}(\phi) \in \bigcup\{\mathbb{F}(\mathsf{Suff}(A)) \mid A \in \mathit{Args}$ *and* $\mathsf{Conc}(A) = \phi\}$.

Necessary explanations are defined by:

- $\mathsf{Acc}(A) = \mathsf{Nec}(A)$;

– $\mathsf{Acc}(\phi) = \bigcap\{\mathbb{F}(\mathsf{Suff}(A)) \mid A \in \mathit{Args} \text{ and } \mathsf{Conc}(A) = \phi\}$.

Example 7. For $\mathcal{AF}(\mathrm{AT}_1)$ we have that sufficient explanations are $\mathsf{Acc}(A) \in \{\{C\}, \{E\}, \{C, E\}, \{C, F\}, \{D, E\}\}$, $\mathsf{Acc}(B) = \{D, F\}$, $\mathsf{Acc}(u) \in \{\{r, s\}, \{r, t\}, \{r, s, t\}\}$ and $\mathsf{Acc}(\neg n(d_1)) = \{v\}$. Moreover, necessary explanations are $\mathsf{Acc}(A) = \emptyset$, $\mathsf{Acc}(B) = \{D, F\}$, $\mathsf{Acc}(u) = \{r\}$ and $\mathsf{Acc}(\neg n(d_1)) = \{v\}$.

Next we show that the sets in $\mathsf{Suff}(A)$ are admissible and contain all the needed arguments. Additionally, we look at conditions under which Suff and Nec are empty, as well as the relation between Suff and Nec. These last results provide the motivation for the necessary formula acceptance explanation.[6]

Proposition 1. *Let $\mathcal{AF} = \langle \mathit{Args}, \mathit{Att}\rangle$ be an AF and let $A \in \mathit{Args}$ be accepted w.r.t. some $\mathsf{Sem} \in \{\mathsf{Adm}, \mathsf{Cmp}, \mathsf{Grd}, \mathsf{Prf}, \mathsf{Sstb}\}$. Then:*

1. *For all $\mathsf{S} \in \mathsf{Suff}(A)$, $\{\mathsf{S}, \mathsf{S} \cup \{A\}\} \subseteq \mathsf{Adm}(\mathcal{AF})$;*
2. *$\mathsf{Suff}(A) = \emptyset$ iff there is no $B \in \mathit{Args}$ such that $(B, A) \in \mathit{Att}$.*
3. *$\mathsf{Nec}(A) = \emptyset$ iff there is no $B \in \mathit{Args}$ such that $(B, A) \in \mathit{Att}$ or $\bigcap \mathsf{Suff}(A) = \emptyset$.*
4. *$\mathsf{Nec}(A) \subseteq \bigcap \mathsf{Suff}(A)$.*

The next proposition relates the introduced notions of necessity and sufficiency with Defending and therefore with the basic explanations from Sect. 3.2.

Proposition 2. *Let $\mathcal{AF} = \langle \mathit{Args}, \mathit{Att}\rangle$ be an AF and let $A \in \mathit{Args}$ be accepted w.r.t. $\mathsf{Sem} \in \{\mathsf{Adm}, \mathsf{Cmp}, \mathsf{Grd}, \mathsf{Prf}, \mathsf{Stb}\}$. Then:*

– *for all $\mathcal{E} \in \mathsf{Sem}(\mathcal{AF})$ such that $A \in \mathcal{E}$, $\mathsf{Defending}(A, \mathcal{E}) \in \mathsf{Suff}(A)$;*
– *$\bigcap_{\mathcal{E} \in \mathsf{Sem}(\mathcal{AF}) \text{ and } A \in \mathcal{E}} \mathsf{Defending}(A, \mathcal{E}) = \mathsf{Nec}(A)$.*

4.2 Necessity and Sufficiency for Non-acceptance

When looking at the non-acceptance of an argument A, the acceptance of any of its direct attackers is a sufficient explanation. However, other arguments (e.g., some of the indirect attackers) might be sufficient as well. An argument is necessary for the non-acceptance of A, when it is relevant and A is accepted in the argumentation framework without it. In what follows we will assume that $(A, A) \notin \mathsf{Att}$, since otherwise A itself is the reason for its non-acceptance.

We need the following definition for our notion of sufficiency.

Definition 11. *Let $\mathcal{AF} = \langle \mathit{Args}, \mathit{Att}\rangle$ be an AF and let $A, B \in \mathit{Args}$ such that A indirectly attacks B, via $C_1, \ldots, C_n \in \mathit{Args}$, i.e., $(A, C_1), (C_1, C_2), \ldots, (C_n, B) \in \mathit{Att}$. It is said that the attack from A on B is* uncontested *if there is no $D \in \mathit{Args}$ such that $(D, C_{2i}) \in \mathit{Att}$ for $i \in \{1, \ldots, \frac{n}{2}\}$. It is* contested *otherwise, in which case it is said that the attack from A is contested in C_{2i}.*

The need for the above definition is illustrated in the next example:

[6] Full proofs of our results can be found in the online technical appendix at: https://nationaal-politielab.sites.uu.nl/necessary-sufficient-explanations-proofs/.

Example 8. In $\mathcal{AF}_1$, the indirect attacks from D and F on A are contested: the attack from D is contested in B, since $(E, B) \in \mathsf{Att}$ and the attack from F is also contested in B since $(C, B) \in \mathsf{Att}$. It is therefore possible that A and D or F are part of the same extension (recall Example 2).

For the definition of necessity for non-acceptance we define subframeworks, which are needed because an argument might be non-accepted since it is attacked by an accepted or by another non-accepted argument.[7]

Definition 12. *Let $\mathcal{AF} = \langle \mathsf{Args}, \mathsf{Att} \rangle$ be an AF and let $A \in \mathsf{Args}$. Then $\mathcal{AF}_{\downarrow A} = \langle \mathsf{Args} \setminus \{A\}, \mathsf{Att} \cap (\mathsf{Args} \setminus \{A\} \times \mathsf{Args} \setminus \{A\}) \rangle$ denotes the AF based on $\mathcal{AF}$ but without A.*

Since indirect attacks might be sufficient for not accepting an argument, but they also might be contested, the definition of sufficiency for non-acceptance is defined inductively.

Definition 13. *Let $\mathcal{AF} = \langle \mathsf{Args}, \mathsf{Att} \rangle$ be an AF and let $A \in \mathsf{Args}$ be non-accepted (w.r.t. Sem). Then:*

- $\mathsf{S} \subseteq \mathsf{Args}$ *is* sufficient for the non-acceptance *of A if S is relevant for A and there is a $B \in \mathsf{S}$ such that:*
  - $(B, A) \in \mathsf{Att}$*; or*
  - *B indirectly attacks A and that attack is uncontested; or*
  - *B indirectly attacks A and for every argument C in which the attack from B on A is contested and every $D \in \mathsf{Args}$ such that $(D, C) \in \mathsf{Att}$, there is an $\mathsf{S}' \subseteq \mathsf{S}$ that is sufficient for the non-acceptance of D.*
- $B \in \mathsf{Args}$ *is* necessary for the non-acceptance *of A if B is relevant for A and A is accepted w.r.t. Sem in $\mathcal{AF}_{\downarrow B}$.*

We denote by $\mathsf{SuffNot}(A) = \{\mathsf{S} \subseteq \mathsf{Args} \mid \mathsf{S}$ is sufficient for the non-acceptance of $A\}$ the set of all sufficient sets of arguments for the non-acceptance of A and by $\mathsf{NecNot}(A) = \{B \in \mathsf{Args} \mid B$ is necessary for the non-acceptance of $A\}$ the set of all necessary arguments for the non-acceptance of A.

Example 9. For $\mathcal{AF}_1$ from Example 1 we have that B is both necessary and sufficient for the non-acceptance of A. Moreover, while D and F are neither sufficient for the non-acceptance of A, $\{D, F\}$ is. For the non-acceptance of B we have that C and E are sufficient, but neither is necessary.

Definition 14. *Let $\mathcal{AF} = \langle \mathsf{Args}, \mathsf{Att} \rangle$ be an AF and let $A \in \mathsf{Args}$ [resp. $\phi \in \mathcal{L}$] be non-accepted. Then sufficient explanations are defined by:*

- $\mathsf{NotAcc}(A) \in \mathsf{SuffNot}(A)$*;*
- $\mathsf{NotAcc}(\phi) \in \bigcup\{\mathbb{F}(\mathsf{SuffNot}(A)) \mid A \in \mathsf{Args}$ *and* $\mathsf{Conc}(A) = \phi\}$.

[7] In terms of labeling semantics (see e.g., [2]) an argument is non-accepted if it is `out` (i.e., attacked by an `in` argument) or `undecided`.

Necessary explanations are defined by:

- $\mathsf{NotAcc}(A) = \mathsf{NecNot}(A)$;
- $\mathsf{NotAcc}(\phi) = \bigcap \{\mathbb{F}(\mathsf{SuffNot}(A)) \mid A \in \mathit{Args} \text{ and } \mathsf{Conc}(A) = \phi\}$.

Example 10. For $\mathcal{AF}_1$ we have, for sufficiency $\mathsf{NotAcc}(A) \in \{\{B\}, \{D, F\}, \{B, D, F\}\}$, $\mathsf{NotAcc}(B) \in \{\{C\}, \{E\}, \{C, E\}\}$, $\mathsf{NotAcc}(u) = \{v\}$ and $\mathsf{NotAcc}(\neg n(d_1)) \in \{\{r, s\}, \{r, t\}, \{r, s, t\}\}$ and for necessity $\mathsf{NotAcc}(B) = \emptyset$ and $\mathsf{NotAcc}(u) = \{v\}$.

The next propositions are the non-acceptance counterparts of Propositions 1 and 2. First some basic properties of sufficiency and necessity for non-acceptance.

Proposition 3. *Let $\mathcal{AF} = \langle \mathit{Args}, \mathit{Att} \rangle$ be an AF and let $A \in \mathit{Args}$ be non-accepted w.r.t.* $\mathsf{Sem} \in \{\mathsf{Adm}, \mathsf{Cmp}, \mathsf{Grd}, \mathsf{Prf}, \mathsf{Sstb}\}$. *Then:* $\mathsf{SuffNot}(A) \neq \emptyset$; *and* $\mathsf{NecNot}(A) = \emptyset$ *implies that there are at least two direct attackers of A.*

Now we show how NoDefAgainst (and hence the basic explanations from Sect. 3.2) is related to our notions of sufficiency and necessity for non-acceptance.

Proposition 4. *Let $\mathcal{AF} = \langle \mathit{Args}, \mathit{Att} \rangle$ be an AF and let $A \in \mathit{Args}$ be an argument that is not accepted w.r.t.* $\mathsf{Sem} \in \{\mathsf{Cmp}, \mathsf{Grd}, \mathsf{Prf}, \mathsf{Sstb}\}$. *Then:*

- *for all* $\mathcal{E} \in \mathsf{Sem}(\mathcal{AF})$ *such that* $A \notin \mathcal{E}$, $\mathsf{NoDefAgainst}(A, \mathcal{E}) \in \mathsf{SuffNot}(A)$;
- $\mathsf{NecNot}(A) \subseteq \bigcap_{\mathcal{E} \in \mathsf{Sem}(\mathcal{AF}) \text{ and } A \notin \mathcal{E}} \mathsf{NoDefAgainst}(A, \mathcal{E})$.

4.3 Necessity, Sufficiency and Minimality

In this paper we have introduced necessity and sufficiency to reduce the size of an explanation. More common in the literature is to place a minimality condition on the explanation [7,8]. In this section we show that our notions of necessity and sufficiency result in explanations that do not contain more arguments than the minimal explanations from [7]. To this end we introduce, for $\preceq \in \{\subseteq, \leq\}$:

- $\mathsf{MinDefending}^{\preceq}(A, \mathcal{E}) = \{\mathsf{S} \in \mathsf{Defending}(A, \mathcal{E}) \mid \nexists \mathsf{S}' \in \mathsf{Defending}(A, \mathcal{E})$ such that $\mathsf{S}' \preceq \mathsf{S}\}$ denotes the $\preceq$-minimal $\mathsf{Defending}(A, \mathcal{E})$ sets.[8]
- $\mathsf{MinNotDefAgainst}^{\preceq}(A, \mathcal{E}) = \{\mathsf{S} \in \mathsf{NoDefAgainst}(A, \mathcal{E}) \mid$ there is no $\mathsf{S}' \in \mathsf{NoDefAgainst}(A, \mathcal{E})$ such that $\mathsf{S}' \preceq \mathsf{S}\}$, denotes the set with all $\preceq$-minimal $\mathsf{NoDefAgainst}(A, \mathcal{E})$ sets.
- $\mathsf{MinSuff}^{\preceq}(A) = \{\mathsf{S} \in \mathsf{Suff}(A) \mid \nexists \mathsf{S}' \in \mathsf{Suff}(A)$ such that $\mathsf{S}' \preceq \mathsf{S}\}$, denotes the set of all $\preceq$-minimally sufficient sets for the acceptance of A.
- $\mathsf{MinSuffNot}^{\preceq}(A) = \{\mathsf{S} \in \mathsf{SuffNot}(A) \mid \nexists \mathsf{S}' \in \mathsf{SuffNot}(A)$ such that $\mathsf{S}' \preceq \mathsf{S}\}$, denotes the set of all $\preceq$-minimally SuffNot sets for the non-acceptance of A.

The next example shows that sufficient explanations can be smaller than minimal basic explanations.

[8] In view of the result in [5], these sets correspond to the minimal ($\preceq = \leq$ and where $\mathsf{S} \leq \mathsf{S}'$ denotes $|\mathsf{S}| \leq |\mathsf{S}'|$) and compact ($\preceq = \subseteq$) explanations from [7].

Example 11. Let $\mathcal{AF}_2 = \langle \mathsf{Args}_2, \mathsf{Att}_2 \rangle$, shown in Fig. 2. Here we have that $\mathsf{Prf}(\mathcal{AF}_2) = \{\{A, B\}, \{C, D\}\}$ and that $\mathsf{PrfAcc}(B) = \{A, B\}$, $\mathsf{PrfAcc}(D) = \{C, D\}$, $\mathsf{PrfNotAcc}(B) = \{C, D\}$ and $\mathsf{PrfNotAcc}(D) = \{A, B\}$. These are the explanations for B and D, whether as defined in Sect. 3.2 or with Defending [resp. NoDefAgainst] replaced by MinDefending [resp. MinNotDefAgainst].

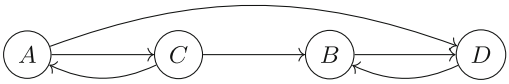

Fig. 2. Graphical representation of $\mathcal{AF}_2$.

When looking at minimally sufficient sets instead, we have that $\mathsf{Acc}(B) = \{A\}$ and $\mathsf{NotAcc}(D) \in \{\{A\}, \{B\}\}$. To see that these explanations are still meaningful, note that A defends B against all of its attackers and as soon as A is accepted under complete semantics, B will be accepted as well. Thus, the minimally sufficient explanations for B and D are $\leq$- and $\subseteq$-smaller than the minimal basic explanations for B and D, but still meaningful.

That minimally sufficient explanations can be smaller than minimal explanations is formalized in the next propositions.

Proposition 5. *Let $\mathcal{AF} = \langle \mathit{Args}, \mathit{Att} \rangle$ be an AF, let $A \in \mathit{Args}$ be accepted w.r.t.* $\mathsf{Sem} \in \{\mathsf{Adm}, \mathsf{Cmp}, \mathsf{Grd}, \mathsf{Prf}, \mathsf{Sstb}\}$ *and let $\preceq \in \{\subseteq, \leq\}$. Then:*

- *for all $\mathcal{E} \in \mathsf{Sem}(\mathcal{AF})$ and all $\mathsf{S} \in \mathsf{MinDefending}^{\preceq}(A, \mathcal{E})$ there is some $\mathsf{S}' \in \mathsf{MinSuff}^{\preceq}(A)$ such that $\mathsf{S}' \preceq \mathsf{S}$;*
- *for all $\mathcal{E} \in \mathsf{Adm}(\mathcal{AF})$ and all $\mathsf{S} \in \mathsf{MinSuff}^{\preceq}(A)$ also $\mathsf{S} \in \mathsf{MinDefending}^{\preceq}(A, \mathcal{E})$;*
- *for all $\mathcal{E} \in \mathsf{Sem}(\mathcal{AF})$ and all $\mathsf{S} \in \mathsf{MinDefending}^{\preceq}(A, \mathcal{E})$, $\mathsf{Nec}(A) \subseteq \mathsf{S}$.*

Proposition 6. *Let $\mathcal{AF} = \langle \mathit{Args}, \mathit{Att} \rangle$ be an AF and let $A \in \mathit{Args}$ be not accepted w.r.t.* $\mathsf{Sem} \in \{\mathsf{Cmp}, \mathsf{Grd}, \mathsf{Prf}, \mathsf{Sstb}\}$. *Then:*

- *for all $\mathcal{E} \in \mathsf{Sem}(\mathcal{AF})$ and all $\mathsf{S} \in \mathsf{MinNotDefAgainst}^{\preceq}(A, \mathcal{E})$, there is some $\mathsf{S}' \in \mathsf{MinSuffNot}^{\preceq}(A)$ such that $\mathsf{S}' \preceq \mathsf{S}$;*
- *for all $\mathcal{E} \in \mathsf{Sem}(\mathcal{AF})$ and all $\mathsf{S} \in \mathsf{MinNotDefAgainst}^{\preceq}(A, \mathcal{E})$, $\mathsf{NecNot}(A) \subseteq \mathsf{S}$.*

5 Applying Necessity and Sufficiency

At the Dutch National Police several argumentation-based applications have been implemented [4]. These applications are aimed at assisting the police at working through high volume tasks, leaving more time for tasks that require human attention. In this section we illustrate how necessity and sufficiency can be applied in the online trade fraud application from [16].

Consider the following language $\mathcal{L}_3$: the complainant delivered (*cd*), the counterparty delivered (*cpd*); the received product seems fake (*fake*); a package is

expected (*pex*); the complainant waited before filing the complaint (*wait*); the received packages is indeed fake (*recfake*); the delivery may still arrive (*deco*); it is a case of fraud (f); and their negations. Based on Dutch Criminal Law (i.e., Article 326) we can derive the following arguments:

$$A_1 : cpd \quad A_2 : \neg cpd \quad A_3 : fake \quad A_4 : \neg fake \quad A_5 : pex \quad A_6 : \neg pex$$
$$A_7 : wait \quad A_8 : \neg wait \quad A_9 : cd \quad A_{10} : \neg cd \quad B_1 : A_1, A_3 \Rightarrow recfake$$
$$B_2 : A_2, A_6 \Rightarrow \neg deco \quad B_3 : A_2, A_5, A_7 \Rightarrow \neg deco \quad B_4 : A_5, A_8 \Rightarrow deco$$
$$C_1 : A_9, B_1 \Rightarrow f \quad C_2 : A_2, A_9, B_2 \Rightarrow f \quad C_3 : A_2, A_9, B_3 \Rightarrow f$$
$$C_4 : A_9, B_4 \Rightarrow \neg f \quad C_5 : A_4, A_9 \Rightarrow \neg f \quad C_6 : A_{10} \Rightarrow \neg f.$$

The above arguments are only a small subset of the possible arguments in the actual application, yet this framework already results in 30 preferred and semi-stable extensions. We can therefore not provide a detailed formal analysis. However, we can already show the usefulness of necessary and sufficient explanations.

The necessary explanation for the acceptance of f is *cd*, while for the acceptance of $\neg f$ the necessary explanation is empty. The reason for this is that, by Article 326, the complainant must have delivered (e.g., sent the goods or money) before it is a case of fraud but $\neg f$ can be accepted for a variety of reasons. In the basic explanations it is not possible to derive this explanation, yet it can be the sole reason for not accepting f. Moreover, minimal sufficient explanations for the acceptance of $\neg f$ when *cd* and $\mathbb{F} = \mathsf{Prem}$ are $\{cd, pex, \neg wait\}$ and $\{cd, \neg fake\}$, these are both $\subset$ and $<$-smaller than any basic explanation for the acceptance of $\neg f$, while still providing the main reasons for the acceptance of $\neg f$.

Therefore, with necessary and sufficient explanations, we can provide compact explanations that only contain the core reasons for a conclusion, something which is not possible with the (minimal) explanations from the basic framework.

6 Conclusion

We have discussed how the explanations from the basic framework introduced in [5] can be adjusted to account for findings from the social sciences on necessary and sufficient explanations [13,14,20]. To this end we have introduced necessary and sufficient sets of arguments for the (non-)acceptance of an argument and formula and integrated these into the explanations definition. The result is a meaningful reduction in the size of an explanation, which almost always exists. Moreover, we have shown that our necessary and sufficient explanations can be smaller than the minimal explanations from [7] and reduce the explanations to the core reasons of (not) accepting a conclusion in a real-life application.

To the best of our knowledge this is the first investigation into necessary and sufficient sets for (non-)acceptance of arguments, especially in the context of integrating findings from the social sciences (e.g., [12]) into (explanations for) argumentation-based conclusions and into a real-life application. In future work we plan to investigate how to integrate further findings, such as contrastiveness and other selection mechanisms.

References

1. Atkinson, K., et al.: Towards artificial argumentation. AI Mag. **38**(3), 25–36 (2017)
2. Baroni, P., Caminada, M., Giacomin, M.: Abstract argumentation frameworks and their semantics. In: Baroni, P., Gabay, D., Giacomin, M., van der Torre, L. (eds.) Handbook of Formal Argumentation, pp. 159–236. College Publications (2018)
3. Besnard, P., et al.: Introduction to structured argumentation. Argum. Comput. **5**(1), 1–4 (2014)
4. Bex, F., Testerink, B., Peters, J.: AI for online criminal complaints: from natural dialogues to structured scenarios. In: Workshop Proceedings of Artificial Intelligence for Justice at ECAI 2016, pp. 22–29 (2016)
5. Borg, A., Bex, F.: A basic framework for explanations in argumentation. IEEE Intell. Syst. **36**(2), 25–35 (2021)
6. Dung, P.M.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artif. Intell. **77**(2), 321–357 (1995)
7. Fan, X., Toni, F.: On computing explanations in argumentation. In: Bonet, B., Koenig, S. (eds.) Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI 2015), pp. 1496–1502. AAAI Press (2015)
8. Fan, X., Toni, F.: On explanations for non-acceptable arguments. In: Black, E., Modgil, S., Oren, N. (eds.) TAFA 2015. LNCS (LNAI), vol. 9524, pp. 112–127. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-28460-6_7
9. García, A., Chesñevar, C., Rotstein, N., Simari, G.: Formalizing dialectical explanation support for argument-based reasoning in knowledge-based systems. Expert Syst. Appl. **40**(8), 3233–3247 (2013)
10. Lacave, C., Diez, F.J.: A review of explanation methods for heuristic expert systems. Knowl. Eng. Rev. **19**(2), 133–146 (2004)
11. Liao, B., van der Torre, L.: Explanation semantics for abstract argumentation. In: Prakken, H., Bistarelli, S., Santini, F., Taticchi, C. (eds.) Proceedings of the 8th International Conference on Computational Models of Argument (COMMA 2020). Frontiers in Artificial Intelligence and Applications, vol. 326, pp. 271–282. IOS Press (2020)
12. Lin, F.: On strongest necessary and weakest sufficient conditions. Artif. Intell. **128**(1), 143–159 (2001)
13. Lipton, P.: Contrastive explanation. R. Inst. Philos. Suppl. **27**, 247–266 (1990)
14. Lombrozo, T.: Causal-explanatory pluralism: how intentions, functions, and mechanisms influence causal ascriptions. Cogn. Psychol. **61**(4), 303–332 (2010)
15. Miller, T.: Explanation in artificial intelligence: insights from the social sciences. Artif. Intell. **267**, 1–38 (2019)
16. Odekerken, D., Borg, A., Bex, F.: Estimating stability for efficient argument-based inquiry. In: Prakken, H., Bistarelli, S., Santini, F., Taticchi, C. (eds.) Proceedings of the 8th International Conference on Computational Models of Argument (COMMA 2020). Frontiers in Artificial Intelligence and Applications, vol. 326, pp. 307–318. IOS Press (2020)
17. Prakken, H.: An abstract framework for argumentation with structured arguments. Argum. Comput. **1**(2), 93–124 (2010)
18. Samek, W., Wiegand, T., Müller, K.R.: Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models. arXiv preprint arXiv:1708.08296 (2017)

19. Saribatur, Z., Wallner, J., Woltran, S.: Explaining non-acceptability in abstract argumentation. In: Proceedings of the 24th European Conference on Artificial Intelligence (ECAI 2020). Frontiers in Artificial Intelligence and Applications, vol. 325, pp. 881–888. IOS Press (2020)
20. Woodward, J.: Sensitive and insensitive causation. Philos. Rev. **115**(1), 1–50 (2006)