Necessary and Sufficient Explanations for Argumentation-Based Conclusions, Extended Abstract*

AnneMarie Borg[1] [0000−0002−7204−6046] and Floris Bex[1,2] [0000−0002−5699−9656]

[1] Department of Information and Computing Sciences, Utrecht University
[2] Tilburg Institute for Law, Technology, and Society, Tilburg University
{A.Borg,F.J.Bex}@UU.nl

Abstract. In this paper, we discuss *necessary* and *sufficient* explanations – the question whether and why a certain argument can be accepted (or not) – for computational 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? Necessity and sufficiency are two of the selection criteria humans use to select *the* explanation from a possibly infinite set of potential explanations. This work is therefore a step towards good, human-understandable argumentation-based explanations.

Keywords: Computational argumentation · Explainable artificial intelligence.

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 [12]. Computational argumentation is one research area in symbolic AI that is frequently mentioned in relation to XAI [7, 20]. For example, arguments can be used to provide reasons for or against decisions [1, 12, 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, e.g., [9–11, 19]. It is the latter type of explanations that is the subject of this paper.

The explanations framework that is introduced in [3] is designed to provide explanations for the (non-)acceptance of arguments. 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 expla-

* This research was partially funded by the Dutch Ministry of Justice and the Netherlands Police. This paper is a shortened version of [5].

nations [15]. Therefore, in [5], we looked at how to select minimal,[3] necessary and sufficient explanations for the (non-)acceptance of an argument.

Minimality is already integrated in some of the existing explanation methods (see, e.g., [6, 9, 10]). However, as we have shown in [5] by introducing the notions necessity and sufficiency we can reduce the size even further, while keeping the explanations meaningful. 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 discuss the potential of the basic framework from [3] when it comes to implementing findings from the social sciences. In particular, this is a shortened version of [5] in which we introduced necessary and sufficient explanations and discussed a real-life application from the Netherlands Police. In addition to making a start at implementing findings from the social sciences, with this work we aim to initiate a discussion on good argumentative explanations.

1 The Framework

We start by recalling the necessary argumentation preliminaries and the basic framework from [3].

Preliminaries. An *abstract argumentation framework* (AF) [8] is a pair $\mathcal{AF} = \langle \text{Args}, \mathcal{A} \rangle$, where Args is a set of *arguments* and $\mathcal{A} \subseteq \text{Args} \times \text{Args}$ is an *attack relation* on these arguments.

Given an AF $\mathcal{AF} = \langle \text{Args}, \mathcal{A} \rangle$, Dung-style semantics can be applied to it, to determine what combination of arguments (called *extensions*) can collectively be accepted from $\mathcal{AF}$. Let $\mathsf{S} \subseteq \text{Args}$ be a set of arguments and let $A \in \text{Args}$. Then S *attacks* A if there is an $A' \in \mathsf{S}$ such that $(A', A) \in \mathcal{A}$; 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 $(A_1, A_2) \in \mathcal{A}$; and S is *admissible* (Adm) if it is conflict-free and it defends all of its elements. An admissible set that contains all the arguments that it defends is a *complete extension* (Cmp) of $\mathcal{AF}$. 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. We denote by $\mathsf{Sem}(\mathcal{AF})$ the set of all the extensions of $\mathcal{AF}$ under the semantics $\mathsf{Sem} \in \{\mathsf{Adm}, \mathsf{Cmp}, \mathsf{Grd}, \mathsf{Prf}\}$.

Where $\mathcal{AF} = \langle \text{Args}, \mathcal{A} \rangle$ is an AF and Sem a semantics, it is said that $A \in \text{Args}$ is *accepted* if $A \in \bigcup \mathsf{Sem}(\mathcal{AF})$ and *non-accepted* if $A \notin \bigcap \mathsf{Sem}(\mathcal{AF})$.

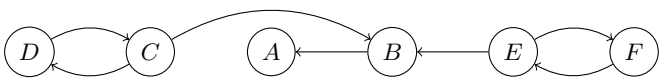

Fig. 1. Graphical representation of the AF $\mathcal{AF}_1$.

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

Example 1. Figure 1 represents $\mathcal{AF}_1 = \langle \text{Args}_1, \mathcal{A}_1 \rangle$ where $\text{Args}_1 = \{A, B, C, D, E, F\}$ and $\mathcal{A}_1 = \{(B, A), (C, B), (C, D), (D, C), (E, B), (E, F), (F, E)\}$. We have that $\mathsf{Grd}(\mathcal{AF}_1) = \{\emptyset\}$ and there are four preferred 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 $\mathsf{Sem} = \mathsf{Prf}$.

Basic definitions of the explanations. In [3] four types of explanations were introduced. These explanations are defined in terms of the function $\mathbb{D}$, which determines the arguments that are in the explanation. For the basic explanations in this paper, we instantiate $\mathbb{D}$ with the following functions, let $A \in \text{Args}$ and $\mathcal{E} \in \mathsf{Prf}(\mathcal{AF})$ for some AF $\mathcal{AF} = \langle \text{Args}, \mathcal{A} \rangle$:[4]

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

The explanations are defined in terms of the above functions, we refer the interested reader to [3] for variations of these functions.

Definition 1. *Let $\mathcal{AF} = \langle Args, \mathcal{A} \rangle$ be an AF and suppose that $A \in Args$ is accepted w.r.t.* Sem. *Then:* $\mathsf{SemAcc}(A) = \{\mathsf{DefBy}(A, \mathcal{E}) \mid \mathcal{E} \in \mathsf{Sem}(\mathcal{AF}) \text{ and } A \in \mathcal{E}\}$.

An acceptance explanation contains all the arguments that defend the argument in an extension.

Definition 2. *Let $\mathcal{AF} = \langle Args, \mathcal{A} \rangle$ be an AF and suppose that $A \in Args$ is non-accepted w.r.t.* Sem. *Then:* $\mathsf{SemNotAcc}(A) = \bigcup_{\mathcal{E} \in \mathsf{Sem}(\mathcal{AF}) \text{ and } A \notin \mathcal{E}} \mathsf{NotDef}(A, \mathcal{E})$.

A non-acceptance explanation contains all the arguments that attack the argument and to which no defense exists in some Sem-extension.

Example 2. Consider the AF $\mathcal{AF}_1$ from Example 1. We have that: $\mathsf{PrfAcc}(A) \in \{\{C\}, \{E\}, \{C, E\}\}$ and $\mathsf{PrfAcc}(B) = \{D, F\}$; $\mathsf{PrfNotAcc}(A) = \{B, D, F\}$ and $\mathsf{PrfNotAcc}(B) = \{C, E\}$.

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. The explanations framework presented in [3] and recalled here is designed to be general in that it can easily be adjusted to incorporate these selection criteria and other findings from the social sciences. We illustrate this here with necessity and sufficiency.

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

2 Studying and Applying the Basic Framework

Necessity and Sufficiency. We will assume that the arguments in an explanation for an argument A are relevant for A: $B \in \text{Args}$ [resp. $\mathsf{S} \subseteq \text{Args}$] 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].

Definition 3. *Given an AF* $\mathcal{AF} = \langle Args, \mathcal{A} \rangle$ *and an argument* A *that is accepted w.r.t. some* $\mathsf{Sem} \in \{\mathsf{Grd}, \mathsf{Cmp}, \mathsf{Prf}\}$ *we say that:*[5]

- $\mathsf{S} \subseteq Args$ *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.*[6] *The set of all sufficienct sets for* A *is denoted by* $\mathsf{Suff}(A)$.
- $B \in Args$ *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* $A \notin \mathcal{E}$. *The set of all necessary arguments for* A *is denoted by* $\mathsf{Nec}(A)$.

The explanations are then defined as: $\mathsf{Acc}(A) \in \mathsf{Suff}(A)$ *for sufficiency and* $\mathsf{Acc}(A) = \mathsf{Nec}(A)$ *for necessity.*

Example 3. 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. We therefore have that sufficient explanations are $\mathsf{Acc}(A) \in \{\{C\}, \{E\}, \{C, E\}, \{C, F\}, \{D, E\}\}$ and $\mathsf{Acc}(B) = \{D, F\}$. Moreover, necessary explanations are $\mathsf{Acc}(A) = \emptyset$ and $\mathsf{Acc}(B) = \{D, F\}$.

In [5] we study several properties of these necessary and sufficient explanations, including that these explanations can be strictly smaller than the minimal (w.r.t. $\subseteq$) and compact (w.r.t. $\subseteq$) explanations from [9].

Application at the Netherlands Police. At the Netherlands Police several argumentation-based applications have been implemented [2]. These applications are aimed at assisting the police at working through high volume tasks, leaving more time for tasks that require human attention. Here we illustrate how necessity and sufficiency can be applied in the online trade fraud application [16].

Consider the following language: 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 (e.g., the complainant did not deliver ($\neg cd$)). A rule set, based on Dutch Criminal Law (i.e., Article 326), allows to derive further conclusions. For example, from *cpd* and *fake* we can derive *recfake*

[5] Due to space restrictions we only discuss sufficiency and necessity for the acceptance of an argument, we refer to [5] for a discussion on non-acceptance and formulas.

[6] In [9] such a set S is called *related admissible* and it is defined as a new semantics.

(argument B_1 below) and from $\neg fake$ and cd we can derive $\neg f$ (argument C_5 below). In particular, we can derive the following set of arguments:

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

These arguments attack each other, based on the arguments from which they are constructed. For example, C_1, C_2 and C_3 attack the arguments C_4, C_5 and C_6 and vice versa in the conclusion, all A_i attack A_{i+1} for $i \in \{1, 3, 5, 7, 9\}$ (and vice versa) as well as the use of these arguments in other arguments: e.g., A_1 attacks A_2 in B_2, B_3, C_2 and C_3 as well.

The above arguments are only a small subset of the possible arguments in the actual application, yet this framework already results in 30 preferred 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 looking at the knowledge base elements 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.

Conclusion. We have discussed how the generality of the basic framework from [3] can be employed to implement findings from the social sciences as surveyed in [15], by integrating necessity and sufficiency [13, 14, 21]. Due to space restrictions we limited this shortened version of [5] to abstract argumentation, for an investigation in a structured setting (i.e., ASPIC$^+$ [17]) we refer to the full paper. Similarly, in [4] we have introduced contrastive explanations. We consider this the beginning of turning argumentation-based explanations into good explanations that can be integrated in applications with human users.

However, if we really want to employ argumentation-based explanations in XAI, as suggested in [7, 20], more research and discussion is needed on argumentative explanations. We therefore invite other researchers to join the discussion on good argumentation-based explanations and the research of integrating findings from the social sciences into the existing work on argumentation-based explanations.

References

1. Atkinson, K., Baroni, P., Giacomin, M., Hunter, A., Prakken, H., Reed, C., Simari, G., Thimm, M., Villata, S.: Towards Artificial Argumentation. AI magazine **38**(3), 25–36 (2017)
2. 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)
3. Borg, A., Bex, F.: A basic framework for explanations in argumentation. IEEE Intelligent Systems **36**(2), 25–35 (2021)
4. Borg, A., Bex, F.: Explaining arguments at the Dutch National Police. In: Palmirani, M., Rodríguez-Doncel, V., Casanovas, P., Pagallo, U., Sartor (eds.) AI Approaches to the Complexity of Legal Systems. Lecture Notes in Artificial Intelligence, Springer (2021), to appear
5. Borg, A., Bex, F.: Necessary and sufficient explanations for argumentation-based conclusions. In: Proceedings of the 16th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU'21) (2021), to appear, for a version with full proofs see https://nationaalpolitielab.sites.uu.nl/necessary-sufficient-explanations-proofs/
6. Caminada, M., Dunne, P.E.: Minimal strong admissibility: A complexity analysis. In: Prakken, H., Bistarelli, S., Santini, F., Taticchi, C. (eds.) Proceedings of the 8th International Conference on Computational Models of Argument (COMMA'20). Frontiers in Artificial Intelligence and Applications, vol. 326, pp. 135–146. IOS Press (2020)
7. Cyras, K., Rago, A., Albini, E., Baroni, P., Toni, F.: Argumentative XAI: A survey. CoRR **abs/2105.11266** (2021), https://arxiv.org/abs/2105.11266
8. Dung, P.M.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artificial Intelligence **77**(2), 321–357 (1995)
9. 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'15). pp. 1496–1502. AAAI Press (2015)
10. Fan, X., Toni, F.: On explanations for non-acceptable arguments. In: Black, E., Modgil, S., Oren, N. (eds.) Proceedings of the 3rd International Workshop on Theory and Applications of Formal Argumentation, (TAFA'15). pp. 112–127. LNCS 9524, Springer (2015)
11. García, A., Chesñevar, C., Rotstein, N., Simari, G.: Formalizing dialectical explanation support for argument-based reasoning in knowledge-based systems. Expert Systems with Applications **40**(8), 3233–3247 (2013)
12. Lacave, C., Diez, F.J.: A review of explanation methods for heuristic expert systems. The Knowledge Engineering Review **19**(2), 133–146 (2004)
13. Lipton, P.: Contrastive explanation. Royal Institute of Philosophy Supplement **27**, 247–266 (1990)
14. Lombrozo, T.: Causal-explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. Cognitive psychology **61**(4), 303–332 (2010)
15. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence **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'20). 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. Argument & Computation **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'20). Frontiers in Artificial Intelligence and Applications, vol. 325, pp. 881–888. IOS Press (2020)
20. 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
21. Woodward, J.: Sensitive and insensitive causation. Philosophical Review **115**(1), 1–50 (2006)