Explaining Arguments at the Dutch National Police

AnneMarie Borg[1] and Floris Bex[1,2]

[1]Department of Information and Computing Sciences, Utrecht University
[2]Tilburg Institute for Law, Technology, and Society, Tilburg University

Abstract

As AI systems are increasingly applied in real-life situations, it is essential that such systems can give explanations that provide insight into the underlying decision models and techniques. Thus, users can understand, trust and validate the system, and experts can verify that the system works as intended. At the Dutch National Police several applications based on computational argumentation are in use, with police analysts and Dutch citizens as possible users. In this paper we show how a basic framework of explanations aimed at explaining argumentation-based conclusions can be applied to these applications at the police.

1 Introduction

Recently *explainable AI* (XAI) has received much attention, mostly directed at new techniques for explaining decisions of machine learning algorithms [21]. However, explanations also play an important role in (symbolic) knowledge-based systems [12]. One area in symbolic AI which has seen a number of real-world applications lately is formal or computational argumentation [1]. Two central concepts in formal argumentation are *abstract argumentation frameworks* [7] – sets of arguments and the attack relations between them – and *structured* or *logical argumentation frameworks* [2] – where arguments are constructed from a knowledge base and a set of rules and the attack relation is based on the individual elements in the arguments. Common for argumentation frameworks, abstract and structured, is that we can determine their extensions, sets of arguments that can collectively be considered as acceptable, under different semantics [7].

The Dutch National Police employs several applications based on structured argumentation frameworks (a variant of ASPIC$^+$ [20]). One such application concerns complaints by citizens about online trade fraud (e.g., a product bought through a web-shop or on eBay turns out to be fake). The system queries the citizen for various observations, and then determines whether the complaint is a case of fraud [3, 19]. Another related example is a classifier for checking fraudulent web-shops, which gathers information about online shops and thus tries to determine whether they are real (bone fide) or fake (mala fide) shops [18]. These applications are aimed at assisting the police at working through high volume tasks, leaving more time for tasks that require human attention.

Argumentation is often considered to be inherently transparent and explainable. A complete argumentation framework and its extensions is a *global* explanation [8]: what can we conclude from the model as a whole? Such global explanations can be used by argumentation experts to check whether the model works as intended. However, as we have noticed when deploying argumentation systems to be used by lay-users (e.g., citizens, police analysts) at the police, more natural and compact explanations are needed. Firstly, we need ways to explain the (non-)acceptability of *individual arguments*, that is, *local* explanations [8] for particular decisions or conclusions. Secondly, explanations should be *compact*, and contain only the *relevant arguments* which are needed in order to draw a conclusion. Finally, explanation should be *tailored to the receiver*. For example,

The final authenticated version is available online at https://doi.org/10.1007/978-3-030-89811-3_13.

in the case of online trade fraud, for a citizen the system should return only the observations provided in the report ("this is presumably a case of fraud because you provided the following facts in your report:..."), but for a police analyst the system should also show which (legal) rules were applied and why there were no exceptions in this case ("this is presumably (not) a case of fraud because the following legal rules are not applicable:...").

In this paper, we show how a variety of different local explanations can be derived from an argumentation framework and we provide motivations for the design options. We start with the basic explanations from [4], which are based on concepts from formal argumentation (Section 3.1). We then discuss how explanations can be selected based on sufficiency and necessity (Section 3.2) and how our explanations can be used to create contrastive explanations (i.e., "why P rather than Q") (Section 3.3). Each of the discussed explanations is based on underlying formal definitions that we cannot introduce here in full detail. We refer the interested reader to [4], [6] and [5] respectively.

Our informal exploration has clear ties to recent more formal work on methods to derive explanations for specific conclusions [9, 10, 11, 13, 22]. We apply and extend the framework from [4] here for several reasons. Often, explanations are only defined for a specific semantics [9, 10] and can usually only be applied to abstract argumentation [10, 13, 22],[1] while our framework 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 [2]), the explanations can be further adjusted (something which is not considered at all in the literature). Moreover, explanations from the literature are usually only for acceptance [9, 13] or non-acceptance [10, 22], while with this framework both acceptance and non-acceptance explanations can be derived in a similar way.[2] Finally, to the best of our knowledge, this is the first approach to local explanations for formal argumentation in which necessary, sufficient and contrastive explanations are considered.

The paper is structured as follows: in the next section we recall some of the most basic and important concepts from formal argumentation. Then, in Section 3, the internet trade fraud scenario and the different possible explanations for the derived conclusions are discussed. We conclude in Section 4.

2 Argumentation Preliminaries

We focus in this paper on the intuition behind the explanations introduced in [6, 4] and the motivation for some of the choices that can be made in the derivation of these explanations. We therefore keep the formal definitions and results limited, leaving more space for an informal discussion.

An *abstract argumentation framework* (AF) [7] 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. An AF can be viewed as a directed graph, in which the nodes represent arguments and the arrows represent attacks between arguments (see, e.g., Figure 1 on page 5). Dung-style semantics can be applied to an AF, to determine what combinations of arguments can collectively be accepted.

Definition 1. For $\mathcal{AF} = \langle \text{Args}, \mathcal{A} \rangle$, $A \in \text{Args}$ *attacks* $B \in \text{Args}$ if $(A, B) \in \mathcal{A}$ and $\mathsf{S} \subseteq \text{Args}$ *attacks* B if there is some $C \in \mathsf{S}$ such that $(C, B) \in \mathcal{A}$; A *defends* B if A attacks an attacker of B and S *defends* B if it attacks every attacker of B;[3] 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* if it is conflict-free and it defends all of its arguments.

[1] 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.

[2] An exception to this might be [11]. However, we consider our framework more easily applicable, since it returns sets of arguments rather than sets of dialectical trees, which might contain many arguments.

[3] In [7], attack and defense are defined from a set of arguments to an argument. In this paper we will mainly rely on attack and defense between arguments, since we are interested in the *arguments that defend* a certain argument, rather than whether that argument is *defended by the set of arguments*.

A $\subseteq$-maximal admissible set is a *preferred extension* (Prf) of $\mathcal{AF}$. The set of all preferred extensions of $\mathcal{AF}$ will be denoted by $\mathsf{Prf}(\mathcal{AF})$.

There are different ways in which the conclusions can be drawn from the extensions of a framework. At the police, when drawing a definite conclusion (e.g., someone is guilty) it is important to be completely certain. This means that the application uses a very skeptical approach towards drawing conclusions: only arguments that are part of every complete set are considered conclusions (i.e., the grounded semantics from [7] is used). When considering whether there is the possibility of the conclusion (e.g., it could be a case of fraud), a more credulous approach can be taken. We follow the latter approach here: an argument that is part of some preferred extension can be considered a conclusion or *accepted*.

For example, in the AF from Figure 1 we have that all arguments are accepted (while under the grounded semantics only C_1 would be accepted). In particular, we have the following preferred extensions: $\{A_1, A_2, A_3\}$, $\{A_1, A_2, A_5\}$, $\{A_1, A_3, A_4\}$, and $\{A_3, A_4, A_6\}$.

In abstract argumentation, as defined above, the arguments are abstract entities and the attack relation is pre-defined. In contrast, in structured argumentation, the arguments are derived from a knowledge base and a set of rules and the attack relation is based on the structure of the arguments. Each of the applications that is in use, is based on a variation of ASPIC$^+$, one of the best-known approaches to structured argumentation [20]. In particular, the notions of a language, axioms and defeasible rules are taken from ASPIC$^+$. See [19] for the formal details.[4] In this paper we only provide the preliminaries that are necessary for the explanations. As we will show in the next section, the AF from Figure 1 is based on a structured setting.

Argumentation frameworks in ASPIC$^+$ are constructed from an *argumentation theory*: $\mathrm{AT} = \langle \mathrm{AS}, \mathcal{K} \rangle$, where $\mathrm{AS} = \langle \mathcal{L}, \mathcal{R}, n \rangle$, an *argumentation system*, is a triple of a formal language, a set of defeasible rules and a naming function for these rules, and $\mathcal{K} = \mathcal{K}_n \cup \mathcal{K}_p$ is the *knowledge base* containing the disjoint sets of axioms ($\mathcal{K}_n$) and ordinary premises ($\mathcal{K}_p$). Arguments are constructued from an argumentation theory as follows:

Definition 2. An *argument* A on the basis of an argumentation theory $\mathrm{AT} = \langle \mathrm{AS}, \mathcal{K} \rangle$, where $\mathrm{AS} = \langle \mathcal{L}, \mathcal{R}, n \rangle$ is:

- ϕ if $\phi \in \mathcal{K}$, where $\mathsf{Prem}(A) = \mathsf{Sub}(A) = \{\phi\}$, $\mathsf{Conc}(A) = \phi$ and $\mathsf{TopRule}(A) =$ undefined;
- $A_1, \ldots, A_n \Rightarrow \psi$, if $A_1, \ldots, A_n$ are arguments such that there is a rule $\mathsf{Conc}(A_1), \ldots, \mathsf{Conc}(A_n) \Rightarrow \psi \in \mathcal{R}$.

  $\mathsf{Prem}(A) = \mathsf{Prem}(A_1) \cup \ldots \cup \mathsf{Prem}(A_n)$, $\mathsf{Sub}(A) = \mathsf{Sub}(A_1) \cup \ldots \cup \mathsf{Sub}(A_n) \cup \{A\}$, $\mathsf{Conc}(A) = \psi$, $\mathsf{TopRule}(A) = \mathsf{Conc}(A_1), \ldots, \mathsf{Conc}(A_n) \Rightarrow \psi$ additionally, we denote $\mathsf{Ant}(\mathsf{TopRule}(A)) = \{\mathsf{Conc}(A_1), \ldots, \mathsf{Conc}(A_n)\}$. Moreover, where S is a set of arguments $\mathsf{Prem}(\mathsf{S}) = \bigcup\{\mathsf{Prem}(A) \mid A \in \mathsf{S}\}$.

Attacks between arguments are based on the premises and conclusions of these arguments.

Definition 3. An argument A *attacks* an argument B iff, (where $\phi = -\psi$ iff $\phi = \neg\psi$ or $\psi = \neg\phi$)

- $\mathsf{Conc}(A) = \neg n(d_i)$, where there is some $B' \in \mathsf{Sub}(B)$ such that $\mathsf{TopRule}(B') = d_i$, it denies a rule; or
- $\mathsf{Conc}(A) = -\phi$, where there is some $B' \in \mathsf{Sub}(B)$ such that $\mathsf{Conc}(B') = \phi$, it denies a conclusion; or
- $\mathsf{Conc}(A) = -\phi$, for some $\phi \in \mathsf{Prem}(B) \setminus \mathcal{K}_n$, it denies a premise.

Dung-style semantics can be applied to argumentation frameworks based on argumentation theories as defined in Definition 1. We will say that a formula ϕ in an argumentation framework

[4] The corresponding demo of [19], demonstrating the argumentation-based part of the application, is available at https://nationaal-politielab.sites.uu.nl/estimating-stability-for-efficient-argument-based-inquiry/.

$\mathcal{AF}(\text{AT})$ is *accepted* if there is some $\mathcal{E} \in \mathsf{Prf}(\mathcal{AF}(\text{AT}))$ with $A \in \mathcal{E}$ such that $\mathsf{Conc}(A) = \phi$ and *non-accepted* if there is some $\mathcal{E} \in \mathsf{Prf}(\mathcal{AF}(\text{AT}))$ such that there is no $A \in \mathcal{E}$ with $\mathsf{Conc}(A) = \phi$.

These basic preliminaries on formal argumentation are enough to illustrate the different possibilities for explaining argumentation-based conclusions derived from the internet trade fraud application at the police.

3 Deriving Explanations

Suppose that the following knowledge base is provided: a citizen has *ordered a product* through an online shop, *paid* for it and *received* a package. However, it is the *wrong product*, it seems *suspicious* as if it might be a replica, rather than a real product. Yet an *investigation* cannot find a problem with the product. Still, the citizen wants to file a complaint of internet trade fraud.

While the citizen provides the information from the described scenario, the system constructs further arguments from this, based on the Dutch law.[5] In particular, the following rules are applied:

R_1 If the complainant *paid* then usually the *complainant delivered*;

R_2 If the *wrong product* was *received* then usually this is *not a case of fraud*;

R_3 If the *wrong product* was *received* then usually the *counter party has delivered*;

R_4 If the product seem *suspicious* then usually the product is *fake*;

R_5 If the product is *fake* then usually the *counter party did not deliver*;

R_6 If an *investigation* shows that there is no problem with the product then usually the product is *not fake*;

R_7 If the *complainant delivered* and the *counter party did not deliver* it is usually *a case of fraud*.

From this we obtain arguments for:[6]

C_1 : the complainant *paid* + $R_1 \Rightarrow$ the *complainant delivered*
A_1 : the *wrong product* was *received* + $R_2 \Rightarrow$ it is *not a case of fraud*
A_2 : the *wrong product* was *received* + $R_3 \Rightarrow$ the *counter party has delivered*
A_3 : the product seems *suspicious* + $R_4 \Rightarrow$ the product is *fake*
A_4 : A_3 + $R_5 \Rightarrow$ the *counter party did not deliver*
A_5 : an *investigation* shows no problems + $R_6 \Rightarrow$ the product is *not fake*
A_6 : C_1 + A_4 + $R_7 \Rightarrow$ it is *a case of fraud*.

Note that the argument A_5 which has conclusion *not fake* will attack any argument with the conclusion *fake* (and vice versa), as well as any argument based on the conclusion *fake* (i.e., A_5 and A_3 attack each other and A_5 attacks A_4 and A_6 because they have *fake* as a sub-conclusion). The graphical representation of the AF, which we will refer as $\mathcal{AF}_1 = \langle \text{Args}_1, \mathcal{A}_1 \rangle$ can be found in Figure 1.

As the aim of the system is to determine whether a particular situation is a case of fraud, we will focus here on the arguments A_1 (*not fraud*) and A_6 (*fraud*). Note that, from an argumentative perspective, both arguments can be accepted, though not simultaneously. For A_1 this is the case since A_1 attacks any argument by which it is attacked (i.e., $(A_1, A_6) \in \mathcal{A}$). For A_6 additional conclusions have to be accepted as well. In particular, one can accept the argument for *fraud* when also accepting the arguments for the *counter party did not deliver* (A_4) and that the *product is*

[5] In order to make the argumentation framework and corresponding explanations more interesting the rules that are applied here are only inspired by the law. The real application is based on slightly different rules [19].

[6] We do not state the arguments based on the knowledge base explicitly, since these neither attack other arguments nor can be attacked themselves and do therefore not influence the acceptability of other arguments.

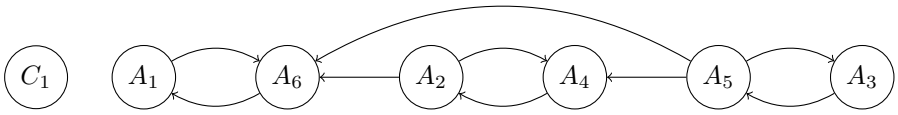

Figure 1: Graphical representation of the argumentation framework $\mathcal{AF}_1$ constructed based on information provided in the complaint.

fake (A_3). This follows since $\{A_3, A_4, A_6\}$ is a preferred extension and A_3 and A_4 attack attackers of A_6 that A_6 would otherwise not be defended against. In what follows we will consider for both A_1 and A_6 explanations for why one could (not) accept them.

3.1 Basic explanations

In [4] skeptical and credulous acceptance and non-acceptance 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 $A \in \text{Args}$ and $\mathcal{E} \in \mathsf{Prf}(\mathcal{AF})$ for some AF $\mathcal{AF} = \langle \text{Args}, \mathcal{A} \rangle$:

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

The explanations are defined for arguments and formulas.

Definition 4. Let $\mathcal{AF} = \langle \text{Args}, \mathcal{A} \rangle$ be an AF and suppose that $A \in \text{Args}$ [resp. $\phi \in \mathcal{L}$] is accepted. Then:

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

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

Definition 5. Let $\mathcal{AF} = \langle \text{Args}, \mathcal{A} \rangle$ be an AF and suppose that $A \in \text{Args}$ [resp. $\phi \in \mathcal{L}$] is non-accepted. Then:

$$\mathsf{NotAcc}(A) = \bigcup_{\mathcal{E} \in \mathsf{Prf}(\mathcal{AF}) \text{ and } A \notin \mathcal{E}} \mathsf{NotDefAgainst}(A, \mathcal{E}).$$
$$\mathsf{NotAcc}(\phi) = \bigcup_{A \in \text{Args and } \mathsf{Conc}(A) = \phi} \quad \bigcup_{\mathcal{E} \in \mathsf{Prf}(\mathcal{AF}) \text{ and } A \notin \mathcal{E}} \mathbb{F}(\mathsf{NotDefAgainst}(A, \mathcal{E})).$$

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 preferred 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 [4]. These will be motivated in the discussions on the different explanations.

- $\mathbb{F} = \mathsf{id}$, 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.
- $\mathbb{F} = \mathsf{ConcSub}$, where $\mathsf{ConcSub}(A) = \{\mathsf{Conc}(B) \mid B \in \mathsf{Sub}(A),\ \mathsf{Conc}(B) \notin \mathcal{K} \cup \{\mathsf{Conc}(A)\}\}$. Then the explanation contains the sub-conclusions that were derived in the construction of the argument.

We can now turn to a discussion on explanations for the (non-)acceptance of (not) fraud.

It is a case of fraud (acceptance of A_6/non-acceptance of A_1). The basic explanation here is that A_6 *can be accepted, when* A_3 *and* A_4 *are accepted as well.* In terms of the conclusions of the arguments, we say that it is *a case of fraud* (A_6), because the product is *fake* (A_3) and the *counter party did not deliver* (A_4). When considering the variations of $\mathbb{F}$, further explanations can be considered. For example, it is a case of fraud, because:

- the *complainant delivered* (C_1) and the *counter party did not deliver* (A_4) and there is a rule (R_7) that states that from these conclusions it can be derived that it is a *case of fraud* (A_6), i.e., $\mathbb{F} = \mathsf{AntTop}$. Such an explanation can be used by an analyst at the police, who is familiar with the rules and wants to understand what parts of the law were applied.
- the *complainant paid* and the product seems *suspicious*, i.e., $\mathbb{F} = \mathsf{Prem}$. At the moment, the system returns this type of explanation, which can be used by the complainant, to understand what parts of the report made the system derive this conclusion.
- the *complainant delivered* (C_1), the *counter party did not deliver* (A_4) and the product is *fake* (A_3), i.e., $\mathbb{F} = \mathsf{ConcSub}$. Explanations like this provide insight into the reasoning process of the system: it shows the sub-steps that were taken. It might be useful for an analyst at the police, who wants more insight into the reasons than only the last step, but also for the complainant, who might not be convinced by an explanation that only contains information provided in the complaint itself.

Similar explanations can be given for not(it is *not a case of fraud*), i.e., that A_1 is not accepted. This follows since the main reason that A_1 cannot be accepted is the fact that A_6 is accepted.

It is not a case of fraud (acceptance of A_1/non-acceptance of A_6). While A_1 can be explained by the acceptance of A_1 (since it can defend itself against the attack from A_6), additional arguments defend A_1 as well (i.e., A_2 and A_5 defend A_1 against the attack from A_6 as well). To give an overview of the possible explanations, we consider here the most extensive set of arguments: A_1, A_2 and A_5. In terms of the conclusions of the arguments, it follows that it is not a case of fraud, because the *counter party has delivered* and the product is *not fake*. Similarly as above, we can also consider other explanations based on elements of arguments: It is not a case of fraud, because:

- the *wrong product* was delivered and there is a rule (R_2) that states that usually, when the wrong product is delivered, it is not a case of fraud, i.e., $\mathbb{F} = \mathsf{AntTop}$. Note that this explanation is the same, whether we consider A_1 to be an explanation for its own acceptance, or the arguments A_2 and A_5 are considered as well.
- the *wrong product* was delivered and an *investigation* shows that there is no problem with the product, i.e., $\mathbb{F} = \mathsf{Prem}$. If A_5 is not a part of the explanation, then this explanation only contains the information that the *wrong product* was delivered.
- the *counter party has delivered* (A_2) and the product is *not fake* (A_5), i.e., $\mathbb{F} = \mathsf{ConcSub}$. Note that, in the case A_1 is its own acceptance explanation, no sub-conclusions are derived in the process.

Like in the case above, the explanations that it is not(*a case of fraud*) is similar to the explanations for *not a case of fraud*. This follows since the argument for *a case of fraud* (A_6) is attacked by each of the arguments considered here (i.e., A_6 is attacked by A_1, A_2 and A_5).

The suggested explanations above are not too extensive for the given example. However, a rule might have many antecedents, a conclusion might be based on many knowledge base elements or the derivation might be long, resulting in many sub-conclusions. It is therefore useful to consider how we can reduce the size of explanations. To this end, it has been argued that humans select their explanations in a biased manner. Selection happens based on e.g., simplicity, generality, robustness – see [17] for an overview on findings for the social sciences on how humans come to their explanations and how this could be applied in artificial intelligence. In the next section we will consider two ways of reducing the size of explanations. Given the space restrictions and since the basic explanations were similar for acceptance and non-acceptance, we only discuss acceptance explanations.

3.2 Necessity and Sufficiency

Necessity and sufficiency in the context of philosophy and cognitive science are discussed in, for example, [14, 15, 23]. 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.[7]

Sufficiency. In terms of arguments, one could say that a set of arguments is *sufficient* for the acceptance of some argument, if by accepting those arguments the argument can also be accepted (i.e., that the set of arguments defends the argument against all its attackers). For example, in the cases above:

- it was already mentioned that the acceptance of A_1 (that it is *not a case of fraud*) can be explained by the argument itself, but also by $\{A_1, A_2\}$, by $\{A_2, A_5\}$ and by $\{A_1, A_2, A_5\}$. Each of these sets is sufficient for the acceptance of A_1. If one were interested in *minimal sufficiency*, then the argument itself would be enough.
- for the argument A_6 (that it *is a case of fraud*) the arguments A_3 and A_4 have to be accepted. Thus there is only one sufficient set: $\{A_3, A_4, A_6\}$.

Formally, given $\mathcal{AF} = \langle \text{Args}, \mathcal{A} \rangle$ and accepted argument $A \in \text{Args}$:

- $\mathsf{S} \subseteq \text{Args}$ is *sufficient for the acceptance* of A if for each $B \in \mathsf{S}$, there is an attack-path from B to A,[8] S is conflict-free and S defends A against all its attackers.

We denote by $\mathsf{Suff}(A) = \{\mathsf{S} \subseteq \text{Args} \mid \mathsf{S} \text{ is sufficient for the acceptance of } A\}$ the set of all sufficient sets of arguments for the acceptance of A. With this sufficient explanations can be defined:

Definition 6. Let $\mathcal{AF} = \langle \text{Args}, \mathcal{A} \rangle$ be an AF and suppose that $A \in \text{Args}$ is accepted. Then: $\mathsf{Acc}(A) \in \mathsf{Suff}(A)$.

For minimally sufficient explanations $\mathsf{Acc}(A) \in \min \mathsf{Suff}(A)$, where minimality can be taken w.r.t. $\subseteq$ or the number of arguments in a set.

The resulting explanations for $\mathcal{AF}_1$ are as described before the formal definitions.

When the structure of the arguments is known we can again look at explanations in terms of the elements of the arguments. Note that when explanations should contain minimal sufficient sets of elements (e.g., minimal sufficient sets of premises or sub-conclusions) one should not simply take the elements of the minimal sufficient set of arguments, but rather compare the sets of elements obtained from each sufficient set and compare those sizes.

[7] See [6] for the technical details, in this paper we focus on the application of necessary and sufficient explanations.

[8] There is an attack path from B to A if there are $C_1, \ldots, C_k \in \text{Args}$ such that $(B, C_1), (C_1, C_2), \ldots, (C_{k-1}, C_k), (C_k, A) \in \mathcal{A}$.

Definition 7. Let $\mathcal{AF}(\text{AT}) = \langle \text{Args}, \mathcal{A} \rangle$ be an AF, based on an argumentation theory AT and suppose that $\phi \in \mathcal{L}$ is accepted. Then:

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

In our example we have that:

- receiving the *wrong product* is sufficient for that it is *not a case of fraud*, if $\mathbb{F} = \mathsf{Prem}$ and, combined with the rule that usually when the wrong product is received it is not a case of fraud, when $\mathbb{F} = \mathsf{AntTop}$.
- the premises that the *complainant paid* and that the product seems *suspicious* are sufficient for that it is *a case of fraud*. When $\mathbb{F} = \mathsf{AntTop}$, the rules from A_3 (if the product seem suspicious then usually the product is fake), A_4 (if the product is fake then usually the counter party did not deliver) and A_6 (if the complainant delivered and the counter party did not deliver it is usually a case of fraud) form the explanation, together with their antecedents that the product seems *suspicious*, the product is *fake*, the *complainant delivered* and the *counter party did not deliver*.

Given the structure of $\mathcal{AF}_1$, there is not much difference between the basic explanations and sufficient explanations. Therefore, we introduce the following example, this time not based on a scenario from the police.

Example 1. Let $\text{AT}_2 = \langle \text{AS}_2, \mathcal{K}_2 \rangle$, where the rules in AS_2 are such that, with $\mathcal{K}_2 = \{r, s, t, v\}$, the following arguments can be derived:[9]

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

See Figure 2 for a graphical representation of the correspond AF $\mathcal{AF}_2$. Note that, like for $\mathcal{AF}_1$, all arguments can be accepted.

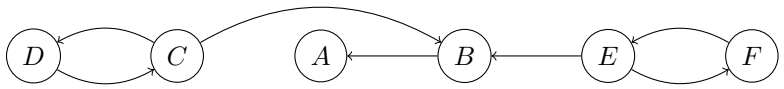

Figure 2: Graphical representation of the abstract argumentation framework $\mathcal{AF}_2$.

On an abstract level, in order to accept A either C or E should be accepted as well. To accept B, one has to accept both D and F. Sufficient explanations for the acceptance of A are $\{C\}$, $\{E\}$, $\{C, E\}$, but also $\{C, F\}$ and $\{D, E\}$ (since these still include C resp. E). Minimally sufficient explanations are $\{C\}$ and $\{E\}$ and $\{D, F\}$ is the only (minimally) sufficient explanation for the acceptance of B.

When looking at the structure of the arguments, taking $\mathbb{F} = \mathsf{Prem}$, we have that $\{r, s\}$, $\{r, t\}$ and $\{r, s, t\}$ are some of the sufficient sets for the acceptance of u and $\{v\}$ is sufficient to accept an exception to the rule d_1.

Necessity. In terms of arguments, an argument can be understood as *necessary* if without that argument, the considered argument could not be accepted. For $\mathcal{AF}_1$, the (minimal) sufficient sets of arguments are also the necessary arguments: A_1 is the only necessary argument for the acceptance of A_1, while there are three arguments necessary for the acceptance of A_6: A_3, A_4 and A_6.

Formally, given an AF $\mathcal{AF} = \langle \text{Args}, \mathcal{A} \rangle$ and $A \in \text{Args}$ an accepted argument:

[9] We ignore again the arguments based on the elements of $\mathcal{K}_2$.

- $B \in$ Args is *necessary for the acceptance* of A if there is an attack-path from B to A and if $B \notin \mathsf{S}$ for some admissible set $\mathsf{S} \subseteq$ Args, then $A \notin \mathsf{S}$.

We denote by $\mathsf{Nec}(A) = \{B \in \text{Args} \mid B \text{ is necessary for the acceptance of } A\}$ the set of all arguments that are necessary for the acceptance of A. With this necessary explanations can be defined:

Definition 8. Let $\mathcal{AF} = \langle \text{Args}, \mathcal{A} \rangle$ be an AF and suppose that $A \in$ Args is accepted. Then: $\mathsf{Acc}(A) = \mathsf{Nec}(A)$.

For an illustration of the difference between sufficiency and necessity, consider the argument A_2. Then $\{A_2\}$ is sufficient for its own acceptance, but $\{A_5\}$ is also sufficient for its acceptance. Therefore, there is no argument that is necessary for the acceptance of A_2 (see also Proposition 2).

Similar reasoning as in the case of sufficiency applies to necessary explanations based on the elements of the arguments. One can collect premises, rules and sub-conclusions from the necessary arguments. However, in terms of elements we can be more detailed. For this we need the following results.

Proposition 1. *Let $\mathcal{AF} = \langle Args, \mathcal{A} \rangle$ and let $A \in Args$ be accepted. Then $\mathsf{Acc}(A) = \emptyset$ iff there is no $B \in Args$ such that $(B, A) \in \mathcal{A}$, where Acc can be defined as in Definition 4 or 6.*

Proposition 2. *Let $\mathcal{AF} = \langle Args, \mathcal{A} \rangle$ and let $A \in Args$ be accepted. Then $\mathsf{Nec}(A) = \emptyset$ if $\bigcap \mathsf{Suff}(A) = \emptyset$.*

While, in view of the above results, a necessary explanation for arguments might be empty, one could still collect necessary premises, rules and sub-conclusions. We therefore define:

Definition 9. Let $\mathcal{AF}(\text{AT}) = \langle \text{Args}, \mathcal{A} \rangle$ be an AF, based on an argumentation theory AT and suppose that $\phi \in \mathcal{L}$ is accepted. Then:

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

To illustrate the difference between necessary and sufficient explanations and the application of the above definition, we return to the AF $\mathcal{AF}_2$ from Example 1.

Example 2. For the AF $\mathcal{AF}_2$ we have that for the acceptance of A no argument is necessary. But, when $\mathbb{F} = \mathsf{Prem}$ we have that r is necessary. For the acceptance of B both D and F are necessary and, when $\mathbb{F} = \mathsf{Prem}$, v is necessary.

3.3 Contrastive explanations

Another relevant way in which humans structure and select their explanations is *contrastiveness* [14, 16, 17]: when people ask 'why P?', they often mean 'why P rather than Q?' – here P is called the fact and Q is called the foil [14]. The answer to the question is then to explain as many of the differences between fact and foil as possible.[10]

When humans provide a contrastive explanation, the foil is not always explicitly stated. While humans are capable of detecting the foil based on context and the way the question is asked, AI-based systems struggle with this.

When the foil is not explicitly stated, formal argumentation has an advantage over some other approaches to AI because it comes with an explicit notion of conflict (i.e., the attack relation). This allows us to derive a foil when none is provided. For example, given an argument one could take as the foil:

- all the arguments that directly attack or defend it;
- all the arguments that directly or indirectly attack or defend it.

[10] See [5] for the technical details, in this paper we focus on the application of contrastive explanations.

In the context of structured arguments, one can also look at the claims of the arguments and take the foil to be arguments with conflicting conclusion.

Given an argument of which the acceptance status should be explained (the fact) and a foil, a contrastive explanation contains those arguments that explain:

- the acceptance of the fact and the non-acceptance of the foil;
- the non-acceptance of the fact and the acceptance of the foil.

Definition 10. Let $\mathcal{AF} = \langle \text{Args}, \mathcal{A} \rangle$ be an AF, let $A \in \text{Args}$ be the fact and $\mathsf{S} \subseteq \text{Args}$ be the foil (for example defined by the direct attacking arguments of A). Suppose that A is accepted [resp. non-accepted] and that each $B \in \mathsf{S}$ is non-accepted [resp. accepted]. Then:

$$\mathsf{Cont}(A, \mathsf{S}) = \mathsf{Acc}(A) \cap \left(\bigcup_{B \in \mathsf{S}} \mathsf{NotAcc}(B) \right)$$

$$\mathsf{ContN}(A, \mathsf{S}) = \mathsf{NotAcc}(A) \cap \left(\bigcup_{B \in \mathsf{S}} \mathsf{Acc}(B) \right).$$

When $\mathsf{Cont}(A, \mathsf{S}) = \emptyset$ (the case for ContN is similar) the explanation will return a pair: $\mathsf{Cont}(A, \mathsf{S}) = \left\langle \mathsf{Acc}(A), \bigcup_{B \in \mathsf{S}} \mathsf{NotAcc}(B) \right\rangle$.

Thus, given explanations for the acceptance [resp. non-acceptance] of the fact and the non-acceptance [resp. acceptance] of the foil the contrastive explanation returns the intersection of these explanations when it is not empty (otherwise it would simply return those two explanations). An empty contrastive explanation rarely happens. In particular:

Proposition 3. *Let $\mathcal{AF} = \langle \text{Args}, \mathcal{A} \rangle$ be an AF. Let $A \in \text{Args}$ and let $\mathsf{S} \subseteq \text{Args}$ be such that for each $B \in \mathsf{S}$, $(B, A) \in \mathcal{A}$. Then $\mathsf{Cont}(A, \mathsf{S}) = \emptyset$ [resp. $\mathsf{ContN}(A, \mathsf{S}) = \emptyset$] implies that $\mathsf{Acc}(A) = \emptyset$ [resp. $\bigcup_{B \in \mathsf{S}} \mathsf{Acc}(B) = \emptyset$].*

Intuitively, this shows that a contrastive explanation is only empty if the fact is not attacked at all [resp. no argument in the set of foils is attacked]. To illustrate contrastive explanations we introduce another scenario, this time about a possible malafide webshop, based on the application in [18].

Example 3. Consider a language $\mathcal{L}_3$, containing the atoms *cf* (a complaint was filed), *m* (the webshop is malafide), *iw* (an investigation is done), *sa* (the url is suspicious), *rc* (the complaint is retracted), *kp* (the webshop owner is known by the police), *ka* (the address is registered at the chamber of commerce), *rr* (the registration was recently retracted) and their negations.

Let $\text{AT}_3 = \langle \text{AS}_3, \mathcal{K}_3 \rangle$, where the rules in AS_2 are such that, with the language $\mathcal{L}_3$ and $\mathcal{K}_3 = \{cf, rc, sa, ka, kp, rr\}$, the following arguments can be derived:

$$\begin{array}{lll} A_1 : cf \quad A_2 : rc & A_3 : sa \quad A_4 : ka & A_5 : kp \quad A_6 : rr \\ B_1 : A_1 \overset{d_1}{\Rightarrow} iw & B_2 : A_2 \overset{d_2}{\Rightarrow} \neg n(d_1) & B_3 : A_5 \overset{d_5}{\Rightarrow} \neg rc \\ B_4 : B_1, A_3 \overset{d_3}{\Rightarrow} m & B_5 : A_4 \overset{d_4}{\Rightarrow} \neg n(d_3) & B_6 : A_6 \overset{d_6}{\Rightarrow} \neg ka. \end{array}$$

See Figure 3 for a graphical representation of the corresponding AF $\mathcal{AF}(\text{AT}_3)$. As in our previous examples, each of the arguments can be accepted.

To start with, we have the following basic explanation for the acceptance of *m* (i.e., the webshop is malafide): the owner of the webshop is known by the police (*kp*) and the registration at the chamber of commerce was recently retracted (*rr*), from which it follows that no exceptions could be derived.

Basic explanations are exhaustive: all the reasons why the webshop is malafide are provided. With our contrastive explanations, the explanation can focus on an explicit contrastive question. For example: the webshop is malafide rather than that there is an exception to rule d_1, since the owner is known by the police (*kp*); and the webshop is malafide rather than that there is an exception to rule d_3, since the registration was recently retracted (*rr*). Thus, the contrastive explanations are better tailored to one question and result in smaller explanations.

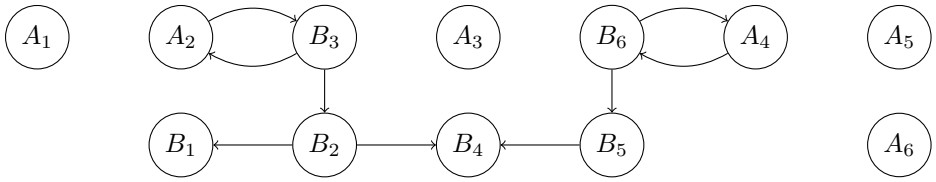

Figure 3: Graphical representation of the AF $\mathcal{AF}(\mathrm{AT}_3)$.

4 Conclusion

In this paper we have discussed how a general framework for explaining conclusions derived from an argumentation framework can be applied on top of the argumentation systems in use at the Dutch National Police. As an example we took the system in use to assist in the processing of complaints on online trade fraud. The ideas presented in this paper can also be applied to the other systems in use at the police as well as any other system based on argumentation frameworks as introduced in [7].

Recall from the introduction that, unlike other approaches to local explanations of argumentation-based conclusions [9, 10, 11, 13, 22], the framework that we applied can capture both acceptance and non-acceptance explanations, is not based on one specific semantics (although we only considered preferred semantics here) and allows to take the structure of arguments into account (i.e., explanations can be sets of premises or rules, rather than just sets of arguments). Moreover, we have shown how our framework can be used to study how findings from the social sciences (those collected in, e.g., [17]) can be implemented. The presented studies of sufficiency, necessity and contrastiveness are just the beginning. On the one hand, especially in the case of contrastive explanations, much more can be said about the individual concepts than we could present here. On the other hand, there are many other aspects of human explanation that have not been investigated yet.

In future work we will continue our study of integrating findings from the social sciences into our explanations. For example, we will study the notion of contrastiveness further, we will look into the robustness of explanations and we will consider further selection criteria. Additionally, for the applications at the Dutch National Police, we will implement the framework and conduct a user study on the best explanations for these specific applications and, possibly, the best explanations for other argumentation-based applications.

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] Besnard, P., Garcia, A., Hunter, A., Modgil, S., Prakken, H., Simari, G., Toni, F.: Introduction to structured argumentation. Arg. & Comp. **5**(1), 1–4 (2014)

[3] 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)

[4] Borg, A., Bex, F.: A basic framework for explanations in argumentation. IEEE Intelligent Systems **36**(2), 25–35 (2021)

[5] Borg, A., Bex, F.: Contrastive explanations for argumentation-based conclusions. arXiv/CoRR **abs/2107.03265** (2021), https://arxiv.org/abs/2107.03265

[6] Borg, A., Bex, F.: Necessary and sufficient explanations for argumentation-based conclusions. In: Vejnarová, J., Wilson, N. (eds.) Proceedings of the 16th European Conference on Symbolic

and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU'21). Lecture Notes in Computer Science, vol. 12897, pp. 45–58. Springer (2021)

[7] 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)

[8] Edwards, L., Veale, M.: Slave to the algorithm: Why a 'right to an explanation' is probably not the remedy you are looking for. Duke Law & Technology Review **16**(1), 18–84 (2017)

[9] Fan, X., Toni, F.: On computing explanations in argumentation. In: Bonet, B., Koenig, S. (eds.) Proceedings of 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 TAFA'15. pp. 112–127. 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] Liao, B., van der Torre, L.: Explanation semantics for abstract argumentation. In: Prakken, H., Bistarelli, S., Santini, F., Taticchi, C. (eds.) Proceedings of COMMA'20. pp. 271–282. IOS Press (2020)

[14] Lipton, P.: Contrastive explanation. Royal Institute of Philosophy Supplement **27**, 247–266 (1990)

[15] Lombrozo, T.: Causal-explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. Cognitive psychology **61**(4), 303–332 (2010)

[16] Miller, T.: Contrastive explanation: A structural-model approach. CoRR **abs/1811.03163** (2018), http://arxiv.org/abs/1811.03163

[17] Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence **267**, 1 – 38 (2019)

[18] Odekerken, D., Bex, F.: Towards transparent human-in-the-loop classification of fraudulent web shops. In: Villata, S., Harašta, J., Křemen, P. (eds.) Proceedings of JURIX 2020. pp. 239–242. IOS Press (2020)

[19] 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 COMMA'20. pp. 307–318. IOS Press (2020)

[20] Prakken, H.: An abstract framework for argumentation with structured arguments. Argument & Computation **1**(2), 93–124 (2010)

[21] Samek, W., Wiegand, T., Müller, K.R.: Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. CoRR **abs/1708.08296** (2017), http://arxiv.org/abs/1708.08296

[22] Saribatur, Z., Wallner, J., Woltran, S.: Explaining non-acceptability in abstract argumentation. In: Proceedings of ECAI'20. pp. 881–888. IOS Press (2020)

[23] Woodward, J.: Sensitive and insensitive causation. Philosophical Review **115**(1), 1–50 (2006)