Explaining Arguments at the Dutch National Police*

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] Department of Law, Technology, Markets and Society, Tilburg University
{a.borg,f.j.bex}@uu.nl

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.

Keywords: Explainable AI · Computational Argumentation

1 Introduction

Recently *explainable AI* (XAI) has received much attention, mostly directed at new techniques for explaining decisions of machine learning algorithms [20]. However, explanations also play an important role in (symbolic) knowledge-based systems [11]. 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* [6] – 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 [6].

The Dutch National Police employs several applications based on structured argumentation frameworks (a variant of ASPIC$^+$ [19]). 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 [3,18]. Another related example is a classifier for checking

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

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 [17]. 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 [7]: 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 [7] 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, 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 in an informal way how a variety of different local explanations can be derived from an argumentation framework. In addition to explanations based on concepts from formal argumentation (e.g., attack and defense), we discuss how explanations can be selected based on sufficiency and necessity. Moreover, we discuss how our explanations can be used to create contrastive explanations (i.e., "why P rather than Q"). We do not present the underlying formal definitions here, these can be found in [4,5].

Our informal exploration has clear ties to recent more formal work on methods to derive explanations for specific conclusions [8,9,10,12,21]. That we introduce a new framework rather than use or modify an existing one has several reasons. Often, explanations are only defined for a specific semantics [8,9] and can usually only be applied to abstract argumentation [9,12,21],[3] 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 [8,12] or non-acceptance [9,21], while we introduce one framework with which both acceptance

[3] 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.

and non-acceptance explanations can be derived in a similar way.[4] 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

In order to present explanations for argumentation-based conclusions, first some basic concepts from formal argumentation have to be introduced. Reasoning based on formal argumentation is based on three concepts: arguments, an attack relation and a notion of defense:

- In abstract argumentation, as introduced in [6], arguments are abstract entities. However, in structured (or deductive) argumentation, arguments can be constructed from a knowledge base and a set of rules.
- Between these arguments an attack relation is defined. Again, in abstract argumentation, this relation is abstract and pre-defined by the user. But in structured argumentation, the attacks between arguments are determined by the underlying structure of the arguments.
- From the attack relation a notion of defense can be derived. An argument A can defend an argument B if it attacks an attacker of B.

An argumentation framework is then a pair of a set of arguments and an attack relation between those arguments. Formally, an *argumentation framework* 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 arguments $A, B \in \text{Args}$, it is said that A *attacks* B iff $(A, B) \in \mathcal{A}$ and A *defends* B if for some $C \in \text{Args}$, $(A, C), (C, B) \in \mathcal{A}$.[5] An argumentation framework can be viewed as a directed graph, in which the nodes represent arguments and the arrows represent attacks between arguments. See Figure 1 (on page 5) for an example.

Conclusions are drawn by selecting sets of arguments that can collectively be considered as acceptable. How such sets are selected depends on the choices of the designer of the system. Common requirements are that the set:

- is conflict-free: there are no attacks between the arguments in the set;

[4] An exception to this might be [10]. However, we consider our framework easier applicable, since it returns sets of arguments rather than sets of dialectical trees, which might contain many arguments.

[5] Our notion of defense, defined between arguments, is different from the one introduced by Dung [6], defined between a set of arguments and an argument. We use this definition since we are interested in the *arguments that defend* a certain argument, rather than whether that argument is *defended by the set of arguments*.

- defends itself: for any attacker of an argument in the set, there is an argument in the set that defends against this attacker;
- is complete: if an argument is defended against all its attackers by the set, then it is contained in the set.

There are different ways in which the conclusions can be drawn from the selected sets of arguments. In the application at the police it is important to only draw conclusions that one can be certain about. 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 [6] is used). For the purpose of this paper, to illustrate the variety of possible explanations, we take a more credulous approach: an argument that is part of some complete set can be considered a conclusion (i.e., the preferred semantics from [6] is used).

Remark 1. For the interested reader, a note on the actual system used in the applications. Each of the applications that is in use, is based on a variation of ASPIC$^+$, one of the best-known approaches to structured argumentation [19]. In particular, the notions of a language, axioms and defeasible rules are taken from ASPIC$^+$ and the conclusions are drawn based on the grounded semantics. See [18] for the formal details.[6]

These basic notions from 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 Internet Trade Fraud

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.[7] 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*;

[6] The corresponding demo of [18], demonstrating the argumentation-based part of the application, is available at https://nationaal-politielab.sites.uu.nl/estimating-stability-for-efficient-argument-based-inquiry/.

[7] 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.

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:

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*.

Attacks between those arguments can be derived from the conflicts between conclusions of the (sub)arguments. For example 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 on their conclusion and A_5 attacks A_4 and A_6 because they have *fake* as a sub-conclusion). The graphical representation of the argumentation framework, which we will refer as $\mathcal{AF}_1$ can be found in Figure 1.

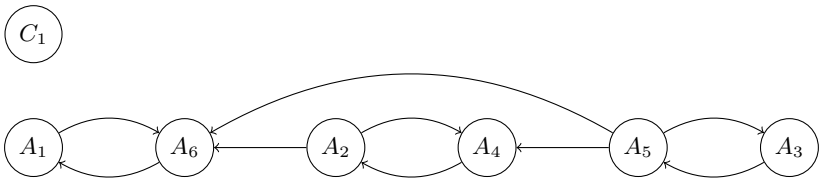

Fig. 1. Graphical representation of the argumentation framework $\mathcal{AF}_1$ constructed based on information provided in the complaint.

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*). From an argumentative perspective a credulous reasoner (a reasoner who wants to accept as many conclusions as possible) can accept both arguments, though not simultaneously. For A_1 this is the case since A_1 attacks any argument by which it is attacked (i.e., there is an attack from A_1 to A_6). 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 fake* (A_3). This follows since A_3, A_4 and A_6 together attack all the arguments that attack them: A_3 attacks A_5 and therefore defends itself, A_4 and A_6 from the attacks by A_5; A_4 attacks A_2 and therefore

defends itself and A_6 from the attacks by A_2; and A_6 attacks A_1 and therefore defends itself against that attack. In what follows we will consider for both A_1 and A_6 explanations for why one could (not) accept them.

It is a case of fraud (acceptance of A_6/non-acceptance of A_1). The 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 individual elements of arguments, 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). Explanations like this take the last rule that was applied in the construction of the argument as well as the antecedents of that rule. Such an explanation can be used by an analyst at the police, who is familiar with the rules and might want to understand what parts of the law were applied to derive the conclusion.
- the *complainant paid* and the product seems *suspicious*. This type of explanation looks at the information provided by the complainant (i.e., the knowledge base) and shows which of this information was used in the derivation of the conclusion. 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). In this explanation all the sub-conclusions are returned that were derived from the information provided by the complainant. 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. This is again the explanation in terms of the last rule applied to derive the argument for *not a case of fraud* and its antecedents. 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. This explanation is about the information provided by the complainant: the elements from the knowledge base. 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). Here the sub-conclusions that were found during the derivation of the argument form the explanation. 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 [16] 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. To illustrate some of the findings and how these can be implemented into our system of explanations, we consider three cases. The first two are often studied together: *necessity* and *sufficiency*.[8] In the context of philosophy and cognitive science, these are discussed in, for example [13,14,22]. The third are *contrastive explanations* [13,15,16]: 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 [13]. The answer to the question is then to explain as many of the differences between fact and foil as possible. We discuss each of the cases separately, in the context of our argumentation setting.

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:

[8] See [4] for the technical details of the necessary and sufficient explanations for abstract argumentation.

- 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. There is therefore only one set of arguments (minimally) sufficient: $\{A_3, A_4, A_6\}$.

Based on these sufficient sets of arguments 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. In the case of our example this does not matter. However, it might be that one sufficient set contains one argument constructed from many premises, while another sufficient set contains several arguments which are constructed from less premises.

In our example we have that:

- receiving the *wrong product* is sufficient for that it is *not a case of fraud*, if we are interested in the premises and, combined with the rule that usually when the wrong product is received it is not a case of fraud, when we are interested in the last rule applied in the construction of the argument.
- the premises that the *complainant paid* and that the product seems *suspicious* are sufficient for that it is *a case of fraud*. When looking at the last rules applied, 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 the argumentation framework on internet trade fraud, there is not much difference between the basic explanations and sufficient explanations. Therefore, we introduce here a second argumentation framework, this time abstract (i.e., no underlying structure in the arguments and not based on a scenario from the police), with which we can show how sufficient explanations might differ.

Example 1. Let $\mathcal{AF}_2 = \langle \text{Args}_2, \mathcal{A}_2 \rangle$ be an argumentation framework where $\text{Args}_2 = \{A, B, C, D, E, F, G\}$ are abstract arguments and where we define the attacks as follows: $\mathcal{A}_2 = \{(B, A), (C, B), (C, D), (D, C), (E, B), (F, E), (F, G), (G, F)\}$. See Figure 2 for a graphical representation.

As in the case of our running example, each argument can be accepted by a credulous reasoner and no argument is accepted by a skeptical reasoner. In order to accept A either C or E should be accepted as well and in order to accept E

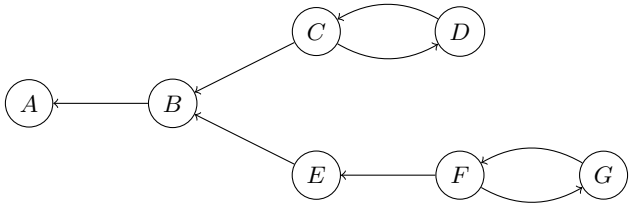

Fig. 2. Graphical representation of the abstract argumentation framework $\mathcal{AF}_2$.

one should accept G. To accept B, one has to accept both D and F. Similarly, in order to not accept A, one has to accept B and therefore both D and F as well, while not accepting B means accepting at least C or E (and possibly both).

Sufficient explanations for the acceptance of A are $\{C\}$, $\{E, G\}$, $\{C, E, G\}$, but also $\{C, F\}$ and $\{D, E, G\}$ (since these still include C resp. E and G). Minimally sufficient explanations are $\{C\}$ and $\{E, G\}$ when minimality is taken w.r.t. $\subseteq$ and only $\{C\}$ when minimality is taken w.r.t. the size of the set. There is only one (minimally) sufficient explanation for the acceptance of B: $\{D, F\}$.

Like in the case of the internet trade fraud example (recall $\mathcal{AF}_1$), the non-acceptance explanations are very similar to the acceptance explanations: the only (minimal) sufficient explanation of the non-acceptance of A is $\{B, D, F\}$ while the sufficient non-acceptance explanation for B can be $\{C\}$, $\{E, G\}$, $\{C, E, G\}$, $\{C, F\}$ and $\{D, E, G\}$.

Necessity. In terms of arguments, an argument can be understood as *necessary* if without that argument, the considered argument could not be accepted. In the case of our example, 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.

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.

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. But there is more possible. Since sometimes there might not be a necessary argument (i.e., when the argument for which the explanation is required is not attacked at all, or the intersection of its sufficient sets is empty)[9] one could still collect necessary premises, rules and

[9] It can be shown formally that there are no sufficient sets of arguments when the argument is not attacked at all and that there are no necessary arguments if the intersection of the sufficient sets is empty (which is at least the case when there are no sufficient sets of arguments).

sub-conclusions (e.g., by taking the intersection of the elements of the sufficient arguments).

For necessity we can also take a look at the abstract argumentation framework $\mathcal{AF}_2$ introduced to illustrate sufficiency. As in the case for our running example on internet trade fraud with the argumentation framework $\mathcal{AF}_1$, when the intersection of the sufficient sets is empty, there are no necessary arguments. For the argumentation framework $\mathcal{AF}_2$ we have that for the acceptance of A no argument is necessary, while for the acceptance of B both D and F are necessary. Similarly, for the non-acceptance of A the arguments B, D and F are necessary, while no argument is necessary for the non-acceptance of B.

Contrastive Explanations. When humans provide a contrastive explanation (they answer the question 'why P rather than Q?' when asked 'why P?'), the foil (i.e., Q) 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 artificial intelligence 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.

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.

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).[10] For example:

- it is a case of *fraud* and not of *not fraud* because of the arguments A_3, A_4 and A_6.
- it is a case of *not fraud* and not a case of *fraud* because of the argument A_1 (and possibly A_2 and A_5).

The explanations in terms of the elements of the arguments have been discussed previously, we will therefore not repeat that discussion here.

[10] It can be shown formally that the intersection is empty when the accepted argument is not attacked or fact and foil are not relevant for each other, i.e., neither does attack the other.

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 [6].

Recall from the introduction that, unlike other approaches to local explanations of argumentation-based conclusions [8,9,10,12,21], our explanation framework can capture both acceptance and non-acceptance explanations, is not based on one specific semantics 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., [16]) 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. Argument & Computation **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.: Necessary and sufficient explanations for formal argumentation. arXiv/CoRR **abs/2011.02414** (2020), https://arxiv.org/abs/2011.02414
5. Borg, A., Bex, F.: A basic framework for explanations in argumentation. IEEE Intelligent Systems (2021), doi: 10.1109/MIS.2021.3053102
6. 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)

7. 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)
8. 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)
9. 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)
10. 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)
11. Lacave, C., Diez, F.J.: A review of explanation methods for heuristic expert systems. The Knowledge Engineering Review **19**(2), 133–146 (2004)
12. 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'20). Frontiers in Artificial Intelligence and Applications, vol. 326, pp. 271–282. IOS Press (2020)
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.: Contrastive explanation: A structural-model approach. CoRR **abs/1811.03163** (2018), http://arxiv.org/abs/1811.03163
16. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence **267**, 1 – 38 (2019)
17. 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 the 33rd International Conference on Legal Knowledge and Information Systems (JURIX'20). Frontiers in Artificial Intelligence and Applications, vol. 334, pp. 239–242. IOS Press (2020)
18. 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)
19. Prakken, H.: An abstract framework for argumentation with structured arguments. Argument & Computation **1**(2), 93–124 (2010)
20. Samek, W., Wiegand, T., Müller, K.R.: Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. ITU Journal: ICT Discoveries - Special Issue 1 - The Impact of Artificial Intelligence (AI) on Communication Networks and Services **1**(1), 39–48 (2018)
21. 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)
22. Woodward, J.: Sensitive and insensitive causation. Philosophical Review **115**(1), 1–50 (2006)