Modeling Contrastiveness in Argumentation

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

Modeling contrastive explanations for the use in artificial intelligence (AI) applications is an important research branch within the field of explainable AI (XAI). However, most of the existing contrastive XAI approaches are not based on the findings in the literature from the social sciences on contrastiveness in human reasoning and human explanations. In this work we collect the various types of contrastiveness proposed in the literature and model these with formal argumentation. The result is a variety of argumentation-based methods for contrastive explanations, based on the available literature and applicable in a wide variety of AI-applications.

1 Introduction

Explainable Artificial Intelligence (XAI) is an important and fast growing research area, contributing to closing the gap between AI-application and its human user. To this end it is essential that XAI approaches incorporate findings from the humanities and social sciences on how humans request, generate, interpret and evaluate explanations (e.g., research on explanations from philosophy [19] or law [4], see [21] for an extensive survey).

Argumentation plays an important role in XAI, because it is central to all human reasoning [20] including explanation [2, 9]. In addition to existing argumentation-based systems being inherently interpretable (cf. [24]), argumentation has also been used to explain the output of other less interpretable AI models such as machine learning classification models (e.g., [23]), deep reinforcement learning models [18] and probabilistic Bayesian Networks [29] (see [13] for a recent overview). Such explanations in terms of arguments can come in many forms, taking into account the above-mentioned literature from the humanities and social sciences. For example, argument-based explanations can be minimal [15], in the form of a dialogue [3], or include only necessary and sufficient reasons [11].

One aspect of everyday explanations that has so far received relatively little attention in the literature on argumentative explanations is that of *contrastiveness*: when asking *why P?* we often expect the explanation to answer *why P rather than Q?* Contrastive explanations can be used to compare a surprising outcome with the expected outcome [19, 31, 17]. Additionally, by using contrastiveness, explanations can be interactively customized and personalized [27], and become easier to derive and cognitively less demanding [19, 32].

There are a lot of studies on contrastive explanations in XAI, but very few of these base their explanation definitions on the theoretical foundations from the humanities and social sciences literature on contrastiveness in human reasoning and human explanation [28]. In this work we study the concept of contrastiveness as can be found in the literature on explanations and formalize our findings in terms of formal argumentation. After a brief introduction to formal argumentation, we discuss the various types of contrastiveness as proposed in the literature as well as the requirements placed on them [19, 28, 17, 32, 5, 26]. We then show how these types of contrastiveness

can be modeled in abstract argumentation [14] (Section 3.2). Due to space restrictions we only study contrastiveness for the (non-)acceptance of arguments in abstract argumentation. However, the proposed explanations could be applied to a structured setting (e.g., one of those in [8]) in a similar way as was done in [10].

With this paper we start a discussion on contrastiveness for argumentative XAI. We provide the basis for this discussion by giving a literature overview and showing how notions from the literature can be implemented in abstract argumentation.

2 Preliminaries: argumentation and explanation

2.1 Abstract Argumentation

An *abstract argumentation framework* (AF) [14] is a pair $\mathcal{AF} = \langle \mathsf{Args}, \mathsf{Att} \rangle$, where Args is a set of *arguments* and $\mathsf{Att} \subseteq \mathsf{Args} \times \mathsf{Args}$ is an *attack relation* on these arguments. An AF can be viewed as a directed graph, see Figure 1 for an example, in which the nodes represent arguments and the arrow represent attacks between arguments.
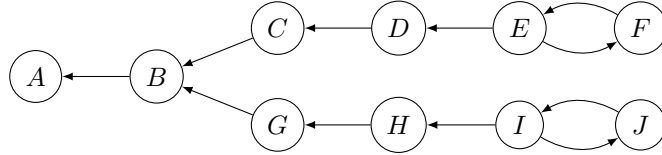

Figure 1: Graphical representation of the argumentation framework $\mathcal{AF}_1$.

Given an argumentation framework $\mathcal{AF}$, Dung-style semantics [14] can be applied to it, to determine what combinations of arguments (called *extensions*) can collectively be accepted.[1]

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

An admissible extension that contains all the arguments that it defends is a complete extension *(*Cmp*). The* grounded extension *(*Grd*) is the minimal (w.r.t.* $\subseteq$*) complete extension and a* preferred extension *(*Prf*) is a maximal (w.r.t.* $\subseteq$*) complete extension.*

We denote by $\mathsf{Sem}(\mathcal{AF})$ the set of all extensions of $\mathcal{AF}$ under the semantics $\mathsf{Sem} \in \{\mathsf{Grd}, \mathsf{Cmp}, \mathsf{Prf}\}$. Based on the semantics, an argument can be accepted and/or not accepted.

Definition 2. *Let* $\mathcal{AF} = \langle \mathit{Args}, \mathit{Att} \rangle$ *be an AF,* $\mathsf{Sem} \in \{\mathsf{Grd}, \mathsf{Cmp}, \mathsf{Prf}\}$ *a semantics and let* $A \in \mathit{Args}$ *be an argument. Then:* A *is* accepted *if there is some* $\mathcal{E} \in \mathsf{Sem}(\mathcal{AF})$ *such that* $A \in \mathcal{E}$; *and* A *is* not accepted *(or* non-accepted*) if there is some* $\mathcal{E} \in \mathsf{Sem}(\mathcal{AF})$ *such that* $A \notin \mathcal{E}$.

Note that, since the grounded extension is unique, if $\mathsf{Sem} = \mathsf{Grd}$, an argument is either accepted or not accepted. For the other semantics an argument can be accepted, not accepted or both (i.e., there might be extensions with the argument and extensions without the argument).

Example 1. *In* $\mathcal{AF}_1$ *from Figure 1 the grounded extension is empty:* $\mathsf{Grd}(\mathcal{AF}_1) = \{\emptyset\}$ *and there are four preferred extensions* $\mathsf{Prf}(\mathcal{AF}_1) = \{\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3, \mathcal{E}_4\}$ *where* $\mathcal{E}_1 = \{A, C, E, G, I\}$, $\mathcal{E}_2 = \{A, C, E, H, J\}$, $\mathcal{E}_3 = \{A, D, F, G, I\}$ *and* $\mathcal{E}_4 = \{B, D, F, H, J\}$. *It follows that all arguments of* $\mathcal{AF}_1$ *are both accepted and non-accepted for* $\mathsf{Sem} \in \{\mathsf{Cmp}, \mathsf{Prf}\}$.

---

[1]Other semantics can be found in, e.g., [6]. We do not discuss these, since the specific semantics is not relevant for our discussion.

2.2 Explanations in Argumentation

We now recall the main definitions for the explanation framework from [10], in which basic explanations for both the acceptance and non-acceptance of some argument A under various semantics were defined. We will apply a general notation for acceptance and non-acceptance explanations, without any restrictions:

Notation 1. *Let $\mathcal{AF} = \langle Args, Att \rangle$ be an AF, $A \in Args$, $\mathsf{Sem} \in \{\mathsf{Grd}, \mathsf{Cmp}, \mathsf{Prf}\}$ be a semantics and $\mathcal{E} \in \mathsf{Sem}(\mathcal{AF})$ an extension. Then:*

- *$\mathsf{SemAcc}(A, \mathcal{E})$ denotes an acceptance explanation for A in the specific extension $\mathcal{E}$; and*

- *$\mathsf{SemNotAcc}(A, \mathcal{E})$ denotes a non-acceptance explanation for A in the specific extension $\mathcal{E}$.*

By using this general notation, other argumentative explanation methods can be used as well, e.g., [15, 16, 25, 30]. Note that these explanations differ slightly from those introduced in [10], in that they also require an extension $\mathcal{E}$ in addition to an argument A – the reason for this will become apparent in Section 3.2. Because explanations in abstract argumentation are defined in terms of attack and defense between arguments, we also need the following definition.

Definition 3. *Given $\mathcal{AF} = \langle Args, Att \rangle$, we say that $A \in Args$ directly defends B if there is some C such that $(C, B) \in Att$ and $(A, C) \in Att$, and A indirectly defends B if A defends C and C defends B. Similarly, A directly attacks $B \in Args$ if $(A, B) \in Att$, A indirectly attacks B if A attacks some $C \in Args$ and C defends B. We will often say that A defends [resp. attacks] B when A directly or indirectly defends [resp. attacks] B.*

Now we can define basic explanations for (non-)accepted arguments as follows.

Definition 4. *Let $\mathcal{AF} = \langle Args, Att \rangle$ be an AF and let $\mathsf{Sem} \in \{\mathsf{Grd}, \mathsf{Cmp}, \mathsf{Prf}\}$. First, let $A \in Args$ be accepted and let $\mathcal{E} \in \mathsf{Sem}(\mathcal{AF})$ such that $A \in \mathcal{E}$, then:*

$$\mathsf{SemAcc}(A, \mathcal{E}) = \mathbb{D}^{\mathsf{acc}}(A, \mathcal{E}).$$

Now, let $B \in Args$ be non-accepted and let $\mathcal{E} \in \mathsf{Sem}(\mathcal{AF})$ such that $B \notin \mathcal{E}$, then:

$$\mathsf{SemNotAcc}(B, \mathcal{E}) = \mathbb{D}^{\mathsf{nacc}}(B, \mathcal{E}).$$

The explanations use a generic function $\mathbb{D}$ that returns a set of arguments as an explanation given a (non-)accepted argument and an extension. This function allows us to vary which attacking or defending arguments are in an explanation for an acceptance explanation ($\mathbb{D}^{\mathsf{acc}}$) or a non-acceptance explanation ($\mathbb{D}^{\mathsf{nacc}}$). The two common instantiations of $\mathbb{D}^{\mathsf{acc}}$ and $\mathbb{D}^{\mathsf{nacc}}$ that we will use throughout this paper are as follows.[2]

- $\mathsf{Defending}(A, \mathcal{E}) = \{B \in \mathcal{E} \mid B \text{ (in)directly defends } A\}$;

- $\mathsf{NoDefAgainst}(A, \mathcal{E}) = \{B \in \mathcal{E} \mid B \text{ attacks } A \text{ and } \mathcal{E} \text{ does not attack } B\}$.

When $\mathbb{D}^{\mathsf{acc}} = \mathsf{Defending}$, an acceptance explanation for argument A in $\mathcal{E}$ is the set of arguments from $\mathcal{E}$ that defends A against all its attackers, and for $\mathbb{D}^{\mathsf{nacc}} = \mathsf{NoDefAgainst}$, a non-acceptance explanation for argument $A \notin \mathcal{E}$ is the set of attackers of A against which $\mathcal{E}$ provides no defence.

Example 2. *Recall that for $\mathcal{AF}_1$ we have the preferred extensions $\mathcal{E}_1$, $\mathcal{E}_2$, $\mathcal{E}_3$ and $\mathcal{E}_4$, then:*

- *$\mathsf{PrfAcc}(A, \mathcal{E}_1) = \{C, E, G, I\}$, $\mathsf{PrfAcc}(A, \mathcal{E}_2) = \{C, E\}$, $\mathsf{PrfAcc}(A, \mathcal{E}_3) = \{G, I\}$ and for argument B $\mathsf{PrfAcc}(B, \mathcal{E}_4) = \{D, F, H, J\}$;*

- *$\mathsf{PrfNotAcc}(A, \mathcal{E}_4) = \{B, D, F, H, J\}$ and $\mathsf{PrfNotAcc}(B, \mathcal{E}_1) = \{C, E, G, I\}$, while for argument B: $\mathsf{PrfNotAcc}(B, \mathcal{E}_2) = \{C, E\}$, $\mathsf{PrfNotAcc}(B, \mathcal{E}_3) = \{G, I\}$.*

---

[2] Other instantiations of $\mathbb{D}$ are, e.g., $\mathsf{DirDefending}$ and $\mathsf{DirAttacking}$, which take only the direct defenders or attackers of A, see [11, 10].

3   Contrastiveness in argumentation

The explanations in Section 2.2 are simple, non-contrastive explanations, answers to a *Why (not) P?* explanation-seeking question, where P is understood as the (non-)acceptance of an argument. In its most basic form a contrastive explanation-seeking question is then of the form *Why P rather than Q?*, where P is called the *fact* and Q is called the *foil* [19]. However, there is no agreement on exactly what is or constitutes a contrastive explanation [28]. Like Rudin [24], we assume that more than one definition is possible, and below we will discuss the different ideas on contrastiveness and how they can be interpreted for argumentative explanations.

The specific contrastive question may vary. First, there is the *negation* question, where the fact is compared with the situation in which it does not occur: *Why P rather than not-P?* [17]. Second, it is possible to compare the *properties* of a certain object, resulting in questions like: *Why P rather than the default value for P?* [17] or *Why does object a have property P, rather than property Q?* [31]. Finally, different *objects* can be compared as well: *Why P rather than Q?* [17] and *Why does object a have property P, while object b has property Q?* [31].

  I. Considering explanations in abstract argumentation, the fact P concerns the (non-)accep-
     tance of an argument, which is also the only property of the argument (i.e., whether it is
     accepted or not). Thus, both negation-type and property-type questions are captured by
     comparing why an argument is accepted as opposed to not accepted. Object-type questions
     then compare different arguments (objects) to each other.

The idea of contrastiveness usually assumes that fact and foil are incompatible [32, 5]. Even for a seemingly compatible fact and foil, what we are really interested in is some sort of (hypothetical) contrast. Take, for example, the question 'Why does John have disease X and Harry does not?' In this case, it would be perfectly possible for Harry to also have disease X, so fact and foil are not inherently incompatible. However, as Ylikoski [32] notes, we are not really interested in Harry (why he does not have X), but rather treat him as a surrogate for John, for whom we want to know why he has X while someone similar to him (Harry) does not.

 II. We will assume that fact and foil are not always compatible. This means that fact and foil
     are not always part of the same extensions, i.e., there is at least one extension in which fact
     or foil is accepted and the other is not. This can be further restricted by, e.g., requiring that
     fact and foil are never part of the same extension (i.e., always incompatible).

Note that it is assumed that fact and foil should be similar or closely related. For example, they should have common presuppositions or a common cause, or share similar causal histories [19, 32, 5]. For example, it makes little sense to ask 'Why does John have disease X while my cat does not?'. We assume that John and Harry are related in that they are comparable persons (age, gender, background, etc.) who share a similar history, i.e., the same (type of) events that have caused John to contract disease X could have also affected Harry.

III. We will assume that fact and foil are *relevant* for each other, which means that they are at
     least connected via the attack relation.

Sokol and Flach [27] differentiate between *Why P despite Q?* and a *Why P given Q?*. The despite-question requires an explanation that is not conditioned on Q, that is, the explanation for the fact P does not take into account the explanation for the foil Q. The given-question requests an explanation in which the explanation for fact P is conditioned on the explanation for foil Q: what do the explanations for fact and foil have in common? Take, for example, the question 'Why did John get a loan and Harry did not?'. A despite-explanation would then be 'because John has an income of more than 100k' (John's income is not directly relevant for Harry), and a given-explanation would be 'because John was first and only one person can get a loan' (John being first is directly relevant for Harry).

IV. For a given-question, the explanation of the fact is conditioned on the foil (the explanation contains the part of the explanations that are the same for both fact and foil). For a despite-question, the explanation of the fact is not conditioned on the foil (the explanation contains the part of the explanation for the fact without the foil).

The final important difference made in the literature is the one between alternative and actual foils, which is made explicit in [32]. On the alternative side there are two incompatible and alternative outcomes of the same process, one the fact and one the foil. On the actual side there are two actual (possibly compatible) outcomes, fact and foil, from two actual and different processes. From these two types of contrasts, two questions are formulated and modeled in [22]. The *alternative* question asks *Why P rather than Q?*, where fact P occurred and foil Q did not, while the *actual* question asks *Why P but Q?*, where P occurred in the current situation and Q occurred in some other situation.

V. We will consider both types of foils. To this end we will take a fixed argumentation framework (i.e., for now we do not allow for dynamic settings) and take extensions within that framework to represent different situations. The alternative question is then modeled given a specific extension (i.e., within an extension), while the actual question is modeled between extensions.

3.1 Relevance and Compatibility

In view of considerations II. and III. above we introduce the notions of compatibility and relevance. Compatibility concerns the mutual acceptance of arguments in the same extensions.

Definition 5. *Let $\mathcal{AF} = \langle Args, Att \rangle$ be an AF and $A, B \in Args$ be two arguments. Then, for* Sem $\in \{$Cmp, Grd, Prf$\}$: *A and B are* compatible *w.r.t.* Sem *if for some $\mathcal{E} \in$ Sem$(\mathcal{AF})$ it holds that $A, B \in \mathcal{E}$; A and B are* incompatible *w.r.t.* Sem *if there is no $\mathcal{E} \in$ Sem$(\mathcal{AF})$ such that $A, B \in \mathcal{E}$; and A and B are* not always compatible *w.r.t.* Sem *if there is some $\mathcal{E} \in$ Sem$(\mathcal{AF})$ such that $A \in \mathcal{E}$ and $B \notin \mathcal{E}$.*

The above can be generalized to sets of arguments. Given two sets of arguments $S_1, S_2 \subseteq Args$: S_1 and S_2 are compatible *w.r.t.* Sem *if for each pair $(A, B) \in S_1 \times S_2$, there is some $\mathcal{E} \in$ Sem$(\mathcal{AF})$ such that $A, B \in \mathcal{E}$; S_1 and S_2 are* incompatible *w.r.t.* Sem *if there is no $\mathcal{E} \in$ Sem$(\mathcal{AF})$ such that $S_1 \cap \mathcal{E} \neq \emptyset$ and $S_2 \cap \mathcal{E} \neq \emptyset$; S_1 and S_2 are* not always compatible *w.r.t.* Sem *if there is some $\mathcal{E} \in$ Sem$(\mathcal{AF})$ such that $S_1 \cap \mathcal{E} \neq \emptyset$ and $S_2 \cap \mathcal{E} = \emptyset$.*

Arguments or sets of arguments are compatible if these can be part of the same extension, incompatible if there is no extension that contains both arguments or arguments from both sets and not always compatible if there is an extension that only contains one argument or arguments from one set and not the other. Given II. we will assume that fact and foil are not always compatible. Given III. we will assume that fact and foil are relevant for each other:

Definition 6. *Let $\mathcal{AF} = \langle Args, Att \rangle$ be an AF and $A, B \in Args$ be two arguments. Then A is* relevant *for B if: A (in)directly attacks or defends B; or there is some $C \in Args$ such that C is relevant for both A and B. We say that A is* conflict-relevant *for B if it (in)directly attacks B and A is* defending-relevant *for B if it (in)directly defends B.*

Arguments are relevant for each other when they are connected by the attack relation, or if there is an argument that is connected by the attack relation to both of the considered arguments.

Example 3. *For the argumentation framework $\mathcal{AF}_1$ we have that E and H are relevant for B and A but not for each other. Also, for* Sem $=$ Prf, *E and H are compatible (there is the extension $\{A, C, E, H, J\}$), but not always (there is the extension $\{A, C, E, G, I\}$) and although H is compatible with B (there is the extension $\{B, D, F, H, J\}$), E is not compatible with B.*

	Neg + despite	Obj + despite	Neg + given	Obj + given
Within	Note 1	Def. 7 & Ex. 4	Note 1	Def. 8 & Ex. 5
Between	Def. 9 & Ex. 7	Def. 9 & Ex. 7	Def. 10 & Ex. 8	Def. 10 & Ex. 8

Table 1: Overview of the types of contrastiveness for considerations I. to V.

3.2 Contrastive Explanations

We can now formalize the notions of contrastiveness discussed earlier. Recall that we consider (I.) negation-type and object-type questions; (II.), (III.) relevant, but not always compatible fact and foil; (IV.) given- and despite-questions; and (V.) alternative (within an extension) and actual (between extensions) questions. The resulting eight types of contrastive explanations are shown in Table 1.

Note 1. *An argument cannot be part of an extension and, at the same time, not be part of that extension. Given II. (not always compatible) and the fact that* within *explanations look at just one extension, a contrastive explanation for* within *will always be an object-type explanation (i.e., between two different arguments).*

We start with the contrastive explanation for *despite* and *within*: *Why A despite B in the current situation?* To ensure a form of contrast, since the fact is accepted, the foil is non-accepted if it is defending-relevant for the fact and the foil is accepted if it is conflict-relevant for the fact. We do not limit the foil to be a single argument, rather, we define the explanations between one argument (the fact) and a set of arguments (the foil).

Definition 7. *Given* $\mathcal{AF} = \langle Args, Att \rangle$, $A \in Args$, $\mathsf{S} \subseteq Args$ *and* $\mathsf{Sem} \in \{\mathsf{Grd}, \mathsf{Cmp}, \mathsf{Prf}\}$. *Let* $\mathsf{Expl} \in \{\mathsf{Acc}, \mathsf{NotAcc}\}$ *determine the type of explanation and let* $\mathcal{E} \in \mathsf{Sem}(\mathcal{AF})$, *then:*

$$\mathsf{SemCont}((A, \mathcal{E}), (\mathsf{S}, \mathcal{E})) = \mathsf{SemAcc}(A, \mathcal{E}) \setminus \left(\bigcup_{B \in \mathsf{S}} \mathsf{SemExpl}(B, \mathcal{E}) \right)$$

$$\mathsf{SemContN}((A, \mathcal{E}), (\mathsf{S}, \mathcal{E})) = \mathsf{SemNotAcc}(A, \mathcal{E}) \setminus \left(\bigcup_{B \in \mathsf{S}} \mathsf{SemExpl}(B, \mathcal{E}) \right).$$

The explanation contains the reasons for the acceptance of the fact that are not part of the acceptance (if $\mathsf{Expl} = \mathsf{Acc}$) or non-acceptance (if $\mathsf{Expl} = \mathsf{NotAcc}$) explanation of the foil and similarly for a non-accepted fact.

Example 4. *In the running example with* $\mathcal{AF}_1$, *there are three preferred extensions in which* A *is accepted and one in which it is not accepted. Now, when* A *is accepted it has to be defended against* B, *which* C *can do. However,* A *might still be accepted without* C *(i.e., in* $\mathcal{E}_3$*).*

- Why A despite not-C? *Recall that* $\mathsf{PrfAcc}(A, \mathcal{E}_3) = \{G, I\}$ *and that* $\mathsf{PrfNotAcc}(C, \mathcal{E}_3) = \{D, F\}$. *We therefore have, for* $\mathsf{Expl} = \mathsf{NotAcc}$ *(i.e., comparing the acceptance of the fact with the non-acceptance of the foil):* $\mathsf{PrfCont}((A, \mathcal{E}_3), (\{C\}, \mathcal{E}_3)) = \{G, I\}$.

- Why not-B despite not-C? *This question is very similar to the above one, where we will now apply* $\mathsf{PrfNotAcc}(B, \mathcal{E}_3) = \{G, I\}$. *We then have, again for* $\mathsf{Expl} = \mathsf{NotAcc}$ *now comparing two non-accepted arguments:* $\mathsf{PrfContN}((B, \mathcal{E}_3), (\{C\}, \mathcal{E}_3)) = \{G, I\}$.

- Why not-B despite not-G? *The explanation for the non-acceptance of* B: $\mathsf{PrfNotAcc}(B, \mathcal{E}_2) = \{C, E\}$. *Similarly to the item above, with* $\mathsf{Expl} = \mathsf{NotAcc}$: $\mathsf{PrfContN}((B, \mathcal{E}_2), (\{G\}, \mathcal{E}_2)) = \{C, E\}$.

For the contrastive explanation *given* and *within* we answer *Why A given its alternative B in the current situation?* Similar to the case of *despite* and *within* contrasts, independent of whether the fact is accepted or non-accepted, the foil can be accepted or non-accepted. However,

6

for accepted [resp. non-accepted] fact, if the foil is conflict-relevant for the fact it should be non-accepted [resp. accepted] and if it is defending-relevant for the fact it should be accepted [resp. non-accepted], this to ensure a contrast between fact and foil.

Definition 8. *Given* $\mathcal{AF} = \langle \mathsf{Args}, \mathsf{Att} \rangle$, $A \in \mathsf{Args}$, $\mathsf{S} \subseteq \mathsf{Args}$ *and* $\mathsf{Sem} \in \{\mathsf{Grd}, \mathsf{Cmp}, \mathsf{Prf}\}$. *Let* $\mathsf{Expl} \in \{\mathsf{Acc}, \mathsf{NotAcc}\}$ *determine the type of explanation and let* $\mathcal{E} \in \mathsf{Sem}(\mathcal{AF})$, *then:*

$$\mathsf{SemCont}((A, \mathcal{E}), (\mathsf{S}, \mathcal{E})) = \mathsf{SemAcc}(A, \mathcal{E}) \cap \left(\bigcup_{B \in \mathsf{S}} \mathsf{SemExpl}(B, \mathcal{E}) \right)$$

$$\mathsf{SemContN}((A, \mathcal{E}), (\mathsf{S}, \mathcal{E})) = \mathsf{SemNotAcc}(A, \mathcal{E}) \cap \left(\bigcup_{B \in \mathsf{S}} \mathsf{SemExpl}(B, \mathcal{E}) \right).$$

While in Definition 7 the explanation contained the reasons for the fact without the reasons for the foil, now the intersection of the explanations for fact and foil are taken. This follows since the question assumes that the foil is *given* and should be part of the explanation.

Example 5. *For the running example with* $\mathcal{AF}_1$:

- Why A given not-B? *Since there are several non-acceptance explanations for B, we have that* $\mathsf{PrfCont}((A, \mathcal{E}_2), (\{B\}, \mathcal{E}_2)) = \{C, E\}$ *and* $\mathsf{PrfCont}((A, \mathcal{E}_3), (\{B\}, \mathcal{E}_3)) = \{G, I\}$.

- Why B given not-G? *Since there is only one extension with B:* $\mathsf{PrfCont}((B, \mathcal{E}_4), (\{G\}, \mathcal{E}_4)) = \{H, J\}$.

For the fact-foil pairs from Example 4, where the foil is now accepted:

- Why A given C? *or* Why not-B given C? *From the acceptance explanation for C (i.e.,* $\mathsf{PrfAcc}(C, \mathcal{E}_2) = \{E\}$), *for* $\mathsf{Expl} = \mathsf{Acc}$: $\mathsf{PrfCont}((A, \mathcal{E}_2), (\{C\}, \mathcal{E}_2)) = \{E\}$, *while in the non-acceptance case:* $\mathsf{PrfContN}((B, \mathcal{E}_2), (\{C\}, \mathcal{E}_2)) = \{E\}$.

- Why not-B given G? *The explanation is very similar as the explanations in the above item, now based on* $\mathsf{PrfAcc}(G, \mathcal{E}_3) = \{I\}$. *Then, for* $\mathsf{Expl} = \mathsf{Acc}$: $\mathsf{PrfContN}((B, \mathcal{E}_3), (\{G\}, \mathcal{E}_3)) = \{I\}$.

We now turn to the comparison *between* extensions. These differ from the contrastive explanations within an extension, in that we do not take the difference or intersection with (a set of) explanations for a foil from the same extension, but rather with the second extension as containing the foil. In other words, we compare the actual situation (extension) of the fact with the actual situation (extension) of the foil.

Example 6. *Suppose we want to compare the acceptance of B in one extension with the acceptance of A in another. For example:* $\mathsf{PrfAcc}(B, \mathcal{E}_4) = \{D, F, H, J\}$ *and* $\mathsf{PrfAcc}(A, \mathcal{E}_2) = \{C, E\}$. *Taking A as the specific foil does not change the explanation, since the explanations do not intersect. If we instead look at $\mathcal{E}_2$ as a whole (recall $\mathcal{E}_2 = \{A, C, E, H, J\}$) we can take the difference and obtain the explanation $\{D, F\}$: B is accepted in $\mathcal{E}_4$ despite A being accepted in $\mathcal{E}_2$, since in $\mathcal{E}_4$ D and F are accepted as well.*

We start with the contrastive explanation for *despite*: *Why A despite B in another situation?* Fact and foil can both be (non-)accepted, since we compare different extensions.

Definition 9. *Given* $\mathcal{AF} = \langle \mathsf{Args}, \mathsf{Att} \rangle$, $A \in \mathsf{Args}$, $\mathsf{S} \subseteq \mathsf{Args}$ *and* $\mathsf{Sem} \in \{\mathsf{Cmp}, \mathsf{Prf}\}$. *Let* $\mathsf{Expl} \in \{\mathsf{Acc}, \mathsf{NotAcc}\}$ *determine the type of explanation and let* $\mathcal{E}, \mathcal{E}' \in \mathsf{Sem}(\mathcal{AF})$, *where* $\mathcal{E} \neq \mathcal{E}'$. *Moreover,* $\mathsf{SemExpl}(B, \mathcal{E}')$ *is an explanation according to Notation 1 for all $B \in \mathsf{S}$. Then:*

$$\mathsf{SemCont}((A, \mathcal{E}), (\mathsf{S}, \mathcal{E}')) = \mathsf{SemAcc}(A, \mathcal{E}) \setminus \mathcal{E}'$$
$$\mathsf{SemContN}((A, \mathcal{E}), (\mathsf{S}, \mathcal{E}')) = \mathsf{SemNotAcc}(A, \mathcal{E}) \setminus \mathcal{E}'.$$

These explanations present the difference of the explanation for the fact in one extension and the extension of the foil. It is assumed that the foil argument(s) are (non-)accepted in the extension $\mathcal{E}'$ and that proper explanations exist. This is necessary to represent the fact-foil construction of contrastive explanations.

Because fact and foil are now in different extensions, we can also ask negation-type questions, where the acceptance of an argument is compared with its non-acceptance.

Example 7. *For the running example with* $\mathcal{AF}_1$, *suppose that we are interested in the extensions* $\mathcal{E}_2$ *and* $\mathcal{E}_4$ *for* Sem = Prf. *We start by comparing the acceptance and non-acceptance of B for these settings:*

- *Why B in* $\mathcal{E}_4$ *despite not-B in* $\mathcal{E}_2$? *Then, for* Expl = NotAcc: PrfCont$((B, \mathcal{E}_4), (\{B\}, \mathcal{E}_2)) =$ $\{D, F\}$.

- *Why not-B in* $\mathcal{E}_2$ *despite B in* $\mathcal{E}_4$? *This results, for* Expl = NotAcc, *in:* PrfContN$((B, \mathcal{E}_2),$ $(\{B\}, \mathcal{E}_4)) = \{C, E\}$.

We can also compare the arguments from A, B, C and H:

- *Why A in* $\mathcal{E}_2$ *despite B in* $\mathcal{E}_4$? *For* Expl = Acc: PrfCont$((A, \mathcal{E}_2), (\{B\}, \mathcal{E}_4)) = \{C, E\}$.

- *Why B in* $\mathcal{E}_4$ *despite A in* $\mathcal{E}_2$? *The explanation is similar to the explanation above, for* Expl = Acc: PrfCont$((B, \mathcal{E}_4), (\{A\}, \mathcal{E}_2)) = \{D, F\}$.

- *Why A in* $\mathcal{E}_2$ *despite not-C in* $\mathcal{E}_3$? *For* Expl = NotAcc: PrfCont$((A, \mathcal{E}_2), (\{C\}, \mathcal{E}_3)) = \{C, E\}$.

- *Why not-A in* $\mathcal{E}_4$ *despite H in* $\mathcal{E}_2$? *For* Expl = Acc: PrfContN$((A, \mathcal{E}_4), (\{H\}, \mathcal{E}_2)) = \{B, D, F\}$.

Note 2. *With these explanation it becomes apparent that the choice of the right foil is essential when formulating the explanation. While the above explanations make sense when an argument is compared with its negation (see the first part of Example 7), this might be less so when two arguments are compared. This has to be accounted for when choosing the formalization, extensions, fact and foil. For example (last two bullets in Example 7), why A in* $\mathcal{E}_2$ *despite not-C in* $\mathcal{E}_3$ *might not be an informative explanation, while why not-A in* $\mathcal{E}_4$ *despite H in* $\mathcal{E}_2$ *might be informative, since H is a reason for the non-acceptance of A in* $\mathcal{E}_4$ *but they are both accepted in* $\mathcal{E}_2$. *What conditions are placed on the foil depends on the applications and situation at hand and is left for future work.*

Finally, we consider the contrastive explanation for *given* and *between*: *Why A given B in another situation?*. Again, rather than taking the intersection with a set of explanations, we take the intersection with the second extension, to ensure a comparison between the fact and its extension with the situation of the foil.

Definition 10. *Given* $\mathcal{AF} = \langle Args, Att \rangle$, $A \in Args$, $S \subseteq Args$ *and* Sem $\in$ {Cmp, Prf}. *Let* Expl $\in$ {Acc, NotAcc} *determine the type of explanation and let* $\mathcal{E}, \mathcal{E}' \in$ Sem$(\mathcal{AF})$, *where* $\mathcal{E} \neq \mathcal{E}'$. *Moreover,* SemExpl$(B, \mathcal{E}')$ *is an explanation according to Notation 1 for all* $B \in S$. *Then:*

$$\text{SemCont}((A, \mathcal{E}), (S, \mathcal{E}')) = \text{SemAcc}(A, \mathcal{E}) \cap \mathcal{E}'$$
$$\text{SemContN}((A, \mathcal{E}), (S, \mathcal{E}')) = \text{SemNotAcc}(A, \mathcal{E}) \cap \mathcal{E}'.$$

In words, these explanations contain the common reasons for the fact in one extension and the arguments in another extension. The other extension is such that it provides proper explanations for the argument(s) in the foil.

Example 8. *We are again interested in comparing* $\mathcal{E}_2$ *and* $\mathcal{E}_4$. *However, rather than taking the difference between the two, we will now take the intersection.*

- Why B in $\mathcal{E}_4$ given not-B in $\mathcal{E}_2$? *The acceptance of B in $\mathcal{E}_4$ has several reasons (i.e., B has to be defended against the attack from both C and G), by specifying the extension with which the explanation has to be compared, we can select the specific reasons. In particular, for* Expl = NotAcc*:* $\mathsf{PrfCont}((B, \mathcal{E}_4), (\{B\}, \mathcal{E}_2)) = \{H, J\}$.

- Why A in $\mathcal{E}_2$ given B in $\mathcal{E}_4$? *We take the intersection of the explanations for A and B in their respective extensions, where* Expl = Acc*:* $\mathsf{PrfCont}((A, \mathcal{E}_2), (\{B\}, \mathcal{E}_4)) = \emptyset$.

4   Conclusion

We have proposed a variety of contrastive explanations in the context of abstract argumentation, based on the existing humanities and social sciences literature on contrastiveness. Contrastive explanations have been studied extensively in the XAI literature [28] and some argumentative formalizations have been proposed (see, e.g., [1, 7, 12]). What is new, and different, about this paper, is that we start from the literature on contrastiveness and propose several ways in which one might turn (argumentative) explanations contrastive. To the best of our knowledge, this is the first work in which the variety of notions on contrastiveness have been formalized in one setting, for use in XAI. It is therefore a starting point for a discussion on contrastiveness in (argumentative) XAI.

With this study we pave the way for other researchers to work with contrastive explanations based on specific notions in the literature. By avoiding restrictions on the (non-)acceptance explanations (recall Notation 1), researchers outside the argumentation community can benefit from this work as well.

Now that we have general notions of contrastiveness, we and other researcher working on contrastive explanations can study when to apply which definition and how to apply them outside abstract argumentation. Moreover, we have now restricted our study to arguments, we can extend this to consider extensions, (dynamic) argumentation frameworks or elements of arguments in the structured setting.

References

[1] Leila Amgoud. Explaining black-box classification models with arguments. In *Proceedings of ICTAI'21*, pages 791–795. IEEE, 2021.

[2] Charles Antaki and Ivan Leudar. Explaining in conversation: Towards an argument model. *European Journal of Social Psychology*, 22(2):181–194, 1992.

[3] Abdallah Arioua and Madalina Croitoru. Formalizing explanatory dialogues. In Christoph Beierle and Alex Dekhtyar, editors, *Scalable Uncertainty Management*, pages 282–297. Springer, 2015.

[4] Katie Atkinson, Trevor Bench-Capon, and Danushka Bollegala. Explanation in AI and law: Past, present and future. *Artificial Intelligence*, 289:103387, 2020.

[5] Eric Barnes. Why P rather than Q? The curiosities of fact and foil. *Philosophical Studies*, 73:35–53, 1994.

[6] Pietro Baroni, Martin Caminada, and Massimilliano Giacomin. Abstract argumentation frameworks and their semantics. In Pietro Baroni, Dov Gabay, Massimiliano Giacomin, and Leon van der Torre, editors, *Handbook of Formal Argumentation*, pages 159–236. College Publications, 2018.

[7] Philippe Besnard, Sylvie Doutre, Théo Duchatelle, and Marie-Christine Lagasquie-Schiex. Question-Based Explainability in Abstract Argumentation. Research Report IRIT/RR–2022–01–FR, IRIT : Institut de Recherche en Informatique de Toulouse, France, April 2022.

[8] Philippe Besnard, Alejandro García, Antony Hunter, Sanjay Modgil, Henry Prakken, Guillermo Simari, and Francesca Toni. Introduction to structured argumentation. *Argument & Computation*, 5(1):1–4, 2014.

[9] Floris Bex and Douglas Walton. Combining explanation and argumentation in dialogue. *Argument & Computation*, 7(1):55–68, 2016.

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

[11] AnneMarie Borg and Floris Bex. Necessary and sufficient explanations for argumentation-based conclusions. In *Proceedings of ECSQARU'21*, pages 45–58. Springer, 2021.

[12] AnneMarie Borg and Floris Bex. Contrastive explanations for argumentation-based conclusions. In *Proceedings of AAMAS'22*, pages 1551–1553, 2022.

[13] Kristijonas Čyras, Antonio Rago, Emanuele Albini, Pietro Baroni, and Francesca Toni. Argumentative XAI: A survey. In Zhi-Hua Zhou, editor, *Proceedings of IJCAI'21*, pages 4392–4399. ijcai.org, 2021.

[14] Phan Minh Dung. 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.

[15] Xiuyi Fan and Francesca Toni. On computing explanations in argumentation. In Blai Bonet and Sven Koenig, editors, *Proceedings of AAAI'15*, pages 1496–1502. AAAI Press, 2015.

[16] Xiuyi Fan and Francesca Toni. On explanations for non-acceptable arguments. In Elizabeth Black, Sanjay Modgil, and Nir Oren, editors, *Proceedings of TAFA'15*, LNCS 9524, pages 112–127. Springer, 2015.

[17] Denis Hilton. Conversational processes and causal explanation. *Psychological Bulletin*, 107(1):65–81, 1990.

[18] Dmitry Kazhdan, Zohreh Shams, and Pietro Liò. Marleme: A multi-agent reinforcement learning model extraction library. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.

[19] Peter Lipton. Contrastive explanation. *Royal Institute of Philosophy Supplement*, 27:247–266, 1990.

[20] Hugo Mercier and Dan Sperber. Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, 34(2):57–74, 2011.

[21] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

[22] Tim Miller. Contrastive explanation: a structural-model approach. *The Knowledge Engineering Review*, 36:e14, 2021.

[23] Henry Prakken and Rosa Ratsma. A top-level model of case-based argumentation for explanation: formalisation and experiments. *Argument & Computation*, (Preprint):1–36, 2021.

[24] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1:206–215, 2019.

[25] Zeynep Saribatur, Johannes Wallner, and Stefan Woltran. Explaining non-acceptability in abstract argumentation. In *Proceedings of the 24th European Conference on Artificial Intelligence (ECAI'20)*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 881–888. IOS Press, 2020.

[26] Kacper Sokol and Peter Flach. Counterfactual explanations of machine learning predictions: Opportunities and challenges for AI safety. In *2019 AAAI Workshop on Artificial Intelligence Safety (SafeAI'19)*, volume 2301. CEUR Workshop Proceedings, 2019.

[27] Kacper Sokol and Peter Flach. One explanation does not fit all: The promise of interactive explanations for machine learning transparency. *Künstliche Intelligenz*, 34:235–250, 2020.

[28] Ilia Stepin, Jose Alonso, Alejandro Catala, and Martin Pereira-Fariña. A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access*, 9:11974–12001, 2021.

[29] Sjoerd Timmer, John-Jules Meyer, Henry Prakken, Silja Renooij, and Bart Verheij. A two-phase method for extracting explanatory arguments from bayesian networks. *International Journal of Approximate Reasoning*, 80:475–494, 2017.

[30] Markus Ulbricht and Johannes P. Wallner. Strong explanations in abstract argumentation. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI'21)*, volume 35, pages 6496–6504, 2021.

[31] Jeroen Van Bouwel and Erik Weber. Remote causes, bad explanations? *Journal for the Theory of Social Behaviour*, 32(4):437–449, 2002.

[32] Petri Ylikoski. The idea of contrastive explanandum. In Johannes Persson and Petri Ylikoski, editors, *Rethinking Explanation*, pages 27–42. Springer, 2007.