Computational Models of Argument
F. Toni et al. (Eds.)

doi:10.3233/FAIA220152

Explaining Change in Quantitative Bipolar Argumentation[1]

Timotheus KAMPIK [a,b], Kristijonas ČYRAS [c]
[a] *Umeå University, Umeå, Sweden*
[b] *SAP Signavio, Berlin, Germany*
[c] *Ericsson Research, Stockholm, Sweden*
ORCiD ID: Timotheus Kampik https://orcid.org/0000-0002-6458-2252, Kristijonas Čyras https://orcid.org/0000-0002-4353-8121

Abstract. This paper presents a formal approach to explaining change of inference in Quantitative Bipolar Argumentation Frameworks (QBAFs). When drawing conclusions from a QBAF and updating the QBAF to then again draw conclusions (and so on), our approach traces changes – which we call *strength inconsistencies* – in the partial order that a semantics establishes on the arguments in the QBAFs. We trace the strength inconsistencies to specific arguments, which then serve as explanations. We identify both sufficient and counterfactual explanations for strength inconsistencies and show that our approach guarantees that explanation arguments exist if and only if an update leads to strength inconsistency.

Keywords. quantitative argumentation, explainable AI, non-monotonic reasoning

1. Introduction

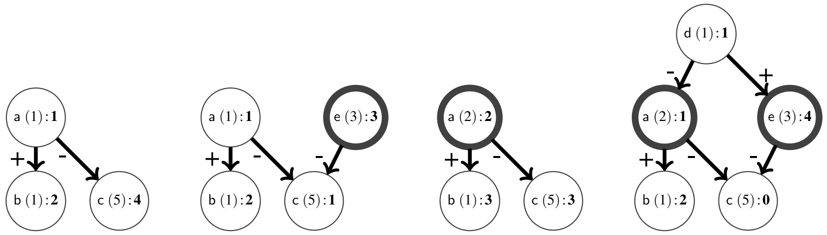
A key challenge in the domain of eXplainable Artificial Intelligence (XAI) is the explanation of an agent's *change of mind*: if the agent has inferred (or decided) A at time t_0, why does she infer A' at t_1? This challenge is reflected in fundamental approaches to decision-making and reasoning. From the perspective of microeconomic decision theory (see, e.g. [1]), a basic assumption is that the agent has *consistent preferences*, i.e. assuming two independent choices A and A', the agent must not decide A and then A' if A' has been available as a decision option all along and also A is still available, as long as no relevant change in circumstances has occurred. If an agent's preferences on the available decision options are not consistent, one would expect an explanation that highlights this relevant change in circumstances that violates the *ceteris paribus*[2] condition. From an automated reasoning perspective, one would expect that an explanation is provided if monotony of entailment is violated, i.e. if the agent first infers A and then A', such that $A \not\subseteq A'$, an explanation of why previously inferred statements are to be rejected should be provided. In this paper, we define such explanations in the setting of evolving Quantitative Bipolar Argumentation Frameworks (QBAFs) [2].

[1] This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

[2] Translates to: "all else unchanged" and is a crucial assumption in classical models of economic rationality.

Specifically, our goal is to explain, in a QBAF that was updated by changing arguments, their initial credences and/or relationships, the relative change in acceptability of specified arguments. We strive for explanations of changes in arguments' relative strengths that pertain to in some sense minimal information causing those changes. We adopt the notions of (attributive) sufficient explanations and counterfactual explanations (see [3] for an excellent overview of counterfactual explanations) to the setting of explaining changes in the partial ordering of argument strengths in evolving QBAFs. In the example below, we give intuitive readings of the introduced concepts; rigorous definitions follow later.

Example 1
*We start with the QBAF depicted in Figure 1.1, which we denote by QBF. We have the nodes (*arguments*)* a *(with initial strength $\tau(\mathsf{a}) = 1$),* b *(with $\tau(\mathsf{b}) = 1$), and* c *(with $\tau(\mathsf{c}) = 5$);* a *supports* b *and attacks* c*. Here,* b *and* c *are topic arguments, i.e. arguments that we want to weigh against each other: the topic argument with the highest Final Strength (FS) can be considered the most* promising. a *is a support argument, i.e. the final strength $\sigma(\mathsf{a})$ is not directly relevant to the decision but impacts the FS of (some) topic arguments. Typically, we determine $\sigma(\mathsf{x})$ of an argument* x *by aggregating the FS of its supporters and attackers. For instance, we can add to $\tau(\mathsf{x})$ the FS of supporters of* x *and subtract the FS of attackers of* x*, iteratively, starting with the neither attacked nor supported leaf arguments (whose FS equals initial strength). Here, we get $\sigma(\mathsf{b})$ by adding $\sigma(\mathsf{a}) = \tau(\mathsf{a})$ to $\tau(\mathsf{b})$: $1+1=2$; and $\sigma(\mathsf{c})$ by subtracting $\sigma(\mathsf{a})$ from $\tau(\mathsf{c})$: $5-1=4$. Consequently,* c *is the is the topic argument with the highest FS (and hence our recommendation).*



1.1: *QBF* 1.2: *QBF*$^{\mathsf{e}}$ 1.3: *QBF*$^{\mathsf{a}}$ 1.4: *QBF*$'$

Figure 1. *QBF* and different updates thereof. Here and henceforth, a node labelled x (i) : **f** carries argument x with initial strength $\tau(\mathsf{x}) = i$ and final strength $\sigma(\mathsf{x}) = \mathbf{f}$. Edges labelled + and − respectively represent attack and support. Arguments with bold borders are strength inconsistency explanation arguments.

Later, our knowledge base receives an update. The update can be of different forms of changes to the QBAF: we give examples of the different resulting situations in Figures 1.2, 1.3, 1.4. As we will spell out shortly, we determine the FSs of b *and* c *(using the same approach as before) in each situation and find that, after any of the updates,* b*, rather than* c*, is the highest-ranking topic argument. We are then interested in explaining why the ranking of* b *relative to* c *has changed.*
(i) In Figure 1.2, the final strength of b *is* 2 *and the final strength of* c *is* 1*. Here, the new argument* e *directly decreases the final strength of* c*. Intuitively, (the addition of)* e *explains the change in the relative ordering of the final strengths for* $\{\mathsf{b}, \mathsf{c}\}$.
(ii) In Figure 1.3, the final strengths of b *and* c *are equal to* 3*. Here, the change in the initial strength of* a *from* 1 *to* 2 *leads to changes to the final strengths of* b *and* c*. Intuitively, (the change in the initial strength of)* a *explains the change in the relative ordering of the final strengths for* $\{\mathsf{b}, \mathsf{c}\}$.

(iii) In Figure 1.4, the final strength of b *is* 2 *and that of* c *is* 0*. Here, we have the addition of new arguments* d *and* e*, as well as a change to the initial strength of* a*, that both influence the final strengths of* b *and* c*. Now, one could say that here all the changes collectively explain the change in the relative ordering of the final strengths for* $\{\mathsf{b},\mathsf{c}\}$*. However, let us search for in some sense* minimal *explanations.*

For instance, the addition of only e suffices *to make* b *stronger than* c*, in the absence of other changes: this is the situation in Figure 1.2. Additionally, since without adding* e *and in the absence of the other changes we would just have QBF we started with as in Figure 1.1, we conclude that* $\{\mathsf{e}\}$ *is a minimal explanation of the change in the relative ordering of the final strengths for* $\{\mathsf{b},\mathsf{c}\}$*.*

Similarly, absent the addition of d *and* e*, with only the change to* a*, we would be in the situation in Figure 1.3, where* c *is not stronger than* b*. Hence,* $\{\mathsf{a}\}$ *is also a minimal explanation of the change in relative ordering of the final strengths for* $\{\mathsf{b},\mathsf{c}\}$*.*

How about other combinations? Absent the change to a *(but with the addition of* e *and* d*), we would find* $\sigma(\mathsf{a}) = 0$ *and thus* $\sigma(\mathsf{b}) = 1 = \sigma(\mathsf{c})$*. I.e. the relative strengths of* b *and* c *would change from QBF. So, intuitively,* $\{\mathsf{d},\mathsf{e}\}$ *also explains the change. But it is not a minimal explanation, because* $\{\mathsf{e}\}$ *is a smaller one. On the other hand, absent the addition of* e*, we would find* $\sigma(\mathsf{a}) = \tau(\mathsf{a}) - \sigma(\mathsf{d}) = 2 - \tau(\mathsf{d}) = 1$*, and the final strengths of* b *and* c *would be* $\sigma(\mathsf{b}) = \tau(\mathsf{b}) + \sigma(\mathsf{a}) = 2$ *and* $\sigma(\mathsf{c}) = \tau(\mathsf{c}) - \sigma(\mathsf{a}) = 4$*, just as in QBF to begin with. So* $\{\mathsf{a},\mathsf{d}\}$ *is not an explanation, for there is no change in the relative strengths of* b *and* c*. Similarly, if the addition of* d *was the only change, we would find* $\sigma(\mathsf{a}) = 0$ *and the final strengths of* b *and* c *equal to their initial strengths. So* d *alone is not an explanation, either.*

In the end, we have two $\subset$*-minimal* sufficient *explanations, namely* $\{\mathsf{a}\}$ *and* $\{\mathsf{e}\}$*, of the change in the relative ordering of the final strengths for* $\{\mathsf{b},\mathsf{c}\}$*. Note, however, that the absence of the changes to* a *does not* counterfactually *restore strength consistency. That is, as shown in the above paragraph, if the initial strength of* a *were* 1 *in* QBF' *(as it is in QBF), we would have* $\sigma(\mathsf{b}) = \sigma(\mathsf{c}) = 1$ *in* QBF'*, whereas* $\sigma(\mathsf{b}) < \sigma(\mathsf{c})$ *in QBF. On the other hand, as shown in the above paragraph, were* e *absent from* QBF'*, we would have* $\sigma(\mathsf{b}) = 2 < 4 = \sigma(\mathsf{c})$ *and strength consistency would be restored: so* $\{\mathsf{e}\}$ *is not only a sufficient explanation, but also a counterfactual one. In fact,* $\{\mathsf{e}\}$ *is a* $\subset$*-minimal counterfactual explanation: any counterfactual explanation entails* e*, because in order to restore strength consistency of* b *and* c *we need to revert (the addition of)* e*.*

The above explanations satisfy the following properties: i) it is sufficient to apply changes to only these arguments (and to ignore the other changes) in the QBAF for the partial order of final strengths to coincide with the one obtained after the actual update (sufficient explanations); ii) in addition to i), reverting the changes made to these arguments only (and keeping all the other changes) restores the original partial order (counterfactual explanations); iii) the set of explanation arguments is $\subset$-minimal among the sets that satisfy i) or ii). These explanations achieve our objective of explaining any change in the partial order that the assignment of the final strengths establishes on a set of arguments of interest, by identifying arguments whose *change* (addition, removal, or change of initial strength) leads to the change in the partial order of the final strengths.

In what follows we formalise the intuition given above by defining and analysing novel forms of explanations in QBAFs. We provide the formal preliminaries in Section 2. We introduce in Section 3 our formal framework for explaining change of inference in

QBAFs. We analyse the properties of our explanations in Section 4. Finally, in Section 5 we discuss our work in the context of related research.

2. Preliminaries

This section introduces the formal preliminaries of our work. Let $\mathbb{I}$ be a set of elements and let $\preceq$ be a preorder on $\mathbb{I}$. Typically, $\mathbb{I} = [0,1]$ is the unit interval[3] and $\preceq = \leqslant$ is the standard less-than-equal ordering. A *quantitative bipolar argumentation framework* contains a set of arguments related by binary *attack* and *support* relations, and assigns an *initial strength* in $\mathbb{I}$ to the arguments. The initial strength can be thought of as initial credence in, or importance of, arguments. Typically, the greater the strength in say the unit interval, the more credible or important the argument is.

Definition 1 (Quantitative Bipolar Argumentation Framework (QBAF) [4,2])
A Quantitative Bipolar Argumentation Framework (QBAF) *is a quadruple* $(Args, \tau, Att, Supp)$ *consisting of a set of arguments* $Args$*, an* attack *relation* $Att \subseteq Args \times Args$*, a* support *relation* $Supp \subseteq Args \times Args$ *and a total function* $\tau : Args \rightarrow \mathbb{I}$ *that assigns the* initial strength $\tau(\mathsf{a})$ *to every* $\mathsf{a} \in Args$*.*

Henceforth, we assume as given a fixed but otherwise arbitrary QBAF $QBF = (Args, \tau, Att, Supp)$, unless specified otherwise. We also assume that $Args$ is finite.

Given $\mathsf{a} \in Args$, the set $Att_{QBF}(\mathsf{a}) := \{\mathsf{b} \mid \mathsf{b} \in Args, (\mathsf{b},\mathsf{a}) \in Att\}$ is the set of attackers of a and each $\mathsf{b} \in Att_{QBF}(\mathsf{a})$ is an *attacker* of a; the set $Supp_{QBF}(\mathsf{a}) := \{\mathsf{c} \mid \mathsf{c} \in Args, (\mathsf{c},\mathsf{a}) \in Supp\}$ is the set of supporters of a and each $\mathsf{c} \in Supp_{QBF}(\mathsf{a})$ is a *supporter* of a. We may drop the subscript $_{QBF}$ when the context is clear.

Reasoning in QBAFs amounts to updating the initial strengths of arguments to their final strengths, taking into account the strengths of attackers and supporters. Specifically, given a QBAF, a strength function assigns final strengths to arguments in the QBAF. Different ways of defining a strength function are called gradual semantics [2,4].

Definition 2 (QBAF Semantics and Strength Functions)
A gradual semantics σ *defines for* $QBF = (Args, \tau, Att, Supp)$ *a* strength function $\sigma_{QBF} : Args \rightarrow \mathbb{I}$ *that assigns the* final strength $\sigma_{QBF}(\mathsf{a})$ *to each argument* $\mathsf{a} \in Args$*.*

For the sake of conciseness, we do not consider the case of a gradual semantics as a partial function that may leave the final strength value of an argument undefined. We may abuse the notation and drop the subscript $_{QBF}$ so that σ denotes the strength function, whenever the context is clear. The (final) strength of an argument can be thought of as its (final) credence or importance. Typically, the greater the strength in $\mathbb{I}$, the more credible or important the argument is. In our examples, we use $\mathbb{I} = \mathbb{R}$.

A gradual semantics can define a strength function as a composition of multivariate real-valued functions that determines the strength of a given argument by aggregating the strengths of its attackers and supporters, taking into account the initial strengths [4]. A strength function so defined is recursive and generally takes iterated updates to produce a sequence of strength vectors, whence the final strengths are defined as the limits (or fixed points) if they exist. However, for *acyclic* QBAFs (without directed cycles) defining a

[3] However, in our examples we use a simplistic semantics and hence a different interval.

semantics and computing the final strengths can be more straightforward: in the topological order of an acyclic QBAF as a graph, start with the leaves,[4] set their final strengths to equal their initial strengths, and then iteratively update the strengths of parents whose all children already have final strengths defined. For instance, in Figure 1.4 from Example 1, we can use the function $\sigma(\mathsf{x}) = \tau(\mathsf{x}) + \left(\sum_{\mathsf{y} \in Supp(\mathsf{x})} \sigma(\mathsf{y}) - \sum_{\mathsf{z} \in Att(\mathsf{x})} \sigma(\mathsf{z})\right)$ defined as a composition, namely sum, of the initial strength ($\tau(\mathsf{x})$) and the difference between the added strengths of the supporters and the added strengths of the attackers ($\sum_{\mathsf{x} \in Supp(\mathsf{x})} \sigma(\mathsf{y}) - \sum_{\mathsf{z} \in Att(\mathsf{x})} \sigma(\mathsf{z})$). It gives final strengths of arguments in the topological order of QBF': first $\sigma(\mathsf{d}) = \tau(\mathsf{d}) = 1$, then $\sigma(\mathsf{a}) = \tau(\mathsf{a}) - \sigma(\mathsf{d}) = 1$ and $\sigma(\mathsf{e}) = \tau(\mathsf{e}) + \sigma(\mathsf{d}) = 4$, and then $\sigma(\mathsf{b}) = \tau(\mathsf{b}) + \sigma(\mathsf{a}) = 2$ and $\sigma(\mathsf{c}) = \tau(\mathsf{c}) - \sigma(\mathsf{a}) - \sigma(\mathsf{e}) = 0$.

While many gradual semantics can be defined for QBAFs in general, their convergence is not always guaranteed in a particular QBAF. For several well-studied semantics, convergence is however always guaranteed in acyclic QBAFs. (See e.g. [4] for a neat exposition of convergence results under various semantics.) In what follows, we restrict our attention to QBAFs for which a fixed but otherwise arbitrary gradual semantics is well-defined. In other words, our study applies to the setting where a gradual semantics σ defines a total strength function σ_{QBF} assigning the final strengths to all arguments of a given QBF. Specifically for illustration purposes to avoid dealing with the sometimes demanding definitions of strength functions, we use acyclic QBAFs and the above strength function σ (in accordance with a topological ordering of an acyclic QBAF). We however note that both the formal definitions and theoretical analysis given in the paper apply to the general setting of well-defined gradual semantics giving total strength functions.

3. Change Explainability in QBAFs

In this section, we introduce our formal approach to change explainability in QBAFs. We start by introducing the notion of *strength consistency*. Henceforth in this section, unless stated otherwise, we let $QBF = (Args, \tau, Att, Supp)$ and $QBF' = (Args', \tau', Att', Supp')$ be QBAFs, let $\mathsf{a}, \mathsf{b}, \mathsf{x}, \mathsf{y} \in Args \cap Args'$, let σ be a strength function, and let $S \subseteq Args \cup Args'$. Let us highlight here that we do not formalise the change operation; instead, we merely assume that we have two QBAFs that have at least two arguments in common, and the second QBAF can be considered a revised (or: *updated*) version of the first one.

Definition 3 (Strength Consistency)
We say that a *is* strength-consistent *w.r.t.* b*, denoted by* $\mathsf{a} \sim_{\sigma,QBF,QBF'} \mathsf{b}$*, iff the following statements hold true:*

- *If* $\sigma_{QBF}(\mathsf{a}) > \sigma_{QBF}(\mathsf{b})$ *then* $\sigma_{QBF'}(\mathsf{a}) > \sigma_{QBF'}(\mathsf{b})$*;*
- *If* $\sigma_{QBF}(\mathsf{a}) < \sigma_{QBF}(\mathsf{b})$ *then* $\sigma_{QBF'}(\mathsf{a}) < \sigma_{QBF'}(\mathsf{b})$*;*
- *If* $\sigma_{QBF}(\mathsf{a}) = \sigma_{QBF}(\mathsf{b})$ *then* $\sigma_{QBF'}(\mathsf{a}) = \sigma_{QBF'}(\mathsf{b})$.

Intuitively, two arguments are strength-consistent only if their relative strengths correspond between the two QBAFs. In an obvious way, $\mathsf{a} \not\sim_{\sigma,QBF,QBF'} \mathsf{b}$ denotes the negation of $\mathsf{a} \sim_{\sigma,QBF,QBF'} \mathsf{b}$ and we say that a and b are *strength-inconsistent*. When there is no ambiguity, we drop the subscripts and write $\mathsf{a} \sim \mathsf{b}$ to denote that a is strength-consistent w.r.t. b, and similarly for the derived notions.

[4] Here, leaves are nodes without incoming edges.

In this work we aim to provide a formal approach to supplying answers to questions regarding changes in arguments' relative strengths in an evolving QBAF. The main objective of this paper is to define explanations as to why, given two QBAFs, any two arguments are strength-inconsistent (or strength-consistent).

As a prerequisite for generating our explanations, we introduce the notion of a QBAF *reversal* with respect to a set of arguments, where such sets of arguments will later play the role of explanations. Colloquially speaking, given QBAFs QBF and its update QBF', a reversal of QBF' to QBF w.r.t. a set of arguments S updates the properties of every argument from S in QBF' so that they reflect the properties of the same argument in QBF: arguments from S that are not in QBF are deleted and arguments from S that are in QBF but not in QBF' are restored.

Definition 4 (QBAF Reversal)
We define the reversal *of* QBF' *to* QBF *w.r.t.* $S \subseteq Args \cup Args'$*, denoted by* $QBF_{\leftarrow QBF'}(S)$*, as a QBAF* $(Args^*, \tau^*, Att^*, Supp^*)$*, where:*

- $Args^* = (Args' \cup S) \setminus (S \setminus Args)$;
- $Att^* = \underbrace{(Att' \setminus (S \times Args))}_{\text{Attacks in } QBF' \text{ that are not from } S \text{ to } Args} \cup \underbrace{(S \times Args^* \cap Att)}_{\text{Attacks in } QBF \text{ from } S \text{ to } Args^*}$;
- $Supp^* = (Supp' \setminus (S \times Args)) \cup (S \times Args^* \cap Supp)$;
- $\tau^* : Args^* \to \mathbb{I}$ *and* $\forall \mathrm{x} \in Args^*$ *the following statement holds true:*

$$\tau^*(\mathrm{x}) = \begin{cases} \tau(\mathrm{x}), & \textit{if } \mathrm{x} \in Args \cap S; \\ \tau'(\mathrm{x}), & \textit{otherwise} . \end{cases}$$

Intuitively: for arguments that were removed (i.e. arguments from $Args \setminus Args'$), those from S are added back; for arguments that were added (i.e. arguments from $Args' \setminus Args$), those from S are removed. The arguments are restored with the associated initial strengths, attacks and supports: in the reversal, we restore "old" attacks and supports from S; we leave "new" attacks and supports unless they are from S to the "old" arguments. For visual intuition, a Venn diagram of the set $Args^*$ is given in Figure 2.

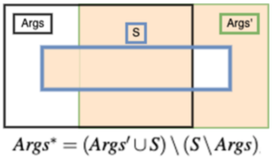


Figure 2. Venn diagram for $Args^*$ (shaded in light and weakly saturated reddish yellow 'sand' colour) in the reversal $QBF_{\leftarrow QBF'}(S) = (Args^*, \tau^*, Att^*, Supp^*)$ of QBF' to QBF w.r.t. $S \subseteq Args' \cup Args$. (Args, Args' and S in small highlighted rectangles are labels of the enclosures highlighted in corresponding colours.)

Using the notion of a QBAF reversal, we introduce different notions of *strength inconsistency explanations*, that are sets of arguments intuitively described as follows:

- $\emptyset$ is both a *sufficient* and a *counterfactual* explanation if we do not find strength inconsistency after the update from QBF to QBF'.
- $S \neq \emptyset$ is a sufficient explanation of strength inconsistency after the update from QBF to QBF' if the inconsistency persists when we reverse everything *except* S back – so changes to S are sufficient for the inconsistency.

- $S \neq \emptyset$ is a counterfactual explanation of strength inconsistency after the update from QBF to QBF' if the inconsistency persists when we reverse everything except S back, but does not persist when we reverse back only S itself – so the absence of changes to S would restore consistency.
- For both sufficient and counterfactual explanations, we define $\subset$-minimal versions.

Definition 5 (Strength Inconsistency Explanations)
We say that $S \subseteq Args' \cup Args$ is a:

- Sufficient Strength Inconsistency (SSI) explanation *of* x *and* y *w.r.t.* σ*,* QBF*, and* QBF' *iff the following statement holds true:*

$$\textit{either} \quad \underbrace{\left(S = \emptyset \textit{ and } \mathsf{x} \sim_{\sigma,QBF,QBF'} \mathsf{y}\right)}_{\mathsf{x} \textit{ and } \mathsf{y} \textit{ are strength-consistent, so empty explanation}}$$

$$\textit{or} \quad \underbrace{\left(\mathsf{x} \not\sim_{\sigma,QBF,QBF'} \mathsf{y} \textit{ and } \mathsf{x} \not\sim_{\sigma,QBF,QBF_{\leftarrow QBF'}((Args \cup Args') \setminus S)} \mathsf{y}\right)}_{\mathsf{x} \textit{ and } \mathsf{y} \textit{ are strength-inconsistent and remain so after reversing everything but S back}}$$

$SX(\mathsf{x} \not\sim_{\sigma,QBF,QBF'} \mathsf{y})$ denotes all SSI explanations of x *and* y *w.r.t.* σ*,* QBF*, and* QBF' *and $SX_{\subset_{\min}}(\mathsf{x} \not\sim_{\sigma,QBF,QBF'} \mathsf{y})$ denotes all $\subset$-minimal SSI explanations of* x *and* y *w.r.t.* σ*,* QBF*, and* QBF'*.*
- Counterfactual Strength Inconsistency (CSI) explanation *of* x *and* y *w.r.t.* σ*,* QBF*, and* QBF' *iff the following statement holds true:*

$$\underbrace{S \in SX(\mathsf{x} \not\sim_{\sigma,QBF,QBF'} \mathsf{y})}_{S \textit{ is an SSI of } \mathsf{x} \textit{ and } \mathsf{y}} \quad \textit{and} \quad \underbrace{\mathsf{x} \sim_{\sigma,QBF,QBF_{\leftarrow QBF'}(S)} \mathsf{y}}_{\mathsf{x} \textit{ and } \mathsf{y} \textit{ become strength-consistent after reversing S}}$$

$CX(\mathsf{x} \not\sim_{\sigma,QBF,QBF'} \mathsf{y})$ denotes all CSI explanations of x *and* y *w.r.t.* σ*,* QBF*, and* QBF' *and $CX_{\subset_{\min}}(\mathsf{x} \not\sim_{\sigma,QBF,QBF'} \mathsf{y})$ denotes all $\subset$-minimal CSI explanations of* x *and* y *w.r.t.* σ*,* QBF*, and* QBF'*.*

Analogously to the case of strength consistency, when there is no ambiguity, we may drop the subscripts and write simply $SX(\mathsf{x} \not\sim \mathsf{y})$ to denote all SSI explanations of x *and* y *(w.r.t. the implicit* σ*,* QBF *and* QBF'*), and similarly for the derived notions.*

Intuitively, a *sufficient* strength inconsistency explanation identifies changes that explain why the relative strengths between two arguments are inconsistent, given an initial QBAF and an update thereof; the changes that a *counterfactual* explanation identifies are – in addition – *counterfactual*, i.e. their absence would restore the initial relative strengths between two arguments. Let us revisit the example from the *Introduction* section to illustrate how strength inconsistency explanations explain change of inference in QBAFs, this time with the formal notation.

Example 2 (Example 1 revisited)
Figures 3.1 and 3.2 depict again the QBAFs $QBF = (\{\mathsf{a},\mathsf{b},\mathsf{c}\}, \tau, \{(\mathsf{a},\mathsf{c})\}, \{(\mathsf{a},\mathsf{b})\})$ and $QBF' = (\{\mathsf{a},\mathsf{b},\mathsf{c},\mathsf{d},\mathsf{e}\}, \tau', \{(\mathsf{a},\mathsf{c}),(\mathsf{e},\mathsf{c}),(\mathsf{d},\mathsf{a})\}, \{(\mathsf{a},\mathsf{b}),(\mathsf{d},\mathsf{e})\})$ from Example 1, where:

- $\tau(\mathsf{a}) = \tau(\mathsf{b}) = 1$ *and* $\tau(\mathsf{c}) = 5$*;*
- $\tau'(\mathsf{a}) = 2$*,* $\tau'(\mathsf{b}) = \tau'(\mathsf{d}) = 1$*,* $\tau'(\mathsf{c}) = 5$ *and* $\tau'(\mathsf{e}) = 3$*.*

Consider the gradual semantics σ defined using the illustrative strength function $\sigma(\mathsf{x}) = \tau(\mathsf{x}) + \left(\sum_{\mathsf{y} \in Supp(\mathsf{x})} \sigma(\mathsf{y}) - \sum_{\mathsf{z} \in Att(\mathsf{x})} \sigma(\mathsf{z})\right)$ that updates the strengths of arguments in an acyclic QBAF according to its topological ordering, as previously discussed. Denote

σ_{QBF} and $\sigma_{QBF'}$ by σ and σ', respectively. Assume we are primarily interested in the final strengths of the arguments b and c: $\sigma(\mathsf{b}) = 2 < 4 = \sigma(\mathsf{c})$. *In contrast,* $\sigma'(\mathsf{b}) = 2 > 0 = \sigma'(\mathsf{c})$. *Hence,* b *is strength-inconsistent w.r.t.* c *(*$\mathsf{b} \not\sim \mathsf{c}$*), for which we have the following explanations: (i)* $SX_{\subset_{\min}}(\mathsf{b} \not\sim \mathsf{c}) = \{\{\mathsf{a}\},\{\mathsf{e}\}\}$*, (ii)* $CX_{\subset_{\min}}(\mathsf{b} \not\sim \mathsf{c}) = \{\{\mathsf{e}\}\}$.

Indeed, for $\{\mathsf{a}\}$*, its relative complement is* $S_{\{\mathsf{a}\}} := (Args \cup Args') \setminus \{\mathsf{a}\} = \{\mathsf{b},\mathsf{c},\mathsf{d},\mathsf{e}\}$*, so that the reversal* $QBF_{\leftarrow QBF'}(S_{\{\mathsf{a}\}})$ *of* QBF' *to* QBF *w.r.t. to* $S_{\{\mathsf{a}\}}$ *has the arguments*

$$\begin{aligned}
&\big(Args' \cup S_{\{\mathsf{a}\}}\big) \setminus \big(S_{\{\mathsf{a}\}} \setminus Args\big) = \\
&(\{\mathsf{a},\mathsf{b},\mathsf{c},\mathsf{d},\mathsf{e}\} \cup \{\mathsf{b},\mathsf{c},\mathsf{d},\mathsf{e}\}) \setminus (\{\mathsf{b},\mathsf{c},\mathsf{d},\mathsf{e}\} \setminus \{\mathsf{a},\mathsf{b},\mathsf{c}\}) = \{\mathsf{a},\mathsf{b},\mathsf{c},\mathsf{d},\mathsf{e}\} \setminus \{\mathsf{d},\mathsf{e}\} = Args.
\end{aligned}$$

Since $Args \cap S_{\{\mathsf{a}\}} = \{\mathsf{b},\mathsf{c}\}$*, it follows that reversing w.r.t. all arguments except* a *yields*

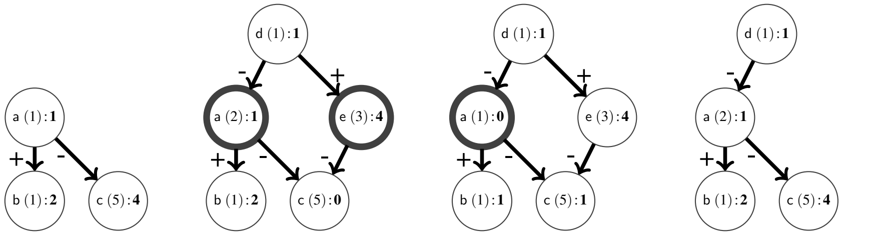
$$\begin{aligned}
&QBF^{\mathsf{a}} := QBF_{\leftarrow QBF'}(S_{\{\mathsf{a}\}}) = (Args^{\mathsf{a}}, \tau^{\mathsf{a}}, Att^{\mathsf{a}}, Supp^{\mathsf{a}}) = \\
&\Big((Args' \cup S_{\{\mathsf{a}\}}) \setminus (S_{\{\mathsf{a}\}} \setminus Args), \tau^{\mathsf{a}}, \big(Att' \setminus (S_{\{\mathsf{a}\}} \times Args)\big) \cup \big((S_{\{\mathsf{a}\}} \times Args^{\mathsf{a}}) \cap Att\big), Supp^{\mathsf{a}}\Big) = \\
&\big(Args, \{(\mathsf{a},\tau'(\mathsf{a})), (\mathsf{b},\tau(\mathsf{b})), (\mathsf{c},\tau(\mathsf{c}))\}, Att' \cup \emptyset, Supp' \cup \emptyset\big) = \\
&(Args, \{(\mathsf{a},2),(\mathsf{b},1),(\mathsf{c},5)\}, \{(\mathsf{a},\mathsf{c})\}, \{(\mathsf{a},\mathsf{b})\}).
\end{aligned}$$

So QBF^{a} *is like* QBF *but with* a*'s initial strength changed to* 2 *(as depicted in Figure 1.3 and discussed in Example 1), thus giving* $\sigma_{QBF^{\mathsf{a}}}(\mathsf{b}) = 3 = \sigma_{QBF^{\mathsf{a}}}(\mathsf{c})$. *So,* b *and* c *are strength-inconsistent (when updating from* QBF *to* QBF'*) and remain so after reversing everything but* $\{\mathsf{a}\}$ *back. Hence,* $\{\mathsf{a}\}$ *is a* $\subset$*-minimal SSI, by Definition 5.*

Now observe that reversing w.r.t. a *yields*

$$\begin{aligned}
&QBF^{*} := QBF_{\leftarrow QBF'}(\{\mathsf{a}\}) = (Args^{*}, \tau^{*}, Att^{*}, Supp^{*}) = \\
&\Big((Args' \cup \{\mathsf{a}\}) \setminus (\{\mathsf{a}\} \setminus Args), \tau^{*}, \big(Att' \setminus (\{\mathsf{a}\} \times Args)\big) \cup ((\{\mathsf{a}\} \times Args^{*}) \cap Att), Supp^{*}\Big) = \\
&\big(Args', \{(\mathsf{a},\tau(\mathsf{a})), (\mathsf{b},\tau'(\mathsf{b})), (\mathsf{c},\tau'(\mathsf{c})), (\mathsf{d},\tau'(\mathsf{d})), (\mathsf{e},\tau'(\mathsf{e}))\}, Att', Supp'\big) = \\
&\big(Args', \{(\mathsf{a},1),(\mathsf{b},1),(\mathsf{c},5),(\mathsf{d},1),(\mathsf{e},3)\}, Att', Supp'\big).
\end{aligned}$$

So QBF^{*} *is like* QBF' *but with* a*'s initial strength unchanged from* 1 *(depicted in Figure 3.3), thus giving* $\sigma_{QBF^{*}}(\mathsf{b}) = 1 = \sigma_{QBF^{*}}(\mathsf{c})$. *That is,* b *and* c *do* ***not*** *become strength-consistent after reversing* $\{\mathsf{a}\}$ *(i.e.* $\mathsf{b} \not\sim_{\sigma,QBF,QBF_{\leftarrow QBF'}(\{\mathsf{a}\})} \mathsf{c}$*), whence* $\{\mathsf{a}\}$ *is* ***not*** *a CSI.*



3.1: QBF 3.2: QBF' 3.3: $QBF_{\leftarrow QBF'}(\{\mathsf{a}\})$. 3.4: $QBF_{\leftarrow QBF'}(\{\mathsf{e}\})$.

Figure 3. QBAFs for explanations from Example 1.

For $\{\mathsf{e}\}$*, with* $S_{\{\mathsf{e}\}} := (Args \cup Args') \setminus \{\mathsf{e}\} = \{\mathsf{a},\mathsf{b},\mathsf{c},\mathsf{d}\}$ *we have that* $\big(Args' \cup S_{\{\mathsf{e}\}}\big) \setminus \big(S_{\{\mathsf{e}\}} \setminus Args\big) = \{\mathsf{a},\mathsf{b},\mathsf{c},\mathsf{d},\mathsf{e}\} \setminus (\{\mathsf{a},\mathsf{b},\mathsf{c},\mathsf{d}\} \setminus \{\mathsf{a},\mathsf{b},\mathsf{c}\}) = \{\mathsf{a},\mathsf{b},\mathsf{c},\mathsf{d},\mathsf{e}\} \setminus \{\mathsf{d}\} = \{\mathsf{a},\mathsf{b},\mathsf{c},\mathsf{e}\}$.

It follows that reversing w.r.t. all arguments except e *yields*

$$QBF^{\mathsf{e}} := QBF_{\leftarrow QBF'}(S_{\{\mathsf{e}\}}) =$$

$$\big(\{\mathsf{a},\mathsf{b},\mathsf{c},\mathsf{e}\},\{(\mathsf{a},\tau(\mathsf{a})),(\mathsf{b},\tau(\mathsf{b})),(\mathsf{c},\tau(\mathsf{c})),(\mathsf{e},\tau'(\mathsf{e}))\},\{(\mathsf{a},\mathsf{c}),(\mathsf{e},\mathsf{c})\},\{(\mathsf{a},\mathsf{b})\}\big) =$$

$$\big(\{\mathsf{a},\mathsf{b},\mathsf{c},\mathsf{e}\},\{(\mathsf{a},1),(\mathsf{b},1),(\mathsf{c},5),(\mathsf{e},3)\},\{(\mathsf{a},\mathsf{c}),(\mathsf{e},\mathsf{c})\},\{(\mathsf{a},\mathsf{b})\}\big).$$

So QBF^{e} *is QBF with* e *and the attack* (e,c) *added (as depicted in Figure 1.2 and discussed in Example 1), thus giving* $\sigma_{QBF^{\mathsf{e}}}(\mathsf{b}) = 2$ *and* $\sigma_{QBF^{\mathsf{e}}}(\mathsf{c}) = 1$*. That is,* b *and* c *remain strength-inconsistent after reversing everything but* $\{\mathsf{e}\}$ *back, and so* $\{\mathsf{e}\}$ *is a* $\subset$*-minimal SSI. Further, reversing w.r.t.* e *yields*

$$QBF^{**} := QBF_{\leftarrow QBF'}(\{\mathsf{e}\}) = (Args^{**},\tau^{**},Att^{**},Supp^{**})$$

$$\big((Args' \cup \{\mathsf{e}\}) \setminus (\{\mathsf{e}\} \setminus Args),\tau^{**},\big(Att' \setminus (\{\mathsf{e}\} \times Args)\big) \cup ((\{\mathsf{e}\} \times Args^{**}) \cap Att),Supp^{**}\big) =$$

$$\big(\{\mathsf{a},\mathsf{b},\mathsf{c},\mathsf{d}\},\{(\mathsf{a},\tau'(\mathsf{a})),(\mathsf{b},\tau'(\mathsf{b})),(\mathsf{c},\tau'(\mathsf{c})),(\mathsf{d},\tau'(\mathsf{d}))\},\{(\mathsf{a},\mathsf{c}),(\mathsf{d},\mathsf{a})\},\{(\mathsf{a},\mathsf{b})\}\big).$$

QBF^{**} *is thus like* QBF' *but without* e *(depicted in Figure 3.4), giving* $\sigma_{QBF^{**}}(\mathsf{b}) = 2$ *and* $\sigma_{QBF^{**}}(\mathsf{c}) = 4$*. So* b *and* c ***do** become strength-consistent after reversing* $\{\mathsf{e}\}$*, whence* $\{\mathsf{e}\}$ ***is** a CSI. Clearly, reversing w.r.t.* $\emptyset$ *yields* QBF'*, so that* $\emptyset$ *is not a CSI, and hence* $\{\mathsf{e}\}$ *is also a* $\subset$*-mininmal CSI.*

Lastly, one can check that $\{\mathsf{d}\}$ *is not an SSI and hence cannot be a CSI: as mentioned in Example 1, adding only* d *leaves the final strengths of* b *and* c *unchanged from their initial strengths, so does not explain anything. Thus,* $\{\mathsf{e}\}$ *is the only* $\subset$*-mininmal CSI.*

4. Theoretical Analysis

In this section, we let $QBF = (Args,\tau,Att,Supp)$ and $QBF' = (Args',\tau',Att',Supp')$ be QBAFs, $\mathsf{x},\mathsf{y} \in Args \cap Args'$, and σ be a strength function. We show that both minimal sufficient and counterfactual explanations are sound and complete: either we have strength inconsistency and at least one non-empty set (and no empty set) of explanation arguments or we have strength consistency explained by the empty set (and only by the empty set).

First, if two arguments are strength-consistent given two QBAFs in which they occur and a gradual semantics, then there is no strength inconsistency to explain and the only explanation is the empty set ($SX_{\subset_{\min}}$-soundness).

Proposition 1 ($SX_{\subset_{\min}}$-Soundness)
If $\mathsf{x} \sim \mathsf{y}$*, then* $SX_{\subset_{\min}}(\mathsf{x} \not\sim \mathsf{y}) = \{\emptyset\}$.

Proof. Let $\mathsf{x} \sim \mathsf{y}$. Then $\emptyset$ is an SSI directly by Definition 5. It is clearly $\subset$-minimal, so $\{\emptyset\} \subseteq SX_{\subset_{\min}}(\mathsf{x} \not\sim \mathsf{y})$. On the other hand, no $S \neq \emptyset$ can be an SSI, by definition, precisely because $\mathsf{x} \sim \mathsf{y}$. So $SX_{\subset_{\min}}(\mathsf{x} \not\sim \mathsf{y}) \subseteq \{\emptyset\}$. Hence, $SX_{\subset_{\min}}(\mathsf{x} \not\sim \mathsf{y}) = \{\emptyset\}$ as required. □

If arguments are strength-inconsistent though, then there exists an explanation, but no empty explanation ($SX_{\subset_{\min}}$-completeness).

Proposition 2 ($SX_{\subset_{\min}}$-Completeness)
If $\mathsf{x} \not\sim \mathsf{y}$*, then* $|SX_{\subset_{\min}}(\mathsf{x} \not\sim \mathsf{y})| \geq 1$ *and* $\emptyset \notin SX_{\subset_{\min}}(\mathsf{x} \not\sim \mathsf{y})$.

Proof. Let $\mathsf{x} \not\sim_{\sigma,QBF,QBF'} \mathsf{y}$.

Proof of $|SX_{\subset_{\min}}(\mathsf{x} \not\sim_{\sigma,QBF,QBF'} \mathsf{y})| \geq 1$. By definition of an SSI, since $\mathsf{x} \not\sim_{\sigma,QBF,QBF'} \mathsf{y}$, any $S \subseteq Args \cup Args'$ is an SSI of x and y (w.r.t. σ, QBF, and QBF') iff $\mathsf{x} \not\sim_{\sigma,QBF,QBF_{\leftarrow QBF'}((Args \cup Args') \setminus S)} \mathsf{y}$. Suppose for a contradiction that such a set S does not exist: $\forall S \subseteq Args \cup Args'$, $\mathsf{x} \sim_{\sigma,QBF,QBF_{\leftarrow QBF'}((Args \cup Args') \setminus S)} \mathsf{y}$. Trivially then, $\mathsf{x} \sim_{\sigma,QBF,QBF_{\leftarrow QBF'}((Args \cup Args') \setminus (Args \cup Args'))} \mathsf{y}$. Since $QBF_{\leftarrow QBF'}(\emptyset) = QBF'$ by definition of QBAF reversal (Definition 4), it follows that $\mathsf{x} \sim_{\sigma,QBF,QBF'} \mathsf{y}$, contradicting $\mathsf{x} \not\sim_{\sigma,QBF,QBF'} \mathsf{y}$. By contradiction, there is at least one $S \in SX_{\subset_{\min}}(\mathsf{x} \not\sim_{\sigma,QBF,QBF'} \mathsf{y})$.

Proof of $\emptyset \notin SX_{\subset_{\min}}(\mathsf{x} \not\sim_{\sigma,QBF,QBF'} \mathsf{y})$. Suppose $\emptyset \in SX(\mathsf{x} \not\sim_{\sigma,QBF,QBF'} \mathsf{y})$ for a contradiction. Since $\mathsf{x} \not\sim_{\sigma,QBF,QBF'} \mathsf{y}$, we have $\mathsf{x} \not\sim_{\sigma,QBF,QBF_{\leftarrow QBF'}((Args \cup Args') \setminus \emptyset)} \mathsf{y}$, by definition of an SSI. As $QBF_{\leftarrow QBF'}(Args \cup Args') = QBF$ by definition of QBAF reversal, it follows that $\mathsf{x} \not\sim_{\sigma,QBF,QBF} \mathsf{y}$. But this is in direct contradiction to the definition of strength consistency (Definition 3). Thus, $\emptyset \notin SX(\mathsf{x} \not\sim_{\sigma,QBF,QBF'} \mathsf{y})$, and hence $\emptyset \notin SX_{\subset_{\min}}(\mathsf{x} \not\sim_{\sigma,QBF,QBF'} \mathsf{y})$. □

We can prove analogous properties for $\subset$-minimal CSIs.

Proposition 3 ($CX_{\subset_{\min}}$-soundness)
If $\mathsf{x} \sim \mathsf{y}$, *then* $CX_{\subset_{\min}}(\mathsf{x} \not\sim \mathsf{y}) = \{\emptyset\}$.

Proof. Let $\mathsf{x} \sim_{\sigma,QBF,QBF'} \mathsf{y}$. By definition, a CSI is an SSI S for which $\mathsf{x} \sim_{\sigma,QBF,QBF_{\leftarrow QBF'}(S)} \mathsf{y}$. Since $QBF_{\leftarrow QBF'}(\emptyset) = QBF'$ and $\mathsf{x} \sim_{\sigma,QBF,QBF'} \mathsf{y}$, we find $\mathsf{x} \sim_{\sigma,QBF,QBF_{\leftarrow QBF'}(\emptyset)} \mathsf{y}$, whence $\emptyset$ is a CSI. Clearly, it is a unique $\subset$-minimal CSI. □

Proposition 4 ($CX_{\subset_{\min}}$-completeness)
If $\mathsf{x} \not\sim \mathsf{y}$, *then* $|CX_{\subset_{\min}}(\mathsf{x} \not\sim \mathsf{y})| \geq 1$ *and* $\emptyset \notin CX_{\subset_{\min}}(\mathsf{x} \not\sim \mathsf{y})$.

Proof. Let $\mathsf{x} \not\sim_{\sigma,QBF,QBF'} \mathsf{y}$.

Proof of $|CX_{\subset_{\min}}(\mathsf{x} \not\sim_{\sigma,QBF,QBF'} \mathsf{y})| \geq 1$. Consider $S = Args \cup Args'$. First note that $QBF_{\leftarrow QBF'}((Args \cup Args') \setminus (Args \cup Args')) = QBF_{\leftarrow QBF'}(\emptyset) = QBF'$, and so $\mathsf{x} \not\sim_{\sigma,QBF,QBF_{\leftarrow QBF'}((Args \cup Args') \setminus (Args \cup Args'))} \mathsf{y}$. Thus, $Args \cup Args' \in SX(\mathsf{x} \not\sim_{\sigma,QBF,QBF'} \mathsf{y})$. Now, since $QBF_{\leftarrow QBF'}(Args \cup Args') = QBF$ and $\mathsf{x} \sim_{\sigma,QBF,QBF} \mathsf{y}$ holds true by definition, we have that $\mathsf{x} \sim_{\sigma,QBF,QBF_{\leftarrow QBF'}(Args \cup Args')} \mathsf{y}$. Thus, by definition, $Args \cup Args' \in CX(\mathsf{x} \not\sim_{\sigma,QBF,QBF'} \mathsf{y})$, so that the non-empty $CX(\mathsf{x} \not\sim_{\sigma,QBF,QBF'} \mathsf{y})$ must have at least one $\subset$-minimal element. Therefore, $|CX_{\subset_{\min}}(\mathsf{x} \not\sim_{\sigma,QBF,QBF'} \mathsf{y})| \geq 1$.

Proof of $\emptyset \notin CX_{\subset_{\min}}(\mathsf{x} \not\sim_{\sigma,QBF,QBF'} \mathsf{y})$. Since a $\subset$-minimal CSI is an SSI, if $\emptyset$ were an SSI, then $\emptyset$ would be a $\subset$-minimal SSI, contradicting Proposition 2. □

The above results show that there are non-trivial (i.e. non-empty) sufficient and counterfactual strength inconsistency explanations if and only if a strength inconsistency results between two arguments after an update to a given QBAF. We deem this a desirable property: one needs to explain only if a change in the relative strengths of arguments actually happens after an update; and if there are explanations of changes in the relative strengths of arguments, then they should correctly refer to such changes.

5. Discussion

In this paper, we introduced explanations for changes in the relative strengths of two arguments after a QBAF update; explanations are in the form of sets of arguments that have been changed (added, removed, changed in their initial score or outgoing attacks and supports). Intuitively, a change by means of a set of arguments E provides a sufficient explanation of an alteration in the relative strengths of some arguments of interest if it suffices to change E without making other changes to obtain the alteration in question. Additionally, E is a counterfactual explanation if the absence of change to E would revert back the alteration in the relative strengths of the arguments of interest, even with all the other changes present. Our approach helps to answer a key explainability question – "why $\mathtt{b}$ and no longer $\mathtt{a}$?" – in dynamic quantitative bipolar argumentation.

To our knowledge, this is the first paper on explainability in quantitative bipolar argumentation. Our explanations are immediately applicable to quantitative (non-bipolar) argumentation, where explainability has not been researched either, with the exception of [5]. There, the authors formalise a notion of *impact* of an argument on the final strength of another argument, roughly as a difference between the final strengths of the latter argument with and without the former argument being present. We instead consider as explanations the changes to arguments that guarantee alterations in the relative strengths of other arguments after a given update to the quantitative argumentation framework.

More generally, our work is positioned at the intersection of argumentation dynamics and explainable argumentation, both of which have been studied in depth: see [6] for a survey on argumentation dynamics, as well as [7] and [8] for surveys on argumentation and explainability. Few works study the intersection of dynamics and explainability *explicitly*. A notable exception is [9], where we studied, in the context of (admissibility-based) abstract argumentation, how the violation of monotony of entailment can be explained in so-called *normal expansion* scenarios, in which new arguments are added to an argumentation framework, but the relation among previously existing arguments remains unchanged. The present work is different in that it i) addresses QBAFs, and ii) explains strength inconsistency (i.e. change in preferences from a decision-theoretical perspective) rather than the violation of monotony of entailment.

However, several argumentation explainability approaches consider dynamics *implicitly*. For instance, assuming some space of modifications in a given argumentation framework, the modifications that would change some topic argument's acceptability status (or strength) can be seen as explanations of such a change [10,11,12]. In particular, a collection of additions or removals of arguments or attacks in an abstract argumentation framework in a way that changes the acceptability of a specific argument is an explanation in e.g. [13,12]. Relatedly, though not directly concerning changes, [14,15,16] define explanations, roughly speaking, as sets of arguments (in non-quantitative argumentation frameworks) that are sufficient for acceptance or rejection of some target argument(s).

Our work is on QBAFs instead, concerning gradual semantics and changes to numerical argument strengths. We also defined counterfactual explanations, rather than necessary ones: for comparison, in Example 1, neither $\mathtt{a}$ nor $\mathtt{e}$ could be said to be necessary explanations, because changing neither one alone is needed for strength inconsistency; rather, $\mathtt{e}$ is counterfactual in that the absence of its change guarantees strength consistency back, given all other changes. Collectively, $\{\mathtt{a}, \mathtt{e}\}$ could be said to be necessary, as changing at least one element therein is needed in any combination of changes that leads to strength inconsistency. We leave formal investigations of this for future work. In the future we can also expand the current perspective on QBAF (change) explainability by in addition providing sub-graphs to trace sets of explanation arguments to topic arguments.

References

[1] M.J. Osborne and A. Rubinstein, *Models in Microeconomic Theory*, Open Book Publishers, 2020. doi:10.11647/OBP.0204.

[2] P. Baroni, A. Rago and F. Toni, From Fine-Grained Properties to Broad Principles for Gradual Argumentation: A Principled Spectrum, *International Journal of Approximate Reasoning* **105** (2019), 252–286. doi:10.1016/j.ijar.2018.11.019.

[3] I. Stepin, J.M. Alonso, A. Catala and M. Pereira-Farina, A Survey of Contrastive and Counterfactual Explanation Generation Methods for Explainable Artificial Intelligence, *IEEE Access* **9** (2021), 11974–12001. doi:10.1109/ACCESS.2021.3051315.

[4] N. Potyka, Extending Modular Semantics for Bipolar Weighted Argumentation, in: *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '19, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 2019, pp. 1722–1730–. ISBN ISBN 9781450363099.

[5] J. Delobelle and S. Villata, Interpretability of Gradual Semantics in Abstract Argumentation, in: *15th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, Vol. 11726 LNAI, G. Kern-Isberner and Z. Ognjanovic, eds, Springer, Belgrade, 2019, pp. 27–38. ISSN 16113349. ISBN ISBN 9783030297640.

[6] S. Doutre and J.-G. Mailly, Constraints and changes: A survey of abstract argumentation dynamics, *Argument & Computation* **9** (2018), 223–248. ISBN ISBN 1946-2174. doi:10.3233/AAC-180425.

[7] A. Vassiliades, N. Bassiliades and T. Patkos, Argumentation and explainable artificial intelligence: a survey, *The Knowledge Engineering Review* **36** (2021), e5. doi:10.1017/S0269888921000011.

[8] K. Čyras, A. Rago, E. Albini, P. Baroni and F. Toni, Argumentative XAI: A Survey, in: *30th International Joint Conference on Artificial Intelligence*, Z.-H. Zhou, ed., IJCAI, Montreal, 2021, pp. 4392–4399. ISBN ISBN 978-0-9992411-9-6. doi:10.24963/ijcai.2021/600.

[9] T. Kampik and K. Čyras, Explanations of Non-Monotonic Inference in Admissibility-based Abstract Argumentation, in: *Logic and Argumentation (to appear)*, P. Baroni, C. Benzmüller and Y.N. Wáng, eds, Springer International Publishing, Cham, 2020.

[10] T. Wakaki, K. Nitta and H. Sawamura, Computing Abductive Argumentation in Answer Set Programming, in: *6th International Workshop on Argumentation in Multi-Agent Systems*, P. McBurney, I. Rahwan, S. Parsons and N. Maudet, eds, Springer, Budapest, 2009, pp. 195–215. doi:10.1007/978-3-642-12805-9_12.

[11] R. Booth, D.M. Gabbay, S. Kaci, T. Rienstra and L. van der Torre, Abduction and Dialogical Proof in Argumentation and Logic Programming, in: *21st European Conference on Artificial Intelligence*, T. Schaub, G. Friedrich and B. O'Sullivan, eds, Frontiers in Artificial Intelligence and Applications, Vol. 263, IOS Press, Prague, 2014, pp. 117–122. doi:10.3233/978-1-61499-419-0-117.

[12] C. Sakama, Abduction in Argumentation Frameworks, *Journal of Applied Non-Classical Logics* **28**(2–3) (2018), 218–239. doi:10.1080/11663081.2018.1487241.

[13] X. Fan and F. Toni, On Computing Explanations for Non-Acceptable Arguments, in: *Theory and Applications of Formal Argumentation - 3rd International Workshop*, E. Black, S. Modgil and N. Oren, eds, Lecture Notes in Computer Science, Vol. 9524, Springer, Buenos Aires, 2015, pp. 112–127. doi:10.1007/978-3-319-28460-6_7.

[14] Z.G. Saribatur, J.P. Wallner and S. Woltran, Explaining Non-Acceptability in Abstract Argumentation, in: *24th European Conference on Artificial Intelligence*, G.D. Giacomo, A. Catalá, B. Dilkina, M. Milano, S. Barro, A. Bugarín and J. Lang, eds, IOS Press, Santiago de Compostela, 2020, pp. 881–888. doi:10.3233/FAIA200179.

[15] M. Ulbricht and J.P. Wallner, Strong Explanations in Abstract Argumentation, in: *35th Conference on Artificial Intelligence*, AAAI, 2021, pp. 6496–6504.

[16] A. Borg and F. Bex, Necessary and Sufficient Explanations for Argumentation-Based Conclusions, in: *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, J. Vejnarová and N. Wilson, eds, Springer International Publishing, Cham, 2021, pp. 45–58. ISBN ISBN 978-3-030-86772-0.