# Change in quantitative bipolar argumentation: Sufficient, necessary, and counterfactual explanations

**Timotheus Kampik**[1,2] , **Kristijonas Čyras**[3] , **José Ruiz Alarcón**[1,4]
[1]Umeå University, Sweden
[2]SAP Signavio, Germany
[3]Ericsson, USA
[4]Ericsson, Sweden
tkampik@cs.umu.se, kristijonas.cyras@ericsson.com, jose.ruiz.alarcon@ericsson.com

## Extended Abstract

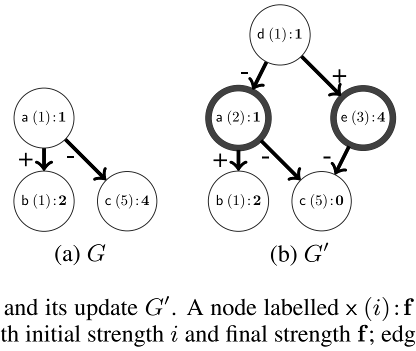
In the 2024 paper authored by Kampik, Čyras, and Alarcón, we address an interesting challenge in the domain of eXplainable Artificial Intelligence (XAI) – we aim to explain an agent's change of mind: if the agent has inferred a set of decisions $A$ at time $t_0$, why does it infer another set of decisions $A'$ at $t_1 > t_0$? We formulate this challenge in the setting of formal argumentation (Bench-Capon and Dunne 2007): we present an approach to *explaining change of inference* in Quantitative Bipolar Argumentation Frameworks (QBAFs) (Baroni, Rago, and Toni 2019). In QBAFs, arguments are assigned initial strengths, i.e. they are nodes with (typically numerical) weights. Two binary relations over the nodes model attacks and supports, respectively, between arguments. To draw conclusions from a QBAF, a gradual argumentation semantics (i.e. an inference function) is applied to the graph to infer the *final strengths* of the arguments, considering the initial strengths and graph topology.

We consider drawing conclusions from a QBAF and *updating the QBAF* to then draw conclusions again. Our goal is to **explain the relative change in the final strengths of specified arguments** in a QBAF that is updated by changing its arguments, their initial strengths and/or relationships. We focus on the partial order over argument strengths that a semantics establishes on arguments of interest, called *topic arguments*. We call the changes that flip the ordering of topic arguments *strength inconsistencies*. We trace the causes of strength inconsistencies to specific arguments, which then serve as explanations. We strive for minimal causes as explanations of changes in arguments' relative strengths.

We adopt the notions of (attributive) sufficient, necessary, and counterfactual explanations from the XAI literature to the setting of explaining changes in the partial ordering of argument strengths in evolving QBAFs. Specifically, we identify sufficient, necessary, and counterfactual explanations for strength inconsistencies. We provide a theoretical analysis showing that strength inconsistency explanations are correct in the sense of being sound and complete: an explanation exists if and only if an update leads to strength inconsistency.

In more detail, our objective is explaining any change in the partial order that the assignment of the final strengths establishes on topic arguments in an updated QBAF. We achieve such explanations by identifying arguments whose *change* (addition, removal, or change of initial strength) is pertinent to the change in the order of the final strengths. Our explanations exhibit the following properties.

Arguments in a *sufficient explanation* are such that it suffices to make changes to these arguments to bring about strength inconsistency, even if no other changes that occur in the update of the QBAF are made. Intuitively, it is sufficient to make changes to only these arguments (and safe to ignore the others) to explain strength inconsistency. Meanwhile, arguments in a *minimal counterfactual explanation* lead to strength inconsistency and are such that reverting changes to exactly these arguments (while keeping the other changes that occur in the update of the QBAF) restores strength consistency. Intuitively, making changes to these arguments explains strength inconsistency while reverting changes to these arguments explains strength consistency. Lastly, arguments in a *necessary explanation* are such that (i) they are sufficient for strength inconsistencyand (ii) it is necessary to make changes to at least one of these arguments to bring about strength inconsistency, whether or not other changes that occur in the update of the QBAF are made. Intuitively, without changes to at least one of these arguments, there would be no strength inconsistency to explain.



(a) $G$ (b) $G'$

Figure 1: $G$ and its update $G'$. A node labelled $\mathsf{x}$ $(i)$ : $\mathbf{f}$ carries argument $\mathsf{x}$ with initial strength $i$ and final strength $\mathbf{f}$; edges labelled $+$ and $-$ represent support and attack. Arguments with bold borders are strength inconsistency explanation arguments, explaining the change in relative strength of the topic arguments $\mathsf{b}$ and $\mathsf{c}$.

By way of an example, consider QBAF $G$ in Figure 1(a) and its update $G'$ in Figure 1(b), with topic arguments $\mathsf{b}$ and $\mathsf{c}$ and final strengths of arguments already determined using a particular semantics (we omit the details). In $G$, $\mathsf{c}$ has the highest final strength. In $G'$, we have both addition of new arguments $\mathsf{d}$ and $\mathsf{e}$ and relationships thereof, and a change to the initial strength of $\mathsf{a}$: the resulting final strength of $\mathsf{b}$ is $2$ and that of $\mathsf{c}$ is $0$. We aim to explain why the ranking of $\mathsf{b}$ relative to $\mathsf{c}$ changed. One could say that all the changes from $G$ to $G'$ collectively explain the change in the relative strengths for $\{\mathsf{b}, \mathsf{c}\}$. However, let us search for sets of arguments that are in some sense *minimal* explanations.

For instance, one can inspect that the addition of only $\mathsf{e}$ *suffices* to make $\mathsf{b}$ stronger than $\mathsf{c}$, in the absence of other changes. Also, without adding $\mathsf{e}$ and in the absence of the other changes we would just have $G$ we started with. We conclude that $\{\mathsf{e}\}$ is a minimal sufficient explanation of the change in the relative ordering of the final strengths for $\{\mathsf{b}, \mathsf{c}\}$. Similarly, without the addition of $\mathsf{d}$ and $\mathsf{e}$, with only the change to $\mathsf{a}$, $\mathsf{c}$ is not stronger than $\mathsf{b}$. Hence, $\{\mathsf{a}\}$ is also a minimal sufficient explanation of the change in relative ordering of the final strengths for $\{\mathsf{b}, \mathsf{c}\}$.

Another kind of explanatory change that we can observe in updating $G$ to $G'$ is a *counterfactual* one: which changes, if reverted back while keeping all the other changes, would annul the relative change in the ranking of the final strengths of $\mathsf{b}$ and $\mathsf{c}$ (i.e. would restore strength consistency)? E.g. to restore strength consistency, it does not suffice to revert (changes to) $\mathsf{a}$ while keeping the other changes. But if $\mathsf{e}$ were absent from $G'$, with $\mathsf{a}$ and $\mathsf{d}$ as they are, we would have $\mathsf{c}$ stronger than $\mathsf{b}$; so, counterfactually, if the addition of $\mathsf{e}$ had not taken place, strength consistency would not have happened. In other words, it suffices to revert (changes to) $\mathsf{e}$ to restore strength consistency assuming that all other changes take place. Thus, $\{\mathsf{e}\}$ is a counterfactual explanation: (the change to) $\mathsf{e}$ both leads to strength inconsistency on its own and, if reverted, would restore strength consistency while keeping the other changes. In fact, $\{\mathsf{e}\}$ is $\subset$-minimally such: when keeping all the other changes, we *must* revert (the addition of) $\mathsf{e}$ in order to restore strength consistency of $\mathsf{b}$ and $\mathsf{c}$ in $G'$; otherwise, we were to revert nothing and thus witness strength inconsistency in $G'$.

Let us now see which changes are actually *necessary* for, i.e. entailed by, the change in the relative strengths of the topic arguments $\mathsf{b}$ and $\mathsf{c}$. First, changes to neither only $\mathsf{a}$ nor only $\mathsf{e}$ could be said to be necessary, because changing neither one specifically is needed for strength inconsistency (precisely because both $\mathsf{a}$ and $\mathsf{e}$ are individually sufficient). Clearly from the above, $\mathsf{d}$ is not necessary either. Instead, collectively $\{\mathsf{a}, \mathsf{e}\}$ can be said to be necessary, as it is needed to make a change with respect to some argument in $\{\mathsf{a}, \mathsf{e}\}$ to explain $\mathsf{c}$ ceasing to be stronger than $\mathsf{b}$ when updating $G$ to $G'$. In other words, if both changes with respect to $\mathsf{a}$ and $\mathsf{e}$ were absent, $\mathsf{c}$ would still be stronger than $\mathsf{b}$. So $\{\mathsf{a}, \mathsf{e}\}$ is a necessary explanation as a set of arguments, changes to at least some of which are *needed* to (entailed by) the change in the relative strengths of the topic arguments, whether or not changes to other arguments happen.

To define explanations, in our work we introduce the notion of a *QBAF reversal*, roughly understood thus: given QBAFs $G$ and its update $G'$, a reversal of $G'$ to $G$ with respect to a set of arguments $S$ updates the properties of every argument from $S$ in $G'$ so that they reflect the properties of the same argument in $G$, namely that arguments from $S$ that are not in $G$ are deleted and arguments from $S$ that are in $G$ but not in $G'$ are restored. We consider acyclic QBAFs and explanations are theoretically applicable to any gradual semantics thereof. In addition to formal definitions and analysis, we implement a method for generating explanations. For that, we formally establish assumptions that speed up the search for explanations and then describe algorithms and a software implementation (in C with Python bindings). Our implementation is applicable to acyclic QBAFs and well-defined semantics that give total strength functions and satisfy directional connectedness (roughly, path-reachable strength dependence). A basic empirical evaluation (with all the code being available at http://s.cs.umu.se/t6xfz2) shows that explanation generation is reasonably fast for QBAFs of smaller sizes that are not particularly densely connected; i.e., the implemented tool can presumably handle argumentation graphs that model the statements and relationships in "human-like" argumentation dialogues reasonably well. Computing explanations given denser or larger QBAFs is costly (and takes substantial amounts of working memory). We explicitly list some potential scaling improvements and consider our work as the first crucial step towards implementing explanations of argument strength changes in QBAFs.

To the best of our knowledge, we were the first to focus on explainability in quantitative bipolar argumentation. More broadly, we focused on explaining change of inference in formal argumentation – colloquially, answering "Why A and no longer B?" instead of simply "Why B?" – that is not limited to a particular set of argumentation semantics. We contributed with novel forms of explanations of inference in dynamically evolving QBAFs; an application of our approach to abstract argumentation can also be found in the paper. Due to the increased research interest in quantitative (bipolar) argumentation, e.g. because of its potential in application scenarios such as explainable recommendation systems, we hope that our work meaningfully complements the research on explainability in formal argumentation. More generally, our assumption is that tracing the reasons for change of inference to the nodes that have been updated in graph-based representations is of general interest to the broader KR&R community as well as to XAI at large.

## References

Baroni, P.; Rago, A.; and Toni, F. 2019. From fine-grained properties to broad principles for gradual argumentation: A principled spectrum. *International Journal of Approximate Reasoning* 105:252–286.

Bench-Capon, T. J. M., and Dunne, P. E. 2007. Argumentation in artificial intelligence. *Artificial Intelligence* 171(10-15):619–641.

Kampik, T.; Čyras, K.; and Alarcón, J. R. 2024. Change in quantitative bipolar argumentation: Sufficient, necessary, and counterfactual explanations. *International Journal of Approximate Reasoning* 164:109066.