Repairing Assumption-Based Argumentation Frameworks

Anna Rapberger[1] and **Markus Ulbricht**[2]
[1]Imperial College London, Department of Computing
[2]Leipzig University, ScaDS.AI
[1]a.rapberger@imperial.ac.uk [2]mulbricht@informatik.uni-leipzig.de

Abstract

The field of formal argumentation is driven by situations where conflicting information need to be balanced out argumentatively. However, if the given knowledge base does not induce any reasonable viewpoint, these methods are stretched to their limits. In this paper, we address this issue in the context of assumption-based argumentation (ABA). More specifically, we study repairing notions for knowledge bases where no assumption can be accepted. We develop genuine repairing techniques for ABA, based on the modification of the building blocks of ABA frameworks, i.e., rules and assumptions. Thereby, we start from basic operators towards more and more fine-grained approaches. We compare their behavior to each other and demonstrate their compliance with suitable repairing desiderata.

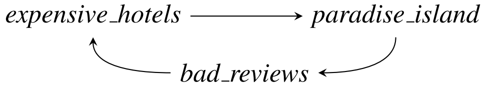
1 Introduction

Reasoning with inconsistent and conflicting knowledge is one of the core competences of knowledge representation and reasoning formalisms (van Harmelen, Lifschitz, and Porter 2008). Providing transparent and reliable ways to manage and resolve inconsistencies is becoming increasingly important; particularly given the potential for invalid and unreliable inferences generated by back-box models and AI-driven products (Huang et al. 2023). Argumentative conflict resolution encompasses non-monotonic and conflict sensitive reasoning which renders formal models of argumentation ideally suited to address these challenges (Bench-Capon and Dunne 2007; Vassiliades, Bassiliades, and Patkos 2021; Rago et al. 2023). Argumentative models have been extensively studied (Gabbay et al. 2021). One of the most prominent formalisms is *assumption-based argumentation (ABA)* (Čyras et al. 2018), a versatile modeling approach that supports applications in, e.g., healthcare (Craven et al. 2012; Cyras et al. 2021), law (Dung, Thang, and Hung 2010) and robotics (Fan et al. 2016), and can be suitably deployed in multi-agent settings to support dialogues (Fan and Toni 2014). The building blocks of ABA frameworks (ABAFs) are assumptions and inference rules; the former are the defeasible elements of the formalism that enable reasoning even in the face of inconsistencies in the knowledge base.

Despite their beneficial and well-explored abilities to handle conflicts, argumentative methods reach their limits when confronted with entirely inconsistent knowledge bases where it is impossible to derive any conclusion. Let us consider the following illustrative example.

Example 1.1. *Our agent, let us call her Alice, plans her summer holiday. Ideally, she would like to spend her holiday at Paradise island; however, it is likely that the hotel prices exceed her budget as she was told that they are quite high. After a bit of research, she comes across a website with disastrous reviews for the island's hotels. Alice is unsure whether she should trust the reviews; an island with such a good reputation will certainly be a decent place to stay. However, given the devastating reviews, Alice starts doubting that the hotel prices are really as expensive as she was told. Taking all information into account, our agent finds herself incapable of reaching a decision.*

We can model the situation in ABA as follows. We consider three assumptions: paradise_island, expensive_hotels and bad_reviews; arrows indicate asymmetric conflicts (attacks) between our assumptions.



A fundamental acceptance criterion in argumentation is admissibility which requires conflict-freeness and the ability to counter-attack each attacker. In Alice's knowledge base, no assumption set is admissible. Consequently, the knowledge base is inconsistent.

Our agent in the example is unable to draw any plausible conclusion; thereby preventing her from reaching a decision. In practice, there are several reasons that can lead to inconsistent knowledge bases such as modeling errors or the merging of different knowledge bases; rendering the given knowledge base inapplicable for further use. To overcome such issues, researchers have invested considerable effort into exploring methods for *repairing* knowledge bases. In his seminal work (Reiter 1987), Reiter introduced diagnostic reasoning to identify inconsistencies and ensure (minimal) change in restoring consistency. Subsequently, extensive research in repairing knowledge bases across various knowledge representation and reasoning formalisms has been conducted, including argumentation (Baumann and Ulbricht 2019; Lehtonen, Niskanen, and Järvisalo 2018; Amgoud and Vesic 2009), description logic (Lembo et al. 2010;

Baader et al. 2018), logic programs (Merhej, Schockaert, and Cock 2017; Gebser et al. 2010; Sakama and Inoue 2003), probabilistic conditional logic (Potyka and Thimm 2014), as well as general non-monotonic logics (Brewka, Thimm, and Ulbricht 2019).

A common approach to restore coherence involves retracting parts of the knowledge base. In the context of our ABA example, this could be achieved in several ways.

Example 1.2. *Taking a closer look at our running example reveals several plausible modifications.*

- *A possible repair for the ABAF would be the* removal of an assumption. *One could argue that the website where Alice found the bad reviews was not very trustworthy; thereby justifying the removal of bad_reviews. As a result, Alice has no doubt to believe that the hotels are expensive.*
- *Alternatively, we could* remove an inference rule *to repair the ABAF. Rather than ignoring the reviews, the modeling issue could be due to an overestimation of market effects. A more nuanced way to repair the ABAF would be the retraction of the inference "bad reviews lead to price drop". Alice can now conclude that the hotels are expensive.*
- *One might, however, argue that there is some truth behind supply and demand mechanisms. A more fine-grained repair operator is therefore to* render rules defeasible *instead of deleting them entirely. This can be achieved by the addition of certain conditions that indicate when or when not a rule is applicable. In this particular situation, the bad reviews might not have shown effect yet, maybe because they are relatively new. Since Alice wants to go on vacation quite soon, she can expect high hotel prices.*

In the field of formal argumentation, research on repairing predominantly focused on abstract argumentation (Dung 1995), where arguments are considered atomic entities. In this paper, we address this issue in the context of ABA. We introduce different repairing notions for knowledge bases where no assumption can be accepted. We identify fundamental properties a repairing operator should satisfy and investigate them w.r.t. our proposed repairing approaches.

In more detail, our main contributions are as follows.

- We discuss how to approach repairing in structured argumentation formalisms and develop desiderata for the behavior of suitable operators. Section 3
- We discuss how to repair by modifying the assumptions of the given ABAF. To this end we propose an operator that removes certain assumptions which cause the semantical collapse within the knowledge base. We show that our operator behaves well w.r.t. our desiderata. Section 4
- We then study repairing in terms of the underlying rules. We propose an operator that renders a given rule set inapplicable. As it turns out, this is a more fine-grained approach compared to our assumption-based operator which also satisfies most of our desiderata. Section 5
- Finally, we consider an operator that renders rules defeasible instead of entirely inapplicable. We formally show that this approach is more flexible. Section 6
- For all our operators we study the computational complexity of verifying minimality of a diagnosis.

2 Background

We recall *assumption-based argumentation (ABA)* (Čyras et al. 2018). We assume a deductive system $(\mathcal{L}, \mathcal{R})$, where $\mathcal{L}$ is a formal language, i.e., a set of sentences, and $\mathcal{R}$ a set of inference rules over $\mathcal{L}$. A rule $r \in \mathcal{R}$ has the form $a_0 \leftarrow a_1, \dots, a_n$ with $a_i \in \mathcal{L}$; $head(r) = a_0$ is the head and $body(r) = \{a_1, \dots, a_n\}$ the (possibly empty) body of r.

Definition 2.1. *An ABA framework (ABAF) is a tuple $(\mathcal{L}, \mathcal{R}, \mathcal{A}, ^{-})$ s.t. $(\mathcal{L}, \mathcal{R})$ is a deductive system, $\mathcal{A} \subseteq \mathcal{L}$ a set of assumptions, and $^{-} : \mathcal{A} \to \mathcal{L}$ a partial contrary function.*

Notice that we deviate from the original ABAF definition by allowing the contrary function to be partial in this work.

Assumption 2.2. *In this work, we focus on* flat *and* finite *ABAFs, i.e., $head(r) \notin \mathcal{A}$ for each $r \in \mathcal{R}$ (no assumption can be derived), and $\mathcal{L}$, $\mathcal{R}$, $\mathcal{A}$ are finite; moreover, each rule is stated explicitly (given as input); and $\mathcal{L}$ is a set of atoms.*

An atom $p \in \mathcal{L}$ is *tree-derivable* from assumptions $S \subseteq \mathcal{A}$ and rules $R \subseteq \mathcal{R}$, denoted $S \vdash_R p$, if p can be derived from the set S of assumptions and the rules in R; i.e.,there is a finite rooted labeled tree $\mathtt{t}$ such that the root is labeled with p, the set of labels for the leaves of $\mathtt{t}$ is equal to S or $S \cup \{\top\}$, and for every inner node v of $\mathtt{t}$ there is a rule $r \in R$ such that v is labelled with $head(r)$, the number of successors of v is $|body(r)|$ and every successor of v is labelled with a distinct $a \in body(r)$ or $\top$ if $body(r) = \emptyset$.

Definition 2.3. *Each tree-derivation $S \vdash_R p$ is an* ABA argument*; $S \vdash_R p$ is a* rule-induced argument *iff $R \neq \emptyset$ and an* assumption argument *iff $R = \emptyset$.*

We drop the subscript R whenever it does not cause confusion. By $Th_{\mathcal{D}}(S) = \{p \in \mathcal{L} \mid \exists S' \subseteq S : S' \vdash p\}$ we denote the set of all conclusions derivable from a set S in an ABAF $\mathcal{D}$. Note that $S \subseteq Th_{\mathcal{D}}(S)$ since, by definition, $\{a\} \vdash_{\emptyset} a$ for each assumption a. We let $\overline{S} = \{\overline{a} \mid a \in S\}$.

Example 2.4. *We model Example 1.1 via $\mathcal{D} = (\mathcal{L}, \mathcal{R}, \mathcal{A}, ^{-})$ with $\mathcal{A} = \{paradise\_island, expensive\_hotels, bad\_reviews\}$, their contraries not_paradise_island, cheap_hotels, and good_reviews, resp., $\mathcal{L} = \mathcal{A} \cup \overline{\mathcal{A}}$, and with rules $\mathcal{R}$:*

$$not\_paradise\_island \leftarrow expensive\_hotels$$
$$good\_reviews \leftarrow paradise\_island$$
$$cheap\_hotels \leftarrow bad\_reviews$$

A set $S \subseteq \mathcal{A}$ *attacks* a set $T \subseteq \mathcal{A}$ if for some $a \in T$ we have $\overline{a} \in Th_{\mathcal{D}}(S)$; moreover, S *defends* T iff S attacks each attacker U of T. A set S is *conflict-free* ($S \in cf(\mathcal{D})$) if it does not attack itself; S is *admissible* ($S \in ad(\mathcal{D})$) if it is conflict-free and defends itself. With a little notational abuse we say S attacks a if S attacks the singleton $\{a\}$.

We next recall grounded, complete, preferred, and stable ABA semantics (abbr. gr, co, pr, stb).

Definition 2.5. *Let $\mathcal{D}$ be an ABAF and let $S \in ad(\mathcal{D})$.*

- *$S \in co(\mathcal{D})$ iff S contains every assumption set it defends;*
- *$S \in gr(\mathcal{D})$ iff S is $\subseteq$-minimal in $co(\mathcal{D})$;*
- *$S \in pr(\mathcal{D})$ iff S is $\subseteq$-maximal in $co(\mathcal{D})$;*
- *$S \in stb(\mathcal{D})$ iff S attacks each $\{x\} \subseteq \mathcal{A} \setminus S$.*

An assumption a is *credulously accepted* wrt. semantics σ in an ABAF D iff $a \in \bigcup \sigma(\mathcal{D})$; we write $a \in \mathit{Cred}_\sigma(\mathcal{D})$.

Example 2.6. *Let us again consider our introductory Example 1.1. As already observed in the introduction, no assumption can be accepted under any of the considered semantics, i.e.,* $\sigma(\mathcal{D}) = \{\emptyset\}$ *for* $\sigma \in \{ad, co, pr\}$ *and* $stb(\mathcal{D}) = \emptyset$*. It follows that no assumption is credulously accepted in* $\mathcal{D}$*.*

Graphical Representations We recall two graphical representations of ABAFs that we will use in the present paper.

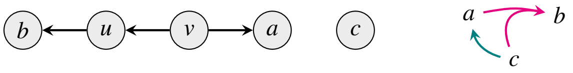
ABAFs are closely related to Dung's *abstract argumentation frameworks (AFs)* (Dung 1995). In brief, an AF is a pair (A, R) consisting of a set of arguments A and an attack relation $R \subseteq A \times A$. We can instantiate an ABAF as AF by setting $A = \{S \vdash p \mid S \subseteq \mathcal{A}, p \in \mathcal{L}\}$ be the set of all ABA arguments and R is the induced attack relation: $S \vdash p$ attacks $T \vdash q$ if $p \in \overline{T}$. Semantics for AFs are defined similar in spirit to ABAFs. It is well-known that the extensions of an ABAF can be computed by evaluating the corresponding AF and computing the extensions; a similar observation can be made for the other direction as well (Čyras et al. 2018).

AFs are well suited to depict the ABA arguments and their relations. *AFs with collective attacks (SETAFs)* (Nielsen and Parsons 2006), on the other hand, are ideally suited to depict the attack structure between the assumptions, as recently outlined (König, Rapberger, and Ulbricht 2022).

Example 2.7. *Consider an ABAF with assumptions* a, b, c*, their contraries* $\overline{a} = s$*,* $\overline{b} = p$*,* $\overline{c} = q$*, and rules* $(p \leftarrow a, c)$ *and* $(s \leftarrow c)$*. We obtain the following arguments:*

$$\{a\} \vdash a \quad \{b\} \vdash b \quad \{c\} \vdash c \quad u = \{a, c\} \vdash p \quad v = \{c\} \vdash s$$

Here, u *and* v *are rule-induced arguments. We can represent the ABAF as AF (left) and SETAF (right) as follows:*



The AF (left) contains all arguments, depicted as black circles with gray background, and the attacks between them (the assumption arguments are called a, b, c, respectively).

The SETAF (right) depicts the attack structure between the assumptions: a *and* c *collectively attack* b *as* $\{a, c\} \vdash p$*, depicted as joint arc; and* c *attacks* a *since* $\{c\} \vdash s$*.*

3 Repairing in Structured Argumentation

In this section, we introduce fundamental concepts for repairing in ABA. In a similar spirit to the research in the AF literature (Baumann and Ulbricht 2019), we consider the following notion of semantical collapse for ABAFs:

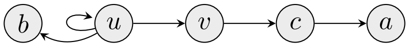
- there is no accepted assumption, i.e., $\mathit{Cred}_\sigma(\mathcal{D}) = \emptyset$.

Note that for stable semantics, this notion amounts to $stb(\mathcal{D}) = \emptyset$ while for the remaining semantics under consideration, this means that only the empty set is an extension. An ABAF can be considered consistent (w.r.t. a given semantics) if some assumption can be accepted.

Definition 3.1. *Let* σ *be any semantics. An ABAF* $\mathcal{D} = (\mathcal{L}, \mathcal{R}, \mathcal{A}, ^-)$ *is called* consistent *iff* $\mathit{Cred}_\sigma(\mathcal{D}) \neq \emptyset$*;* D *is called* inconsistent *iff* $\mathit{Cred}_\sigma(\mathcal{D}) = \emptyset$*.*

Repairing and Abstract Argumentation Research on repairing AFs mainly focuses on removing *arguments* or *attacks* in order to restore consistency (Baumann and Ulbricht 2019; Niskanen and Järvisalo 2020). In this context, many convenient theoretical properties hold, but unfortunately this approach is not applicable to structured argumentation.

Example 3.2. *Consider an ABAF with assumptions a, b, c; contraries* $\overline{a} = c$*,* $\overline{b} = p$ *and* $\overline{c} = q$*; and rules* $r_1 = p \leftarrow a, q$ *and* $r_2 = q \leftarrow b$*. It can be checked that the ABAF is inconsistent w.r.t. any considered semantics. The corresponding AF with* $u = \{a, b\} \vdash p$ *and* $v = \{b\} \vdash q$ *is depicted below.*



A possible repair of the AF would be to remove the attack (u, v)*; then* $\{v, a\}$ *is acceptable w.r.t. admissible semantics. What is the corresponding modification in the knowledge base? In order to remove the attack from* u *to* v*, it is necessary to remove* b *from the set of assumptions in* u*. This can only be achieved by modifying the rule* r_2 *or by removing* b *entirely which causes further severe modifications of the AF.*

It turns out that the abstraction level of AFs is simply too high to apply AF repairing methods. We will thus develop genuine repairing notions for ABA in this work.

Repairing Desiderata An ABA knowledge base consists of multiple parts (literals, assumptions, their contraries, and rules). Consequently, there are multiple conceivable ways to repair a given ABAF. In order to proceed in a structured way, let us formalize desiderata for the behavior of repairing operators first. We assume we are given an arbitrary but fixed repairing operator rep which can turn a given ABAF $\mathcal{D}$ into different ABAFs $img(rep(\mathcal{D}))$.

Example 3.3. *Consider, for instance, a simple repairing operator* rep_c *which, given ABAF* $\mathcal{D}$ *and a set* S *of assumptions, removes the contraries of* S *from the partial contrary function in* $\mathcal{D}$*, i.e., given* $\mathcal{D} = (\mathcal{L}, \mathcal{R}, \mathcal{A}, ^-)$*, we have* $rep_c(\mathcal{D}, S) = (\mathcal{L}, \mathcal{R}, \mathcal{A}, ^-|_{\mathcal{A} \setminus S})$*. Then* $img(rep_c(\mathcal{D}))$ *is the set of all ABAFs we can attain by applying* rep_c *to different* $S \subseteq \mathcal{A}$*, i.e.,* $img(rep_c(\mathcal{D})) = \{rep_c(\mathcal{D}, S) \mid S \subseteq \mathcal{A}\}$*.*

In general, not each attainable ABAF $\mathcal{D}' \in img(rep(\mathcal{D}))$ will be consistent. In fact, even our drastic rep_c does not have this property. However, a repairing operator should be effective in the sense that at least some repair always exists.

(E) Effectiveness: For any ABAF $\mathcal{D}$, there should be some repair, i.e., at least one $\mathcal{D}' \in img(rep(\mathcal{D}))$ is consistent.

Example 3.4. *Given any ABAF* $\mathcal{D} = (\mathcal{L}, \mathcal{R}, \mathcal{A}, ^-)$*, we can consider* $rep_c(\mathcal{D}, \mathcal{A})$*, i.e., we remove any contrary occurring in* $\mathcal{D}$*. Clearly, the resulting ABAF is consistent (each assumption is acceptable). Consequently,* rep_c *satisfies (E).*

Moreover, repairing should be efficient in the sense that the "repaired" ABA $\mathcal{D}'$ can be computed in polynomial time.

(T) Tractable: Given $\mathcal{D}$, any particular $\mathcal{D}' \in img(rep(\mathcal{D}))$ can be computed in polynomial time.

For instance, our artificial operator rep_c satisfies (T) since altering the contrary function can be done efficiently.

Intuitively, a repairing operator should never introduce novel conflicts into the knowledge base. This is, however, hard to formalize because argumentation is driven by conflicting viewpoints. We can still capture the following notion of *self-controversial* assumptions we seek to avoid.

Definition 3.5. *Let σ be any semantics and $\mathcal{D} = (\mathcal{L}, \mathcal{R}, \mathcal{A}, \overline{})$ be an ABAF. An assumption $a \in \mathcal{A}$ is called* self-controversial *if $\overline{a} \in Th_{\mathcal{D}}(a)$. The ABAF $\mathcal{D}$ is called* self-controversial *if each $a \in \mathcal{A}$ is.*

A repairing operator should push $\mathcal{D}$ towards consistency and one might argue that for this, it is never beneficial to introduce self-controversial assumptions.

(SC) Self-Controversial: If a is self-controversial in $\mathcal{D}' \in img(rep(\mathcal{D}))$, then it is already self-controversial in $\mathcal{D}$.

Example 3.6. *Suppose we are given an ABAF $\mathcal{D}$ with some rule $\overline{a} \leftarrow a$ for some assumption $a \in \mathcal{A}$. Then, a is clearly self-controversial. While rep_c can remove self-controversial assumptions (because in $rep_c(\mathcal{D}, \{a\})$, a is not self-controversial anymore), no novel self-controversial assumption can arise. Thus, rep_c satisfies (SC).*

Ideally, a repair operator should preserve every conclusion that can be derived. We consider the following desideratum.

(DP) Derivation Persistence: If $S \vdash p$ in $\mathcal{D}$, then the same is true in each $\mathcal{D}' \in img(rep(\mathcal{D}))$.

On the other hand, we do not want to introduce new derivations when repairing an ABAF. Ideally, the modification should not alter the underlying logic of our knowledge base.

(CM) Conclusion Monotonicity: If $S \vdash p$ in some $\mathcal{D}' \in img(rep(\mathcal{D}))$, then $S \vdash p$ in $\mathcal{D}$.

Example 3.7. *The operator rep_c satisfies (CM) because no new rules are introduced. It also satisfies (DP) because it does not alter the existing derivations.*

Our last desideratum highlights that our auxiliary rep_c is too drastic as it undermines intrinsic properties of the given atoms, instead of fixing modeling errors.

(CP) Contrary Persistence: If $p = \overline{a}$ in $\mathcal{D}$, then the same is true in each $\mathcal{D}' \in img(rep(\mathcal{D}))$.

To see this, let us head back to our introductory example.

Example 3.8. *A possible repair in Example 2.4 is to simply forget that cheap_hotels is the contrary of expensive_hotels. Then bad_reviews does not attack expensive_hotels anymore; despite the presence of cheap_hotels ← bad_reviews. Now, our repaired ABAF no longer aligns well with our intuition.*

Outline In the subsequent sections, we will examine modifications of the acceptability status of assumptions and rules. Starting with the most basic idea, we will modify the ABAF in a way that certain assumptions are rendered "false" and remove them from the knowledge base entirely. This, in turn, frees other assumptions to become acceptable (Section 4). As it will turn out, a more fine-grained approach is removing rules instead (Section 5). We will show that, under mild conditions, rendering assumptions false can be captured by removing rules, thus the latter approach is a faithful generalization of the former. Finally, we discuss techniques to make rules defeasible instead of removing them entirely.

4 Assumption-Based Repairing

In this section we manipulate the assumptions in $\mathcal{D}$ in order to restore consistency. To this end let us introduce the notion of an assumption-based repairing operator.

Definition 4.1. *An* assumption-based repairing operator *is a mapping $(\mathcal{D}, S) \mapsto rep(\mathcal{D}, S)$ that takes as an input an ABAF $\mathcal{D} = (\mathcal{L}, \mathcal{R}, \mathcal{A}, \overline{})$ and a set $S \subseteq \mathcal{A}$ and returns a modified ABAF $rep(\mathcal{D}, S)$.*

That is, we can specify a set S of assumptions that are supposed to be modified by the operator rep.

Definition 4.2. *Let σ be any semantics. Let $\mathcal{D} = (\mathcal{L}, \mathcal{R}, \mathcal{A}, \overline{})$ be an inconsistent ABAF and $S \subseteq \mathcal{A}$. We call S a (minimal)* assumption-based diagnosis *for $\mathcal{D}$ w.r.t rep iff (it is $\subseteq$-minimal s.t.) $rep(\mathcal{D}, S)$ is consistent.*

If no assumption can be accepted, a natural strategy to repair $\mathcal{D}$ is to take a set S of assumptions and set S to "false" manually. This can be done as follows: remove S from the set $\mathcal{A}$ of assumptions (and from the set of literals) and remove each rule making use of S (since these rules are not applicable). The following operator implements this idea.

Definition 4.3. *Let $\mathcal{D} = (\mathcal{L}, \mathcal{R}, \mathcal{A}, \overline{})$ be an ABAF and $S \subseteq \mathcal{A}$ a set of assumptions. We let $rma(\mathcal{D}, S) = (\mathcal{L}', \mathcal{R}', \mathcal{A}', \overline{}')$ be the ABAF with $\mathcal{L}' = \mathcal{L} \setminus S$, $\mathcal{A}' = \mathcal{A} \setminus S$, $\overline{}' = \overline{}|_{\mathcal{A}'}$, and $\mathcal{R}' = \{r \in \mathcal{R} \mid body(r) \cap S = \emptyset\}$*

Let us head back to our introductory example. This operator indeed captures the approach to disregard the reviews.

Example 4.4. *Recall our introductory Example 2.4 where no assumption can be accepted. As already suggested in the introduction, the website where Alice found the reviews may not have been very trustworthy. We therefore simply remove the assumption; we apply rma to $\mathcal{D}$ in order to get rid of bad_reviews. The resulting ABAF $\mathcal{D}'$ has assumptions $\mathcal{A}' = \{$paradise_island, expensive_hotels$\}$, and rules $\mathcal{R}'$:*

not_paradise_island ← expensive_hotels

good_reviews ← paradise_island

The attack structure of $\mathcal{D}'$ looks as follows.

$$\mathit{expensive\_hotels} \longrightarrow \mathit{paradise\_island}$$

We accept $S = \{$expensive_hotels$\}$ (under any semantics). As a result, Alice believes the hotels at the island are expensive, giving her a reason to reconsider her vacation choices.

4.1 Basic Properties

We establish that rma is capable of repairing almost any ABAF. To this end our first observation is that rma does not alter whether or not some assumption is self-controversial.

Lemma 4.5 (Self-Controversial Assumptions)**.** *Any $a \in \mathcal{A}$ is self-controversial in $rma(\mathcal{D}, S)$ iff $a \notin S$ and a is self-controversial in $\mathcal{D}$.*

This means, however, that rma cannot always repair a given ABAF: if each assumption is self-controversial in $\mathcal{D}$, then the same will always be true in $rma(\mathcal{D}, S)$. On the other hand, if $a \in \mathcal{A}$ is not self-controversial, then $S = \mathcal{A} \setminus \{a\}$ trivially repairs $\mathcal{D}$. Consequently, a repair exists under mild conditions. In the context of our repairing desiderata, this means Effectiveness (E) is satisfied in most cases.

Proposition 4.6 (Effectiveness)**.** *Let σ be any semantics. There is a diagnosis S w.r.t. rma for $\mathcal{D}$ iff $\mathcal{D}$ is not self-controversial.*

We next observe that $rma(\mathcal{D}, S)$ preserves all tree-derivations that do not make use of any assumption in S (and does not add any novel derivations). In this sense, it adheres to both Derivation Persistence (DP) and Conclusion Monotonicity (CM) as formalized in Section 3.

Proposition 4.7. *For any ABAF $\mathcal{D}$ and set S of assumptions, there is a tree-derivation $T \vdash p$ in $rma(\mathcal{D}, S)$ iff $T \vdash p$ is a tree-derivation in $\mathcal{D}$ and $T \cap S = \emptyset$.*

Example 4.8. *Recall Example 3.2 before and after applying rma to $S = \{a\}$. In $\mathcal{D}' = rma(\mathcal{D}, S)$ we can derive $\{b\} \vdash b$, $\{c\} \vdash c$, and $\{b\} \vdash q$ which are exactly the derivations in $\mathcal{D}$ that do not rely on the assumption a.*

Furthermore, rma does not change the contrary function, so Contrary Persistence (CP) holds under mild conditions.

Fact 4.9. *Let $\mathcal{D} = (\mathcal{L}, \mathcal{R}, \mathcal{A}, {}^{-})$ be an ABAF and $S \subseteq \mathcal{A}$. If $p = \overline{a}$ and $a \notin S$, then the same is true in $rma(\mathcal{D}, S)$.*

Moreover, rma represents a simple syntactical modification of $\mathcal{D}$ which can be performed easily. Consequently, the tractability (T) requirement is also satisfied.

Fact 4.10 (Tractability)**.** *For any ABAF $\mathcal{D}$ and set S of assumptions, the ABAF $rma(\mathcal{D}, S)$ can be computed in* P.

In summary, the assumption-based operator rma satisfies most of the desiderata under consideration.

Summary 4.11. *The operator rma satisfies **(SC)**, **(CM)**, and **(T)**; a version of **(DP)**; moreover, **(E)** and **(CP)** are satisfied under mild conditions.*

4.2 Minimal Repairs

As the previous fact states, we can compute rma efficiently. Moreover, the reason for Proposition 4.6 to hold is that $\mathcal{A} \setminus \{a\}$ is always a diagnosis for $\mathcal{D}$ as long as a is no self-attacker. However, this is a rather drastic and unnatural way to repair $\mathcal{D}$. Consequently, let us now examine minimal modifications to a given ABAF $\mathcal{D}$. The bad news is that in general, this is a hard problem for most semantics. The reason for this is twofold: First, the underlying decision problem ("*Is S a diagnosis for $\mathcal{D}$?*") is as hard as reasoning in an ABAF (because the special case $S = \emptyset$ amounts to checking whether any assumption is acceptable).

Fact 4.12. *Deciding whether a set S of assumptions is a diagnosis w.r.t. rma for an ABAF $\mathcal{D}$ is i)* NP*-complete for $\sigma \in \{ad, co, pr, stb\}$ and ii) in* P *for gr.*

Second, suppose we are given a diagnosis S and want to test it for minimality. In case of gr semantics this is still tractable; the reason is the following underlying characterization for consistency w.r.t. grounded semantics.

Lemma 4.13. *An ABAF $\mathcal{D} = (\mathcal{L}, \mathcal{R}, \mathcal{A}, {}^{-})$ is consistent w.r.t. gr semantics iff there is some $a \in \mathcal{A}$ s.t. $\emptyset$ defends a in $\mathcal{D}$.*

For the other semantics, however, testing minimality is a hard problem. The reason for this is the non-monotonic behavior of this operator. We end up with the following complexity for testing minimality.

Theorem 4.14. *Deciding whether a set S of assumptions is a minimal diagnosis w.r.t. rma for an ABAF $\mathcal{D}$ is i)* DP*-complete for $\sigma \in \{ad, co, pr, stb\}$ and ii) in* P *for gr.*

We can benefit from the positive result for gr as follows. Due to finiteness of $\mathcal{A}$, it is clear that each diagnosis S can be reduced to a minimal one. This helps reducing the search space when striving for a minimal diagnosis.

Proposition 4.15. *Let $\mathcal{D} = (\mathcal{L}, \mathcal{R}, \mathcal{A}, {}^{-})$ be an inconsistent ABAF and S be a gr diagnosis. There is some $S' \subseteq S$ s.t. S' is a minimal diagnosis w.r.t. rma for $\sigma \in \{ad, co, pr\}$.*

Example 4.16. *Consider $\mathcal{D}$ having assumptions a, b, c; contraries $\overline{a} = a_c$, $\overline{b} = b_c$, and $\overline{c} = c$; and rules $\mathcal{R}$:*

$$r_1 = \overline{a} \leftarrow b \quad r_2 = \overline{b} \leftarrow a \quad r_3 = \overline{a} \leftarrow c \quad r_4 = \overline{b} \leftarrow c$$

Take $S = \{b, c\}$ leaving us with an unattacked assumption a. Clearly, $\{a\} \in gr(rma(\mathcal{D}, S))$ proving that S is a gr-diagnosis. Consequently, S is also a co-diagnosis. However, it is not minimal as removing b was not necessary. We can therefore enhance our diagnosis to $S' = \{c\}$.

5 Rule-Based Repairing

In the previous section, we considered an operator which renders assumptions false. As a side effect, certain rules (determined by the set S of assumptions) are rendered inapplicable as well. In this section, we consider a more fine-grained approach by specifying the rules in question directly.

Definition 5.1. *A* rule-based repairing operator *is a mapping $(\mathcal{D}, R) \mapsto rep(\mathcal{D}, R)$ that takes an ABAF $\mathcal{D} = (\mathcal{L}, \mathcal{R}, \mathcal{A}, {}^{-})$ and a set $R \subseteq \mathcal{R}$ and returns a modified ABAF $rep(\mathcal{D}, R)$.*

Analogous to an assumption-based operator, we can specify a set R of rules supposed to be modified by rep.

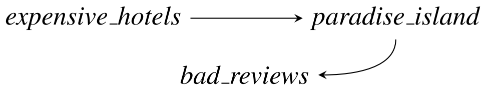
Definition 5.2. *Let σ be any semantics. Let $\mathcal{D} = (\mathcal{L}, \mathcal{R}, \mathcal{A}, {}^{-})$ be an inconsistent ABAF and $R \subseteq \mathcal{R}$. We call R a (minimal)* rule-based diagnosis *for $\mathcal{D}$ w.r.t rep iff (it is $\subseteq$-minimal s.t.) $rep(\mathcal{D}, R)$ is consistent.*

We consider the following rule-based repairing operator.

Definition 5.3. *Let $\mathcal{D} = (\mathcal{L}, \mathcal{R}, \mathcal{A}, {}^{-})$ be an ABAF and let $R \subseteq \mathcal{R}$ be a set of rules. We define*

$$rmr(\mathcal{D}, R) = (\mathcal{L}, \mathcal{R} \setminus R, \mathcal{A}, {}^{-}).$$

Example 5.4. *In Example 4.4, we fixed the ABAF from Example 2.4 by applying rma to remove bad_reviews. Arguably, it is a somewhat drastic step to disregard the reviews of the website as a whole. Maybe, the issue lies in the overestimation of market effects. A more nuanced way to repair the ABAF is to reconsider the validity of the rule cheap_hotels $\leftarrow$ bad_reviews by applying rmr to get rid of the rule. The resulting ABAF $\mathcal{D}''$ still has all assumptions, but again only the rules $\mathcal{R}'$ from the ABAF $\mathcal{D}'$ from Example 4.4, yielding the following attack structure:*



As a result, we accept $S = \{$expensive_hotels, bad_reviews$\}$. Now, our agent can conclude that the hotel prices at the island are expensive although they are rated poorly; Alice might now conclude that the trip is too expensive for her.

5.1 Rule-Based Repairing and Assumption Attacks

Let us now delve deeper into the behavior of our repair operator. First, we observe that rmr does not alter the contraries of the assumptions, therefore satisfying Contrary Persistence (CP). In contrast to our assumption-based repairing operator, we obtain that the contraries of the assumptions are preserved since the assumptions are left untouched.

Fact 5.5 (Contrary Persistence)**.** *For each assumption a, it holds that $p = \overline{a}$ in $\mathcal{D}$ iff $p = \overline{a}$ in $rmr(\mathcal{D}, R)$.*

We furthermore note that rmr satisfies (SC) since it does not introduce self-controversial assumptions.

Lemma 5.6 (Self-Controversial Assumptions)**.** *If an assumption $a \in \mathcal{A}$ is self-controversial in $rmr(\mathcal{D}, R)$ then a is self-controversial in $\mathcal{D}$.*

The other direction, however, is not always true. The removal of a rule can make a controversial assumption uncontroversial. In contrast to our assumption-based repair operator, self-controversiality is no longer the limiting factor for repairing an ABAF—as long as the assumptions are not their own contrary (i.e., as long as $\overline{a} \neq a$). This leads to a crucial observation: the removal of rules cannot break an attack cycle which is caused by the contrary function.

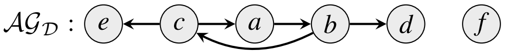
Example 5.7. *Consider an ABAF $\mathcal{D}$ with assumptions a, b, and c with contraries $\overline{a} = b$, $\overline{b} = c$, and $\overline{c} = a$. In an intuitive sense which we will formalize below, the assumptions form an odd cycle. Since $\mathcal{D}$ is flat, the assumptions are not attacked by any other set of assumptions (there is no rule with head $\overline{a}$ because $\overline{a} = b$ is an assumption). Therefore, none of the assumptions can be accepted in $\mathcal{D}$. Removing any rule from $\mathcal{D}$ does not influence their acceptability.*

The repairability of an ABAF $\mathcal{D}$ w.r.t. rmr is closely tied to the attack structure imposed by the contrary function. Striving to formalize this, let us define the so-called *assumption attack graph* of an ABAF $\mathcal{D}$.

Definition 5.8. *For an ABAF $\mathcal{D} = (\mathcal{L}, \mathcal{A}, \mathcal{R}, \overline{})$ the* assumption attack graph *is a directed graph $\mathcal{AG}_{\mathcal{D}} = (V, E)$ where $V = \mathcal{A}$ and $(a, b) \in E$ iff $\overline{b} = a$.*

Intuitively, $\mathcal{AG}_{\mathcal{D}}$ is a graphical depiction of the relation among the contraries of the assumptions in $\mathcal{D}$.

Example 5.9. *Let $\mathcal{D} = (\mathcal{L}, \mathcal{A}, \mathcal{R}, \overline{})$ be the ABAF where $\mathcal{A} = \{a, b, c, d, e, f\}$ and $\overline{a} = c$, $\overline{b} = a$, $\overline{c} = b$, $\overline{d} = b$, $\overline{e} = c$, $\overline{f} = p$, for some atom p which is no assumption. Independent of the rule set $\mathcal{R}$, $\mathcal{AG}_{\mathcal{D}}$ is given as follows:*



We can make several observations about the whole ABAF:

- *Consider e.g. a. Since $\mathcal{D}$ is flat no rule entails $\overline{a} = c$; i.e, the only way to attack a is via the assumption c. Consequently, there is no argument $S \vdash \overline{a}$ in $\mathcal{D}$ except $\{c\} \vdash c$.*
- *Every $a \in \mathcal{A}$ has at most one in-coming attack in $\mathcal{AG}_{\mathcal{D}}$.*
- *We do not know how many tree derivations entail $p = \overline{f}$.*

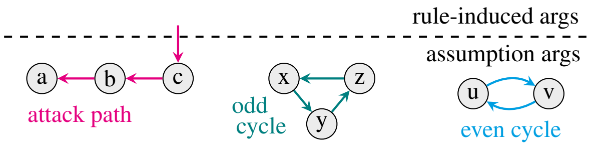


Figure 1: Exemplified attack structures between assumptions

Let us formalize the possible attack configurations between assumptions in an ABAF. We identify three main configurations: even cycles, odd cycles, and proper paths.

Definition 5.10. *Let $\mathcal{D} = (\mathcal{L}, \mathcal{A}, \mathcal{R}, \overline{})$ be an ABAF. A sequence $a_1, \ldots, a_n$ of assumptions with $\overline{a_{i+1}} = a_i$ for all $i < n$ and $a_k \neq a_j$ for all $k, j \leq n$ is called an* attack path. *It is called*

- odd attack cycle *iff $\overline{a_1} = a_n$ and n is odd;*
- even attack cycle *iff $\overline{a_1} = a_n$ and n is even;*
- rooted attack path *iff $\overline{a_1} \notin \mathcal{A}$.*

Note that the assumption a_1 in a rooted attack path is either attacked by a literal $p \in \mathcal{L} \setminus \mathcal{A}$ or unattacked (since the contrary function is partial).

Figure 1 gives examples of the three attack configurations. The dashed line separates rule-induced arguments from assumption arguments. Note that in Figure 1, only c receives attacks from rule-induced arguments. Similar as in Example 5.9, we note that the remaining assumption arguments in the figure receive no further attacks since their contrary is an assumption itself. Now, since every assumption has at most one contrary, we make the following observation.

Fact 5.11. *Each assumption a in an ABAF $\mathcal{D}$ receives only one incoming attack in $\mathcal{AG}_{\mathcal{D}}$.*

It follows that cycles in $\mathcal{AG}_{\mathcal{D}}$ have no further incoming attacks. Hence, the existence of an even attack cycle guarantees the existence of an admissible extension.

Corollary 5.12. *Let $\sigma \in \{ad, co, pr\}$. Each ABAF s.t. $\mathcal{AG}_{\mathcal{D}}$ contains an even cycle is consistent w.r.t. σ.*

In the same vein, if $\mathcal{AG}_{\mathcal{D}}$ contains an odd cycle, then it cannot be broken apart anymore. This causes stable semantics to collapse.

Corollary 5.13. *No ABAF s.t. $\mathcal{AG}_{\mathcal{D}}$ contains an odd cycle has a stable extension.*

On the other hand, an assumption argument $(\{a\} \vdash a)$ can have several outgoing attacks in case a is the contrary of several assumptions. Overall, each assumption can either be traced back to an even cycle, an odd cycle, or a rooted attack path. The next lemma summarizes this observation.

Lemma 5.14. *For each assumption a in an ABAF $\mathcal{D}$, precisely one of the following options is satisfied.*

1. *$\mathcal{AG}_{\mathcal{D}}$ contains an odd cycle with an attack path to a;*
2. *$\mathcal{AG}_{\mathcal{D}}$ contains an even cycle with an attack path to a;*
3. *$\mathcal{AG}_{\mathcal{D}}$ contains a rooted attack path containing a.*

As was already hinted at by Corollary 5.13, we need to be aware of the odd cycles in $\mathcal{AG}_{\mathcal{D}}$. We introduce the notion of *contradictory ABAFs* motivated by this observation.

Definition 5.15. *An ABAF $\mathcal{D} = (\mathcal{L}, \mathcal{A}, \mathcal{R}, ^{-})$ is* contradictory *iff for each $a \in \mathcal{A}$ there is an odd attack cycle with an attack path to a.*

Example 5.16. *The ABAF in Example 5.9 is not contradictory, but only due to f. If f was not present, then any assumption would be attacked by the odd cycle a, b, c.*

The presence of a certain attack structure determines the repairability of an ABAF. The following proposition characterizes the existence of a diagnosis under a given semantics. In terms of our desiderata, this means that rmr satisfies Effectiveness (E) under the specified conditions.

Proposition 5.17 (Effectiveness)**.** *Let $\mathcal{D}$ be an ABAF.*

- *Let $\sigma \in \{ad, co, pr\}$, then $\mathcal{D}$ has a diagnosis $R \subseteq \mathcal{R}$ w.r.t. rmr iff $\mathcal{D}$ is not contradictory.*
- *Let $\sigma = gr$, then $\mathcal{D}$ has a diagnosis $R \subseteq \mathcal{R}$ w.r.t. rmr iff $\mathcal{D}$ contains a rooted attack path.*
- *Let $\sigma = stb$, then $\mathcal{D}$ has a diagnosis $R \subseteq \mathcal{R}$ w.r.t. rmr iff $\mathcal{D}$ contains no odd attack cycle.*

Regarding the remaining desiderata, rmr does not satisfy Derivation Persistence (DP) as it removes rules.

Example 5.18. *Let $\mathcal{D}$ be an ABAF with assumptions a, b and rule $r = (p \leftarrow a, b)$. In $\mathcal{D}$, we have $\{a, b\} \vdash p$; however, we cannot construct the argument in $rmr(\mathcal{D}, \{r\})$ anymore.*

Conclusion Monotonicity (CM), on the other hand, is satisfied. When applying our rmr operator, we do not introduce new derivations in the knowledge base.

Proposition 5.19 (Conclusion Monotonicity)**.** *Let $\mathcal{D} = (\mathcal{L}, \mathcal{A}, \mathcal{R}, ^{-})$ be an ABAF and let $R \subseteq \mathcal{R}$ denote a set of rules. For any $S \subseteq \mathcal{A}$ and $p \in \mathcal{L}$, if $S \vdash p$ in $rmr(\mathcal{D}, r)$ then $S \vdash p$ in $\mathcal{D}$.*

The operator rmr satisfies Tractability (T) since the removal of rules is a simple syntactical modification.

Fact 5.20 (Tractability)**.** *For any ABAF $\mathcal{D}$ and set R of rules, the ABAF $rmr(\mathcal{D}, R)$ can be computed in polynomial time.*

In summary, our rule-based repair operator satisfies most of the desiderata under consideration except (DP). The following theorem summarizes our findings of this section.

Summary 5.21. *The operator rmr satisfies **(CP)**, **(SC)**, **(CM)**, and **(T)**; **(E)** is satisfied under mild conditions.*

5.2 Minimal Repairs

Similar as in the case of assumption-based repairs, we are interested in minimizing the number of rules we remove. This will, however, turn out to be a hard problem again. First of all, the remark pertaining to verifying some repair of course persists, as it is based on the complexity of ABA reasoning.

Fact 5.22. *Deciding whether a set R of rules is a diagnosis w.r.t. rmr for $\mathcal{D}$ is i)* NP*-complete for $\sigma \in \{ad, co, pr, stb\}$ and ii) in* P *for gr.*

The complexity of verifying a minimal rule-based diagnosis is as hard as in the assumption-based case with one noteworthy exception: the gr case is not tractable anymore, but coNP-complete. This confirms the intuition that rmr is more fine-grained compared to rma and thus, the additional possibilities cause the gr case to become harder.

Theorem 5.23. *Deciding whether a set R of rules is a minimal diagnosis w.r.t. rmr for an ABAF $\mathcal{D}$ is i)* DP*-complete for $\sigma \in \{ad, co, pr, stb\}$ and ii)* coNP*-complete for gr.*

However, similar to Section 4, we can use gr-diagnoses to reduce the search space for minimal σ-diagnoses for $\sigma \in \{ad, co, pr\}$. The theoretical underpinning is the following observation, similar to the previous setting.

Fact 5.24. *Let $(\mathcal{L}, \mathcal{R}, \mathcal{A}, ^{-})$ be an ABAF and R be a rule-based gr diagnosis w.r.t. rmr. There is some $R' \subseteq R$ s.t. R' is a minimal rule-based diagnosis for $\sigma \in \{ad, co, pr\}$.*

This is especially interesting in view of the following observation: while verifying minimality of a *given* diagnosis is hard for gr, we can *compute* certain particular minimal ones efficiently. This way, we can make efficient use of Fact 5.24.

Proposition 5.25. *There is a polynomial algorithm that, on input an ABAF $\mathcal{D}$, returns a minimal rule-based gr-diagnosis or detects if none such exists.*

5.3 Connection to Assumption-based Repair

The conditions for repairs to exist differ for rma and rmr. As noted in Proposition 4.6, rma is capable of repairing any ABAF that is not self-controversial, whereas rmr relies on more involved characterizations (cf. Proposition 5.17). On the other hand, as evident when comparing Examples 4.4 and 5.4, it becomes apparent that our assumption-based operator rma in some sense makes use of rmr as it removes rules. It turns out that in the fragment of ABAFs with *separated contraries* in which each ABAF $(\mathcal{L}, \mathcal{A}, \mathcal{R}, ^{-})$ satisfies $\mathcal{A} \cap \overline{\mathcal{A}} = \emptyset$ (Rapberger and Ulbricht 2023), we have indeed a strong connection: here, a diagnosis S w.r.t. rma induces a diagnosis $\mathcal{R}' = \{r \in \mathcal{R} \mid body(r) \cap S \neq \emptyset\}$ w.r.t. rmr for a given ABAF $\mathcal{D}$ under the semantics-specific restrictions we identified in Proposition 5.17.

Proposition 5.26. *Let $\mathcal{D} = (\mathcal{L}, \mathcal{A}, \mathcal{R}, ^{-})$ be an ABAF with separated contraries and σ be any semantics. If $S \subseteq \mathcal{A}$ is a diagnosis w.r.t. rma for $\mathcal{D}$, then $R = \{r \in \mathcal{R} \mid body(r) \cap S \neq \emptyset\}$ is a diagnosis w.r.t. rmr for $\mathcal{D}$.*

An ABAF with separated contraries is characterized by $\mathcal{AG}_{\mathcal{D}} = (V, E)$ being a directed graph with $E = \emptyset$. Due to this, according to Proposition 5.17, such ABAFs can always be repaired w.r.t rmr, but not necessarily w.r.t. rma.

In the general case, the correspondence stated in Proposition 5.26 cannot be obtained, as we illustrate below.

Example 5.27. *Let $\mathcal{D}$ be an ABAF with assumptions $\mathcal{A} = \{a, b, c, d\}$; the assumptions a, b and c are in an odd attack cycle, i.e., $\overline{a} = b$, $\overline{b} = c$ and $\overline{c} = a$. The ABAF contains the rule $r = (\overline{d} \leftarrow d)$, that is, d is self-controversial. We note that $\mathcal{D}$ has a diagnosis w.r.t. both rma and rmr under most considered semantics ($\mathcal{D}$ cannot be repaired w.r.t. rmr under stable semantics by Proposition 5.17). Any diagnosis $S \subseteq \mathcal{A}$ w.r.t. rma is a subset of $\{a, b, c\}$ while the unique diagnosis w.r.t. rmr is $\{r\}$.*

For stable semantics, let $\mathcal{D}'$ be an ABAF with assumptions $\mathcal{A} = \{a, b, c, d\}$ with $\overline{a} = b$ and $\overline{b} = c$, and rules $r_1 = (\overline{d} \leftarrow d)$ (d is self-controversial) and $r_2 = (\overline{d} \leftarrow b)$. A diagnosis for $\mathcal{D}$ w.r.t. rma is $\{c\}$; however, the corresponding induced rule set $R = \emptyset$ does not repair $\mathcal{D}'$ w.r.t. rmr.

Although rma includes the rule removal operation, we cannot infer a diagnosis w.r.t. rmr in general as the examples above demonstrate.

6 Defeasibility-based Repairing

In the previous section, we considered a repair notion which operates on the rules of the ABAF. However, the operation is rather coarse; it removes the rules in the diagnosis entirely, thus preventing it from further use.

In this section, we discuss a more fine-grained notion which preserves the rule but makes it *defeasible*. In contrast to other rule-based formalisms, each rule in (standard) ABA is strict. Defeasibilty of rules can be modeled by adding an assumption to the body of the rule which serves as potential weak point of it. By attacking this dedicated assumption the rule can be refuted/prevented from taking effect.

Definition 6.1. *A* defeasible repairing operator *is a mapping $(\mathcal{D}, R) \mapsto rep(\mathcal{D}, (R, f))$ that takes as an input an ABAF $\mathcal{D} = (\mathcal{L}, \mathcal{R}, \mathcal{A}, ^-)$, a set $R \subseteq \mathcal{R}$ and a function $f : R \to \mathcal{L}$ and returns a modified ABAF $rep(\mathcal{D}, (R, f))$.*

A diagnosis is a tuple (R, f) with rules R and function f.

Definition 6.2. *Let $\mathcal{D} = (\mathcal{L}, \mathcal{R}, \mathcal{A}, ^-)$ be an inconsistent ABAF, σ be any semantics, $R \subseteq \mathcal{R}$, and $f : R \to \mathcal{L}$. We call (R, f) a (minimal)* rule-defeasibility diagnosis *for $\mathcal{D}$ w.r.t rep iff (R is $\subseteq$-minimal s.t.) $rep(\mathcal{D}, (R, f))$ is consistent.*

We define a repair operator that takes a set of rules together with a function that is intuitively understood as the contrary function of the rules. For each rule, we introduce a fresh assumption and add it to the body.

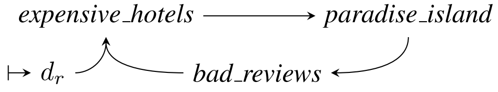
Definition 6.3. *Let $\mathcal{D} = (\mathcal{L}, \mathcal{R}, \mathcal{A}, ^-)$ be an ABAF, $R \subseteq \mathcal{R}$, and $f : R \to \mathcal{L}$ be a function. We let $def(\mathcal{D}, (R, f)) = (\mathcal{L}', \mathcal{R}', \mathcal{A}', ^{-\prime})$ be the ABAF where*

$$\begin{aligned}
\mathcal{L}' &= \mathcal{L} \cup \{d_r \mid r \in R\} \\
\mathcal{A}' &= \mathcal{A} \cup \{d_r \mid r \in R\} \\
^{-\prime} &= ^- \cup \{(d_r, f(r)) \mid r \in R\} \\
\mathcal{R}' &= (\mathcal{R} \setminus R) \cup \{head(r) \leftarrow body(r) \cup \{d_r\} \mid r \in R\}
\end{aligned}$$

Example 6.4. *In Example 5.4, we used rmr ot delete the rule cheap_hotels $\leftarrow$ bad_reviews due to our doubts in market effects. Nevertheless, one might argue that there is some truth behind supply and demand mechanisms. It could be reasonable to make the rule defeasible instead of deleting it. In this particular situation, the reviews might not have shown effect yet, maybe because they are relatively new; however, we can expect that the hotel prices will drop when enough time has passed. We use the operator def:*

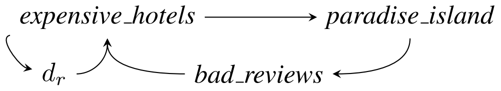
$$cheap\_hotels \leftarrow bad\_reviews, d_r$$

The new assumption d_r can capture, for instance, that the rule applies only in the future (d_r = time_has_passed). A suitable contrary for d_r would be a fact (present $\leftarrow$). Assuming this rule is contained in our knowledge base, we obtain the following attack structure:



Now, S = {expensive_hotels, bad_reviews} can be accepted; d_r is attacked by a fact and d_r and bad_reviews jointly attack expensive_hotels.

Alternatively, also the assumption expensive_hotels can be considered as reasonable contrary of d_r; since it shows that not enough time has passed for the market mechanisms to come into effect; yielding the following attack structure.



Again, we can accept S = {expensive_hotels, bad_reviews}.

In the above example, we exemplified different ways to utilize def for repairing the knowledge base, yielding the accepted set of assumptions {*expensive_hotels*, *bad_reviews*}. The choice of the contrary of the newly introduced semantics is however not uniquely determined. Now, we might end up in the undesired situation that the repair operator returns a consistent ABAF by accepting the newly introduced assumptions only.

Example 6.5. *Let us again consider our running example. Now, let $\overline{d_r} = p$ for some $p \notin Th_{\mathcal{D}}(\mathcal{A})$. Then d_r is unattacked in the modified framework and $\{d_r\}$ is the only accepted set of the ABAF.*

While we do not want to disregard our newly introduced assumptions entirely, we would preferably be able to accept assumptions from the original ABAF as well. We therefore stipulate that a successful repair is required to enable some of the original assumptions to become acceptable.

Definition 6.6. *Let $\mathcal{D} = (\mathcal{L}, \mathcal{A}, \mathcal{R}, ^-)$ be an ABAF and let σ be a semantics. A diagnosis (R, f) w.r.t. def is* successful *iff there is $E \in \sigma(def(\mathcal{D}, (R, f)))$ such that $E \cap \mathcal{A} \neq \emptyset$.*

6.1 Basic Properties

Let us study fundamental properties of our new repair operator. First note that def satisfies Contrary Persistence (CP).

Fact 6.7 (Contrary Persistence)**.** *Let $\mathcal{D} = (\mathcal{L}, \mathcal{A}, \mathcal{R}, ^-)$ be an ABAF. For each assumption $a \in \mathcal{A}$, it holds that $p = \overline{a}$ in $\mathcal{D}$ iff $p = \overline{a}$ in $rmr(\mathcal{D}, R)$.*

Moreover, changing the defeasibility of the rules does not introduce new self-controversial assumptions. As it was the case for rmr, the other direction does not necessarily hold.

Lemma 6.8 (Self-Controversial Assumptions)**.** *If $a \in \mathcal{A}$ is self-controversial in $def(\mathcal{D}, (R, f))$ then a is self-controversial in $\mathcal{D}$.*

Turning now to the Effectiveness (E) of our new operator, we observe many similarities to rmr. In the case of grounded semantics, however, we additionally require that an inconsistent ABAF $\mathcal{D}$ contains an atom $p \in Th_{\mathcal{D}}(\emptyset)$ which can be used to defend some assumption.

Proposition 6.9 (Effectiveness)**.** *Let $\mathcal{D}$ be an ABAF.*

- *For $\sigma \in \{ad, co, pr\}$, $\mathcal{D}$ has a successful diagnosis (R, f), $R \subseteq \mathcal{R}$, $f : R \to \mathcal{L}$, w.r.t. def iff $\mathcal{D}$ is not contradictory.*
- *For $\sigma = gr$, $\mathcal{D}$ has a successful diagnosis (R, f), $R \subseteq \mathcal{R}$, $f : R \to \mathcal{L}$, w.r.t. def iff $\mathcal{D}$ is consistent or there is $p \in Th_{\mathcal{D}}(\emptyset)$ and a rooted attack path with root a s.t. $p \neq \overline{a}$.*

- *For $\sigma = stb$, $\mathcal{D}$ has a successful diagnosis (R, f), $R \subseteq \mathcal{R}$, $f : R \to \mathcal{L}$, w.r.t. def iff $\mathcal{D}$ contains no odd attack cycle.*

In contrast to our rule-based operator rmr, a weaker form of derivation persistence can be established for def.

Example 6.10. *Consider our ABAF $\mathcal{D}$ from Example 5.18 with assumptions a, b and rule $r = (p \leftarrow a, b)$, yielding the tree-based argument $\{a, b\} \vdash p$. In $rmr(\mathcal{D}, \{r\})$, we cannot construct the argument anymore. If we make r defeasible instead of deleting it, by adding the new assumption d_r, we still can derive p as long as we extend the original set with d_r; i.e., we have $\{a, b, d_r\} \vdash p$ in $def(\mathcal{D}, (\{r\}, f))$ (this holds for an arbitrary contrary function f).*

We extend this idea to the general case and obtain the following result, showing that a form of Derivation Persistence (DP) and Conclusion Monotonicity (CM) hold for def.

Proposition 6.11 (Weak DP and CM). *For an ABAF $\mathcal{D} = (\mathcal{L}, \mathcal{R}, \mathcal{A}, \overline{})$, $R \subseteq \mathcal{R}$, and $f : R \to \mathcal{L}$. it holds that*

$$T \vdash_U p \text{ in } \mathcal{D} \text{ iff } T \cup \{d_r \mid r \in U \cap R\} \vdash_{U'} p \text{ in } def(\mathcal{D}, R, f),$$

$$U' = (U \setminus R) \cup \{head(r) \leftarrow body(r) \cup \{d_r\} \mid r \in U \cap R\}.$$

Moreover, the operator def satisfies Tractability (T) since it is a simple syntactical modification.

Fact 6.12 (Tractability). *For any ABAF $\mathcal{D} = (\mathcal{L}, \mathcal{A}, \mathcal{R}, \overline{})$, set $R \subseteq \mathcal{R}$ of rules, and function $f : R \to \mathcal{L}$, the ABAF $def(\mathcal{D}, (R, f))$ can be computed in polynomial time.*

Overall, the following desiderata are satisfied by def.

Summary 6.13. *The operator def satisfies **(CP)**, **(SC)**, and **(T)**; also, def satisfies weak versions of **(DP)** and **(CM)**; moreover, **(E)** is satisfied under mild conditions.*

6.2 Minimal Repairs

We turn to the investigation of minimal repairs. In a nutshell, we obtain similar results for def as in Section 5.

Theorem 6.14. *Deciding whether $R \subseteq \mathcal{R}$ is a minimal successful diagnosis w.r.t. def is i) DP-complete for $\sigma \in \{ad, co, pr, stb\}$ and ii) coNP-complete for gr.*

However, we can again compute a minimal gr-diagnosis efficiently; thus overall the situations are comparable.

Proposition 6.15. *There is a polynomial algorithm that, on input an ABAF $\mathcal{D}$, returns a minimal successful rule-defeasibility gr-diagnosis (R, f) or detects if no such exists.*

Let us note that in this setting there are various conceivable notions as to how to define a "minimal" diagnosis, apart from minimality w.r.t. the set of rules. We leave a thorough investigation of other minimality notions for future work.

6.3 On the Relation of Defeasibility and Deletion

We study the relation between rmr and def. As it turns out, both repairing notions are closely related. We show that each diagnosis w.r.t. rmr can be used to obtain a diagnosis w.r.t. def (for grounded semantics, we additionally require the existence of some factual atom p in $\mathcal{D}$ and $rmr(\mathcal{D}, R)$).

Proposition 6.16. *Let $\mathcal{D} = (\mathcal{L}, \mathcal{R}, \mathcal{A}, \overline{})$ be an ABAF.*

- *Let $\sigma \in \{ad, co, pr, stb\}$. If R is a diagnosis for $\mathcal{D}$ w.r.t. rmr then there is $f : R \to \mathcal{L}$ s.t. (R, f) is a successful diagnosis w.r.t. def.*
- *Let $\sigma = gr$. If R is a diagnosis for $\mathcal{D}$ w.r.t. rmr and there is $p \in Th_{\mathcal{D}}(\emptyset) \cap Th_{rmr(\mathcal{D},R)}(\emptyset)$ then there is $f : R \to \mathcal{L}$ s.t. (R, f) is a successful diagnosis w.r.t. def.*

7 Discussion

In this paper we studied the problem of repairing a semantical collapse in ABA, one of the most prominent structured argumentation formalisms. After noting that known results on repairing in AFs are not suitable for our setting, we developed genuine ABA repairing notions. We investigated approaches based on i) removing assumptions from the knowledge base, ii) deleting rules, and iii) making rules defeasible. To formalize how an intuitive repairing operator should behave, we developed several desiderata for our setting. We investigated the compliance of our operators with these desiderata, as well as further theoretical and computational properties. Moreover, we showed that all of them are capable of repairing any knowledge base under mild conditions. A noteworthy observation is that finding *minimal* diagnosis is computationally hard in all cases; but we identified gr repairs as a suitable starting point to reduce the search space for the other semantics.

7.1 Related Work

Reparing inconsistent knowledge bases has been considered in many non-monotonic reasoning formalisms. The approach which is most closely related to the present work is research on repairing in abstract argumentation (Baumann and Ulbricht 2019). Although AFs and ABA are closely related, this approach cannot be utilized as we discussed in Section 3. Moreover, the paper (Baumann and Ulbricht 2019) does not consider any desiderata and only the simple setting of removing arguments or attacks. Due the structural simplicity of AFs, however, the authors obtain stronger results compared to our study in the context of ABA. Their work also considers structural properties like hitting sets or splittings (Baumann 2011) which would be worthwhile studying in ABA as well.

The connection of ABAFs and logic programs (LPs) has been thoroughly studied (Caminada and Schulz 2017); in brief, each LP can be seen as instance of ABA where the default negated literals are the assumptions. Among other techniques, researchers proposed the removal of rules to repair inconsistent programs (Janota and Marques-Silva 2017; Sakama and Inoue 2003), similar in spirit to our rule-based repair operator rmr. Investigating repair operators for LPs inspired by our assumption- and defeasibility-based techniques could yield valuable insights.

Repairing inconsistent knowledge bases is closely linked to dynamics in argumentation; specifically, to enforcement (Baumann and Brewka 2010; Baumann 2012; Rapberger and Ulbricht 2023) where frameworks are manipulated to enforce acceptance of a given argument. In contrast to the present work, enforcement typically focuses on expanding knowledge bases, as e.g., studied by Cayrol, de

Saint-Cyr, and Lagasquie-Schiex (2010) and by Oikarinen and Woltran (2011) in the context of strong equivalence. Furthermore of relevance in this context is research on forgetting (Baumann and Berthold 2022; Berthold, Rapberger, and Ulbricht 2023) where framework modifications are used to remove information from a given knowledge base.

An orthogonal approach to repairing inconsistent instances in several knowledge representation and reasoning domains is to weakening semantics, if the considered semantics turns out to be too demanding. Historically, the motivation for exploring weaker versions of semantics has been, and continues to be, to enable reasoning when traditional evaluation methods fail. Prominent examples include the—by now—classical three-valued semantics (Dung 1995); more recent approaches are, e.g., semi-stable (Caminada 2006) or weakly admissible semantics (Baumann, Brewka, and Ulbricht 2022). The primary difference to repairing knowledge bases through modifications lies in identifying the issue within the specific instance itself, e.g., caused by a modeling error. Hence, repairing becomes crucial when traditional argumentation mechanisms fail.

7.2 Future Work

Our study lays theoretical foundations for repairing in ABA, with numerous conceivable future work directions. First of all, the development of further operators and desiderata would broaden our study. Also, exploring combined operators can provide greater flexibility. An inconsistent instance has usually several possible repairs; a crucial question is how to select the most appropriate one.

In order to ensure applicability, especially with regard to large ABA instances, efficient algorithms are important. In the context of ABA, a rich literature on efficient algorithms is already available (Lehtonen, Wallner, and Järvisalo 2021; Lehtonen, Wallner, and Järvisalo 2022; Lehtonen et al. 2023), so we are convinced that repairs can also be computed with (modifications of) the available techniques.

A crucial next step is to study repairs for more expressive fragments of ABA to be able to apply our repairing operators to non-flat ABA instances or when preferences are taken into account. Moreover, we believe that our approach can be extended to further structured argumentation formalisms like ASPIC$^+$ (Modgil and Prakken 2014), logic-based argumentation (Besnard and Hunter 2001; Amgoud 2014), or defeasible logic programming (García and Simari 2004).

Reparing with semi-abstract representations Our observation that results on AF repairing are not applicable is not very surprising, since this is often the case in dynamic argumentation scenarios (Rapberger and Ulbricht 2023; Prakken 2023). Several techniques have been studied to overcome these issues by extending the AF in a suitable way (Baumann, Rapberger, and Ulbricht 2023; Rapberger and Ulbricht 2023; Prakken 2023; Dvorák, Rapberger, and Woltran 2023; Rocha and Cozman 2022; Bernreiter et al. 2023). These *semi-abstract formalisms* have the advantage that they better connect the knowledge base and its corresponding abstract representation. An interesting future work direction would be to study whether such techniques are also applicable to our setting. For instance in (Rapberger and Ulbricht 2023) each argument is augmented with its *claim* and its *vulnerabilites* which, in the context of ABA, directly correspond to the conclusion resp. the assumption set which is necessary to entail the conclusion. It would thus be interesting to see whether such techniques might help us in applying AF research to repairing ABAFs.

Acknowledgements

This work was supported by the German Federal Ministry of Education and Research (BMBF, 01/S18026A-F) by funding the competence center for Big Data and AI "ScaDS.AI" Dresden/Leipzig.

References

Amgoud, L., and Vesic, S. 2009. Repairing preference-based argumentation frameworks. In Boutilier, C., ed., *IJCAI 2009, Proceedings*, 665–670.

Amgoud, L. 2014. Postulates for logic-based argumentation systems. *Int. J. Approx. Reason.* 55(9):2028–2048.

Baader, F.; Kriegel, F.; Nuradiansyah, A.; and Peñaloza, R. 2018. Making repairs in description logics more gentle. In *KR 2018, Proceedings*, 319–328. AAAI Press.

Baumann, R., and Berthold, M. 2022. Limits and possibilities of forgetting in abstract argumentation. In *Proc. IJCAI*, 2539–2545. ijcai.org.

Baumann, R., and Brewka, G. 2010. Expanding argumentation frameworks: Enforcing and monotonicity results. In *COMMA 2010, Proceedings*, volume 216 of *FAIA*, 75–86. IOS Press.

Baumann, R., and Ulbricht, M. 2019. If nothing is accepted - repairing argumentation frameworks. *J. Artif. Intell. Res.* 66:1099–1145.

Baumann, R.; Brewka, G.; and Ulbricht, M. 2022. Shedding new light on the foundations of abstract argumentation: Modularization and weak admissibility. *Artif. Intell.* 310:103742.

Baumann, R.; Rapberger, A.; and Ulbricht, M. 2023. Equivalence in argumentation frameworks with a claim-centric view: Classical results with novel ingredients. *J. Artif. Intell. Res.* 77:891–948.

Baumann, R. 2011. Splitting an argumentation framework. In *LPNMR 2011, Proceedings*, 40–53. Springer.

Baumann, R. 2012. What does it take to enforce an argument? Minimal change in abstract argumentation. In *ECAI 2012, Proceedings*, volume 242 of *FAIA*, 127–132. IOS Press.

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

Bernreiter, M.; Dvorák, W.; Rapberger, A.; and Woltran, S. 2023. The effect of preferences in abstract argumentation under a claim-centric view. In *AAAI 2023, Proceedings*, 6253–6261. AAAI Press.

Berthold, M.; Rapberger, A.; and Ulbricht, M. 2023. Forgetting aspects in assumption-based argumentation. In *KR 2023, Proceedings*, 86–96.

Besnard, P., and Hunter, A. 2001. A logic-based theory of deductive arguments. *Artif. Intell.* 128(1-2):203–235.

Brewka, G.; Thimm, M.; and Ulbricht, M. 2019. Strong inconsistency. *Artif. Intell.* 267:78–117.

Caminada, M., and Schulz, C. 2017. On the equivalence between assumption-based argumentation and logic programming. *J. Artif. Intell. Res.* 60:779–825.

Caminada, M. 2006. Semi-stable semantics. In *COMMA 2006, Proceedings*, volume 144 of *FAIA*, 121–130. IOS Press.

Cayrol, C.; de Saint-Cyr, F. D.; and Lagasquie-Schiex, M. 2010. Change in abstract argumentation frameworks: Adding an argument. *J. Artif. Intell. Res.* 38:49–84.

Craven, R.; Toni, F.; Cadar, C.; Hadad, A.; and Williams, M. 2012. Efficient argumentation for medical decision-making. In *KR 2012, Proceedings*. AAAI Press.

Čyras, K.; Fan, X.; Schulz, C.; and Toni, F. 2018. Assumption-based argumentation: Disputes, explanations, preferences. In *Handbook of Formal Argumentation*. College Publications. chapter 7, 365–408.

Cyras, K.; Oliveira, T.; Karamlou, A.; and Toni, F. 2021. Assumption-based argumentation with preferences and goals for patient-centric reasoning with interacting clinical guidelines. *Argument Comput.* 12(2):149–189.

Dung, P. M.; Thang, P. M.; and Hung, N. D. 2010. Modular argumentation for modelling legal doctrines of performance relief. *Argument Comput.* 1(1):47–69.

Dung, P. M. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artif. Intell.* 77(2):321–357.

Dvorák, W.; Rapberger, A.; and Woltran, S. 2023. A claim-centric perspective on abstract argumentation semantics: Claim-defeat, principles, and expressiveness. *Artif. Intell.* 324:104011.

Fan, X., and Toni, F. 2014. A general framework for sound assumption-based argumentation dialogues. *Artif. Intell.* 216:20–54.

Fan, X.; Liu, S.; Zhang, H.; Leung, C.; and Miao, C. 2016. Explained activity recognition with computational assumption-based argumentation. In *ECAI 2016, Proceedings*, volume 285 of *FAIA*, 1590–1591. IOS Press.

Gabbay, D.; Giacomin, M.; Simari, G. R.; and Thimm, M., eds. 2021. *Handbook of Formal Argumentation*, volume 2. College Publications.

García, A. J., and Simari, G. R. 2004. Defeasible logic programming: An argumentative approach. *Theory Pract. Log. Program.* 4(1-2):95–138.

Gebser, M.; Guziolowski, C.; Ivanchev, M.; Schaub, T.; Siegel, A.; Thiele, S.; and Veber, P. 2010. Repair and prediction (under inconsistency) in large biological networks with answer set programming. In *KR 2010, Proceedings*. AAAI Press.

Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; and Liu, T. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *CoRR* abs/2311.05232.

Janota, M., and Marques-Silva, J. 2017. On minimal corrections in ASP. In *RCRA 2017*, volume 2011 of *CEUR Workshop Proceedings*, 45–54. CEUR-WS.org.

König, M.; Rapberger, A.; and Ulbricht, M. 2022. Just a matter of perspective. In *Proc. COMMA*, volume 353 of *Frontiers in Artificial Intelligence and Applications*, 212–223. IOS Press.

Lehtonen, T.; Rapberger, A.; Ulbricht, M.; and Wallner, J. P. 2023. Argumentation frameworks induced by assumption-based argumentation: Relating size and complexity. In *Proc. KR*, 440–450.

Lehtonen, T.; Niskanen, A.; and Järvisalo, M. 2018. Sat-based approaches to adjusting, repairing, and computing largest extensions of argumentation frameworks. In *COMMA 2018, Proceedings*, volume 305 of *FAIA*, 193–204. IOS Press.

Lehtonen, T.; Wallner, J. P.; and Järvisalo, M. 2021. Declarative algorithms and complexity results for assumption-based argumentation. *J. Artif. Intell. Res.* 71:265–318.

Lehtonen, T.; Wallner, J. P.; and Järvisalo, M. 2022. Algorithms for reasoning in a default logic instantiation of assumption-based argumentation. In *COMMA 2022, Proceedings*, volume 353 of *FAIA*, 236–247. IOS Press.

Lembo, D.; Lenzerini, M.; Rosati, R.; Ruzzi, M.; and Savo, D. F. 2010. Inconsistency-tolerant semantics for description logics. In *RR 2010, Proceedings*, volume 6333 of *LNCS*, 103–117. Springer.

Merhej, E.; Schockaert, S.; and Cock, M. D. 2017. Repairing inconsistent answer set programs using rules of thumb: A gene regulatory networks case study. *Int. J. Approx. Reason.* 83:243–264.

Modgil, S., and Prakken, H. 2014. The *ASPIC*$^+$ framework for structured argumentation: a tutorial. *Argument & Computation* 5(1):31–62.

Nielsen, S. H., and Parsons, S. 2006. A generalization of Dung's abstract framework for argumentation: Arguing with sets of attacking arguments. In *ArgMAS 2006, Proceedings*, volume 4766 of *LNCS*, 54–73. Springer.

Niskanen, A., and Järvisalo, M. 2020. Smallest explanations and diagnoses of rejection in abstract argumentation. In *KR 2020, Proceedings*, 667–671.

Oikarinen, E., and Woltran, S. 2011. Characterizing strong equivalence for argumentation frameworks. *Artificial intelligence* 175(14):1985–2009.

Potyka, N., and Thimm, M. 2014. Consolidation of probabilistic knowledge bases by inconsistency minimization. In *ECAI 2014, Proceedings*, volume 263 of *Frontiers in Artificial Intelligence and Applications*, 729–734. IOS Press.

Prakken, H. 2023. Relating abstract and structured accounts of argumentation dynamics: the case of expansions. In *KR 2023, Proceedings*, 562–571.

Rago, A.; Russo, F.; Albini, E.; Toni, F.; and Baroni, P. 2023. Explaining classifiers' outputs with causal models and argumentation. *FLAP* 10(3):421–509.

Rapberger, A., and Ulbricht, M. 2023. On dynamics in structured argumentation formalisms. *J. Artif. Intell. Res.* 77:563–643.

Reiter, R. 1987. A theory of diagnosis from first principles. *Artif. Intell.* 32(1):57–95.

Rocha, V. H. N., and Cozman, F. G. 2022. Bipolar argumentation frameworks with explicit conclusions: Connecting argumentation and logic programming. In *NMR 2022, Proceedings*, volume 3197 of *CEUR Workshop Proceedings*, 49–60. CEUR-WS.org.

Sakama, C., and Inoue, K. 2003. An abductive framework for computing knowledge base updates. *Theory Pract. Log. Program.* 3(6):671–713.

van Harmelen, F.; Lifschitz, V.; and Porter, B. W., eds. 2008. *Handbook of Knowledge Representation*, volume 3 of *Foundations of Artificial Intelligence*. Elsevier.

Vassiliades, A.; Bassiliades, N.; and Patkos, T. 2021. Argumentation and explainable artificial intelligence: a survey. *The Knowledge Engineering Review* 36:e5.