Weak Admissibility for ABA via Abstract Set-Attacks

Lydia Blümel[1] , **Matthias König**[2] , **Markus Ulbricht**[3]
[1]Artificial Intelligence Group, University of Hagen
[2]TU Wien, Institute of Logic and Computation
[3]University of Leipzig, ScaDS.AI
lydia.bluemel@fernuni-hagen.de, matthias.koenig@tuwien.ac.at, mulbricht.informatik@uni-leipzig.de

Abstract

Is an argument acceptable if all potential counter-arguments are unacceptable themselves? In standard models of argumentation, the answer to this question is counter-intuitively not necessarily yes. However, based on the notion of weak admissibility, a family of semantics has been established where these unreasonable attacks do not successfully counter otherwise strong arguments. While in the abstract setting weak admissibility is well-understood, a similar issue arises in the context of structured argumentation formalisms like assumption based argumentation (ABA). It is well known that under standard argumentation semantics, ABA frameworks can be reduced to abstract argumentation frameworks (AFs), however, it turns out that in the case of weak admissibility this approach surprisingly fails. We instead propose to utilize a recently published instantiation technique utilizing collective attacks (SETAFs). We first define weak admissibility for SETAFs and study basic properties; afterwards, we push our proposal to the structured setting. We show that via our approach the characteristic properties of weak admissibility carry over to ABA, and thus establish a basis for further studies of these common scenarios also in ABA and related structured argumentation formalisms.

1 Introduction

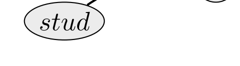
Formal argumentation constitutes a vibrant research area in AI, covering various aspects of knowledge representation, non-monotonic reasoning, and multi-agent systems (Baroni et al. 2018). It deals with computational models of arguments and argumentative reasoning workflows. Thereby, the goal is to determine reasonable viewpoints, i. e. jointly acceptable sets of arguments in an automated way. In research on argumentation, two main branches have emerged, namely *structured* and *abstract* argumentation.

Without any doubt the main booster for the development of the latter was Dung's seminal paper on abstract argumentation frameworks (AFs) (Dung 1995). In an abstract argumentation framework, arguments are viewed as atomic entities, and the conflicts between them are viewed as directed edges. Consequently, Dung obtains a representation as a directed graph $F = (A, R)$ formalizing the given debate. AFs have been thoroughly investigated ever since, covering various aspects such as evaluation techniques (Charwat et al. 2015), dynamic reasoning environments (Gabbay et al. 2021), computational aspects (Dvořák and Dunne 2018), and graph-theoretical properties (Dunne 2007).

In structured argumentation formalisms, a given knowledge base is evaluated by means of argumentative reasoning techniques: the incorporated, possibly conflicting, information gives rise to arguments and attacks among them, constructed in a specific way. This way, structured argumentation covers entire argumentative workflows (Baroni et al. 2018). A prominent example is assumption-based argumentation (ABA), which is well developed theoretically (Cyras et al. 2018) and finds applications in e.g. medical reasoning (Cyras et al. 2021) or planning (Fan 2018).

In both structured as well as abstract argumentation, a key concept is the notion of *admissibility*. Loosely speaking, admissibility formalizes that, in a given debate, an acceptable set S of arguments should (i) not contain internal conflicts and (ii) be able to refute arguments raised against S. However, as pointed out in several works (Dung 1995; Dondio and Longo 2019; Baumann, Brewka, and Ulbricht 2022), this requirement is sometimes too strong, especially in the presence of paradoxical arguments as in the following.

Example 1. *Suppose our agent participates in a debate about climate change. The first argument brought forward is that "Climate change is happening due to mankind emitting carbon dioxide." (yielding an assumption for climate change, cc). Another argument confirms this, stating that "According to numerous studies, climate change is happening." (yielding an assumption in favor of studies, $stud$). However, another participant counters this by arguing that "I read on social media that everything written on the internet is false." (yielding an assumption in favor of distrusting information, dis). If we distrust every information on the internet, this together with the fact that the studies on climate change are available online constitutes a collective attack from dis and $stud$ towards cc: if both of these assumptions are accepted, we have to disregard cc. On the other hand, dis is a self-attacking argument, because if everything on the internet is false, then also this same information found on social media. We obtain the following attack structure.*

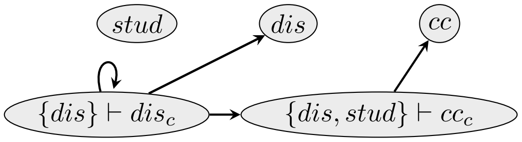


As we will see in Example 15, such scenarios can easily occur in ABAFs. Clearly, in this scenario we would like to disregard the self-attacking argument that represents distrusting reasonable information. However, no commonly agreed semantics for ABAFs can handle this in a satisfactory way.

In the context of AFs, a quite successful solution to this issue has been proposed. Based on the notion of so-called *weak admissibility*, Baumann, Brewka, and Ulbricht (2020b) propose several semantics that are able to neglect paradoxical arguments such as dis in a reasonable way, while keeping reasonable ones acceptable. Intuitively, weak admissibility does not require the set S to refute *every* argument; it suffices to counter-attack *reasonable* viewpoints.

Weak admissibility is well understood (Baumann, Brewka, and Ulbricht 2022; Dauphin, Rienstra, and van der Torre 2021; Blümel and Ulbricht 2022a); and ABAF and AFs are closely related (König, Rapberger, and Ulbricht 2022; Caminada et al. 2024). Consequently, a natural idea would be to try and benefit from weak admissibility for AFs out of the box: we instantiate the given knowledge base D and evaluate the argumentation graph F_D according to the weak semantics family. This direct approach, however, does not work as we demonstrate next.

Example 2. *Instantiating the aforementioned ABAF as an AF yields the following argumentation graph (we will formalize this in Example 15 below):*



Instead of disregarding the self-attacking dis, weak admissibility disregards the argument "$c(dis) \leftarrow dis$" which explicates that this assumption is self-attacking. Consequently, dis and $stud$ are accepted whereas cc is not. This is surprisingly far away from what we want to achieve; after all, weak admissibility handles (abstract) self-attackers quite well.

Given that the common AF instantiation does not help here, we need another strategy to obtain a reasonable notion of "weak" ABA semantics. An abstract formalism which is much closer in spirit to ABA compared to AFs are *argumentation frameworks with collective attacks* (SETAFs), introduced by Nielsen and Parsons (2006). SETAFs and their semantics are well-studied (Flouris and Bikakis 2019; Bikakis et al. 2021), their relation to ABA is understood (Caminada et al. 2024; Dimopoulos et al. 2024), computational aspects have been investigated (Dvořák, Greßler, and Woltran 2018; Dvořák, König, and Woltran 2022a;2022b), and even a suitable notion of a *reduct* is already available (Dvořák et al. 2024), which is a core concept towards defining the weak semantics family.

Our strategy is thus as follows: we first define weak admissibility for SETAFs and study its properties on an abstract level. Then, we showcase how to transfer this approach to structured argumentation by translating these semantics to ABA. In order to assess our approach, we propose desiderata as to how a "good" notion of weak admissibility for ABA should behave.

More specifically, our main contributions are:

- We develop desiderata for a "weak" semantics family in the context of ABA and demonstrate why the standard ABA instantiation is insufficient. Section 3
- We develop weak admissibility for SETAFs. We study formal properties and show that many properties known to hold for AFs generalize well to this setting. Section 4
- We propose weak admissibility for ABA by means of the SETAF SF_D induced by some knowledge base D. We show that these novel ABA semantics indeed satisfy the previously formulated desiderata. Section 5

2 Background

2.1 Abstract Argumentation

We briefly recall AFs and SETAFs as our utilized notions of abstract argumentation. Since SETAFs generalize AFs, we first introduce SETAFs and thereby subsume the AF notions.

Definition 3 (SETAF). *A SETAF $SF = (A, R)$ is a pair, consisting of a finite set of arguments A and an attack relation $R \subseteq 2^A \times A$ that contains attacks from a set of arguments towards a single argument.*

Definition 4. *A set $S \subseteq A$ is called* conflicting *if for some $(T, h) \in R$ holds $T \cup \{h\} \subseteq S$, otherwise it is* conflict-free. *We denote the set of conflict-free sets by $cf(SF)$. We say S defends an argument $a \in A$ if for each $(T, a) \in R$ holds $(S', t) \in R$ for some $S' \subseteq S$, $t \in T$. S defends $U \subseteq A$ if S defends each $u \in U$.*

We set $E_R^+ = \{h \in A \mid \exists (T, h) \in R : T \subseteq E\}$ and $E_R^\oplus = E \cup E_R^+$. We recall grounded, admissible, complete, and preferred semantics (*grd*, *adm*, *com*, and *pref* respectively (Nielsen and Parsons 2006; Dvořák, Greßler, and Woltran 2018; Flouris and Bikakis 2019)).

Definition 5. *Given a SETAF $SF = (A, R)$ and a conflict-free set $S \in cf(SF)$. Then,*

- *$S \in adm(SF)$, if S defends itself in SF,*
- *$S \in com(SF)$, if $S \in adm(SF)$ and $a \in S$ for all $a \in A$ defended by S,*
- *$S \in grd(SF)$, if S is $\subseteq$-minimal in $com(SF)$, and*
- *$S \in pref(SF)$, if S is $\subseteq$-maximal in $adm(SF)$.*

We introduce AFs (Dung 1995) as a special case of SETAFs; the terminology and semantics carry over.

Definition 6 (AF). *An AF $F = (A, R)$ is a pair, consisting of a finite set of arguments A and an attack relation $R \subseteq A \times A$ that contains attacks from a single argument towards a single argument.*

2.2 Weak Admissibility

As we generalize weak admissibility and related notions to SETAFs in this paper, we now briefly recall the relevant notions on AFs (Baumann, Brewka, and Ulbricht 2020b). We start with the E-reduct as the underlying foundation.

Definition 7. *Given an AF $F = (A, R)$ and $E \subseteq A$, the E-*reduct *of F is the AF $F^E = (A', R')$, with*

$$A' = A \setminus E_R^\oplus \qquad R' = R \cap (A' \times A')$$

We next recall weak admissibility in the context of AFs (Baumann, Brewka, and Ulbricht 2020b).

Definition 8. *Let $F = (A, R)$ be an AF, let $E \subseteq A$ be a set of arguments, and $F^E = (A', R')$ its E-reduct. E is called* weakly admissible *in F ($E \in adm^w(F)$) iff*

1. *$E \in cf(F)$, and*
2. *for each $(t, h) \in R$ with $h \in E$ it holds $t \notin \bigcup adm^w(F^E)$.*

We also recall weak defense and the induced semantics.

Definition 9. *Let $F = (A, R)$ be an AF. Given two sets $E, X \subseteq A$, E weakly defends X iff for any attacker y of X,*

1. *E attacks y, or*
2. *$y \notin E$, $y \notin \bigcup adm^w\left(SF^E\right)$ and $X \subseteq X' \in adm^w(SF)$.*

Definition 10. *Let $F = (A, R)$ be an AF. A set $E \subseteq A$ is called weakly complete in F ($E \in com^w(SF)$) iff $E \in adm^w(SF)$ and for any superset $X \supseteq E$ weakly defended by E, we have that $X = E$. Moreover, E is weakly grounded iff it is minimal in $com^w(F)$ as well as* weakly preferred *iff it is maximal in $adm^w(SF)$.*

Definition 11. *Let $SF = (A, R)$ be a SETAF and let $S \subseteq A$ be a set of arguments. We define the* projection *to S as the SETAF $SF{\downarrow_S} = (S, R \cap (2^S \times S))$.*[1]

2.3 Assumption-Based Argumentation

We recall underlying definitions of ABA (Cyras et al. 2018).

Definition 12 (ABAF)**.** *An* ABA framework (ABAF) *is a tuple $D = (\mathcal{L}, \mathcal{R}, \mathcal{A}, c)$, where $\mathcal{L}$ is a set of atoms, $\mathcal{R}$ a set of inference rules over $\mathcal{L}$ of the form $a_0 \leftarrow a_1, ..., a_n$ with $a_i \in \mathcal{L}$ for $0 \leq i \leq n$, $\mathcal{A} \subseteq \mathcal{L}$ a set of assumptions and $c : \mathcal{A} \to \mathcal{L}$ a contrary function.*

We focus on finite, flat ABA frameworks, i. e. where $\mathcal{L}, \mathcal{R}, \mathcal{A}$ are finite sets and for each $r \in \mathcal{R}$ it holds that $head(r) \notin \mathcal{A}$. We say that an atom $p \in \mathcal{L}$ is *derivable* from assumptions $S \subseteq \mathcal{A}$ and rules $R \subseteq \mathcal{R}$, if p can be derived from the set S of assumptions and the rules in R in the natural way. That is, 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 $body(r) = \emptyset$. We denote such a derivation by $S \vdash_{\mathtt{t}} p$ and drop the subscript $\mathtt{t}$ whenever the underlying tree is unimportant. The set of all atoms which are derivable from S is denoted by $th_D(S)$.

Let $S, T \subseteq \mathcal{A}$ for a given ABAF $D = (\mathcal{L}, \mathcal{R}, \mathcal{A}, c)$. Set S *attacks* T if there are $S' \subseteq S$ and $a \in T$ s. t. $S' \vdash c(a)$. A set of assumptions $S \subseteq \mathcal{A}$ is *conflict-free*, $S \in cf(D)$, iff it does not attack itself. We say $S \subseteq \mathcal{A}$ *defends* some assumption $a \in \mathcal{A}$ whenever $T \vdash c(a)$ implies $S \vdash c(t)$ for some $t \in T$. By Γ_D we denote the *characteristic function* of D, i. e. we let $\Gamma_D(S) = \{a \in \mathcal{A} \mid S \text{ defends } a\}$. We drop the subscript D whenever it is clear from the context.

Definition 13. *Let $S \in cf(D)$ for an ABAF $D = (\mathcal{L}, \mathcal{R}, \mathcal{A}, c)$. Then*

- *$S \in adm(D)$ iff $\Gamma(S) \subseteq S$,*
- *$S \in com(D)$ iff $\Gamma(S) = S$,*
- *$S \in grd(D)$ iff S is $\subseteq$-minimal in $com(D)$,*
- *$S \in pref(D)$ iff S is $\subseteq$-maximal in $adm(D)$.*

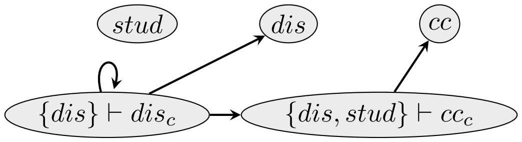
AFs and ABAFs are closely related. From an ABAF D, we can construct a semantics-preserving AF F_D as follows.

Definition 14. *Let $D = (\mathcal{L}, \mathcal{R}, \mathcal{A}, c)$ be an ABAF. The associated AF $F_D = (A_D, R_D)$ is given by*

$$A_D = \{S \vdash_{\mathtt{t}} a \mid S \vdash_{\mathtt{t}} a \text{ is a tree derivation in } D\}$$
$$R_D = \{(S \vdash_t a, T \vdash_t b) \mid a \in c(T)\}.$$

Throughout our examples, we depict tree-based arguments $S \vdash_{\mathtt{t}} p$ as tuples (S, p) which is known to be an equivalent representation (see e.g. (Lehtonen et al. 2023)).

Example 15. *Let us model the introductory Example 1 as an ABAF $D = (\mathcal{L}, \mathcal{R}, \mathcal{A}, c)$ with the sets $\mathcal{A} = \{stud, dis, cc\}$, $\mathcal{L} = \mathcal{A} \cup \{a_c \mid a \in \mathcal{A}\}$, contraries $c(a) = a_c$ for each assumption and rules $\mathcal{R} = \{(dis_c \leftarrow dis), (cc_c \leftarrow dis, stud)\}$. Indeed, as we already hinted at in Example 2, the associated AF F_D is given as follows (each assumption entails a canonical argument for itself and both rules induces arguments with out-going attacks).*



Indeed, the argument in favor of $\{dis, stud\} \vdash cc_c$ is weakly admissible as it is only attacked by some self-attacker. Consequently, $\{cc\} \notin adm^w(F_D)$ which is not desired here.

3 ABA and Weak Admissibility

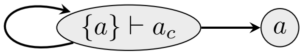
The introduction of weak admissibility was motivated by the desire to accept arguments which cannot defend themselves against non-serious attackers, for instance a self-attacker or a member of an isolated odd-length-cycle. These problems occur in ABAFs as well. Rules like $c(a) \leftarrow a$, which has the contrary of one of its own assumptions as its head, can occur and induce arguments, that are not acceptable themselves, but still capable of attacking other arguments which were derived from a consistent subset of rules and would otherwise be accepted. Instead of coming up with an artificial version of "weak admissibility" for ABA, let us examine the behavior we expect.

Example 16. *Consider the following simple ABAF $D = (\mathcal{L}, \mathcal{R}, \mathcal{A}, c)$ with $\mathcal{L} = \{a, a_c\}$, $\mathcal{A} = \{a\}$, contrary $c(a) = a_c$, and the rule $\mathcal{R} = \{a_c \leftarrow a\}$.*

It is easy to spot that a is a paradoxical assumption as it derives its own contrary. We would therefore expect no argument to be weakly admissible, i. e. intuitively we strive to achieve $adm^w(D) = \{\emptyset\}$ in this example. Indeed, if we remove the assumption causing the issue, a, we are left with an empty ABAF, as no rule is left.

The associated AF F_D, however, is given as follows (we depict the canonical assumption argument just by a).

[1] Note that our notion of projection differs from the one in (Dvořák et al. 2024): in our case, if we remove an argument a, we remove every attack $(T \cup \{a\}, h)$ entirely, whereas in (Dvořák et al. 2024) parts of the attack remain (akin to the reduct).



Consider the argument representing the assumption a. Its only attacker is the argument $\{a\} \vdash a_c$ which in turn is a self-attacker. By the nature of weak admissibility, this self-attacker can be disregarded and consequently, $\{a\} \in adm^w(F_D)$. Thus, while a is the culprit *of the irrational argument (via the rule $a_c \leftarrow a$), it is treated like the* victim *in F_D and rendered acceptable. This is contrary to the intended intuition.*

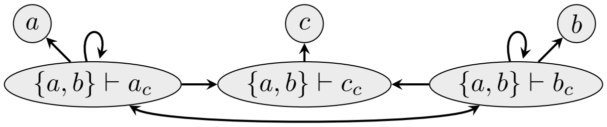
Note that in Example 16 the classic admissible semantics produces the desired result, i. e. both the assumption a and the derived argument $\{a\} \vdash a_c$ are rejected. As the following example shows, applying admissible semantics does not always yield the desired outcome.

Example 17. *Consider the ABAF $D = (\mathcal{L}, \mathcal{R}, \mathcal{A}, c)$ with $\mathcal{L} = \{a, b, c, a_c, b_c, c_c\}$, $\mathcal{A} = \{a, b\}$, contraries $c(a) = a_c$ as well as $c(b) = b_c$, $c(c) = c_c$, and the following rules $\mathcal{R}$:*

$$a_c \leftarrow a, b \qquad b_c \leftarrow a, b \qquad c_c \leftarrow a, b$$

Here we have that c is only attacked by the paradoxical assumption set $\{a, b\}$, so we expect $\{c\} \in adm^w(D)$ to hold.

However, the associated AF F_D is given as follows.



Here c is attacked by the argument $\{a, b\} \vdash c_c$. Since the former is only attacked by two self-attackers, $\{a, b\} \vdash c_c$ is weakly admissible in F_D. It follows that $\{c\} \notin adm^w(F_D)$.

Consequently, the problematic behavior observed in Example 16 now causes additional harm: the argument $\{a, b\} \vdash c_c$ stemming from the paradoxical assumption set $\{a, b\}$ is rendered accepted, while c is rejected. Again, this is contrary to our intuition. The situation is no better under classic admissibility: here, c is also rejected.

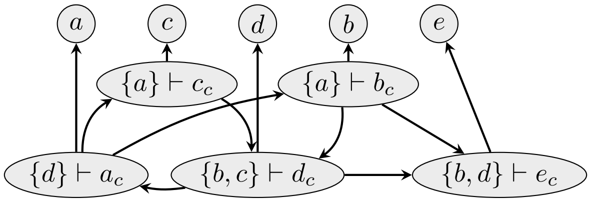
Both classic admissible semantics *adm* and weakly admissible semantics adm^w therefore fail to treat the cause and the consequences of inconsistent arguments resp. assumption sets in ABA Frameworks, when applied to the instantiated AF F_D. It is evident the cause of the undesired behavior lies in the representation of D. The same issues arise in the presence of odd-length cycles, ruling out exception-handling of self-attackers as a solution.

Example 18. *Consider the following ABAF $D = (\mathcal{L}, \mathcal{R}, \mathcal{A}, c)$ with $\mathcal{L} = \{a, b, c, d, e, a_c, b_c, c_c, d_c, c(e)\}$, $\mathcal{A} = \{a, b, c, d, e\}$, the usual contraries, and rules $\mathcal{R}$:*

$$b_c \leftarrow a \quad c_c \leftarrow a \quad d_c \leftarrow b, c \quad e_c \leftarrow b, d \quad a_c \leftarrow d$$

At first glance, this situation is more involved compared to Example 16, but it is actually quite similar. Again, we have a set of assumptions that contradict themselves, here indirectly through an odd-length cycle of attacks. We also have an innocent bystander, the assumption e is not involved in the odd cycle of attacks and therefore a promising candidate for acceptance, since its only attackers b, d are paradoxical as they are indirectly attacking themselves.

However, in the AF F_D, we again have the problem that e can neither be accepted under adm nor adm^w.



The argument e is not acceptable under classic admissibility, because it has an attacker $\{b, d\} \vdash e$, and the two possible defenders $\{a\} \vdash b$ and $\{b, c\} \vdash d$ are stuck in a cycle of length 3, unable to defend themselves. Furthermore, since both its attackers are part of an unattacked odd cycle, the argument $\{b, d\} \vdash e$ is weakly admissible, so e, which is attacked by it and has no defender, is not. Again, the attacking set of assumptions $\{b, d\}$ is the actual problem, but not punished by weak admissibility in the AF-instantiation. Note that $\{b, d\}$ is not attacking itself directly, but is not acceptable due to attacking its defenders.

Desiderata The main issue raised by these examples is that in ABA a set of assumptions might not be paradoxical to begin with, but only contradicts itself when instantiated with rules that induce direct or indirect self-contradictions. A weak admissibility semantics for ABA has to account for this difference from AFs. Bearing that in mind, we formulate desiderata to capture the desired behavior of the weak admissibility semantics family for ABA.

First of all, weak admissibility is supposed to be a weaker version of Dungs admissibility. Consequently, we expect the following to hold for a natural definition of adm^w.

(L) Liberalization: It holds that $\bigcup adm(D) \subseteq \bigcup adm^w(D)$.

In abstract argumentation, Liberalization has been utilized to describe inter-semantical relations before (Blümel and Ulbricht 2022a). Note that the above formalization emphasizes the accepted arguments, and does not make any direct statement about the individual extensions.

Along these lines, but from a slightly more technical point of view, we recall Dung's fundamental lemma stating that, if $E \in adm(F)$ and E defends a, then $E \cup \{a\}$ is admissible as well. Since weak admissibility is supposed to be a liberalization of *adm*, we expect a similar behavior here:

(F) Fundamental Lemma: If $S \in adm^w(D)$ defends a, then $D \cup \{a\} \in adm^w(D)$.

The main goal driving weak admissibility is to neglect the impact self-attacking arguments have on the remaining knowledge base. In fact, we want to be able to ignore an assumption that derives its own contrary completely. To formalize this desideratum we have to talk about ABAFs induced by subsets of the assumptions. We define the projection of an ABAF to a subset of its assumptions as follows:

Definition 19. *Let $D = (\mathcal{L}, \mathcal{R}, \mathcal{A}, c)$ be an ABAF, $T \subseteq \mathcal{A}$. The projection $D{\downarrow_T} = (\mathcal{L}', \mathcal{R}', \mathcal{A}', c')$ of D to T is defined* *by* $\mathcal{L}' = (\mathcal{L} \setminus (\mathcal{A} \setminus T)) \cup \{x_c\}$*,* $\mathcal{A}' = T$*, the rules are* $\mathcal{R}' = \mathcal{R} \setminus \{r \mid body(r) \cap (A \setminus T) \neq \emptyset\}$*, and for all* $a \in T$

$$c'(a) = \begin{cases} c(a) & if\ c(a) \in \mathcal{L}' \\ x_c & otherwise \end{cases}$$

Recall that we study flat ABAFs here, so even if an assumption is the contrary of another assumption it never occurs in the head of a rule and is therefore never derived from the rule set of the projection. Therefore replacing it by the placeholder x_c does not cause any harm. We are now ready to define our desideratum for paradoxical assumptions.

(P) Paradoxical Assumptions: If $\{a\} \vdash c(a)$, then it holds that $adm^w(D) = adm^w(D \downarrow_{\mathcal{A}\setminus\{a\}})$.

If our notion of weak admissibility adheres to this principle, then we have formally established the neglection of self-attacking assumptions, as it is the case for *adm*^{w} in AFs, where self-attackers can simply be removed under *adm*^{w}.

A conceptual difference to AFs is that in ABA, assumptions sometimes collectively attack others, e. g. when a rule requires multiple assumptions in order to derive a contrary. Hence, a set of assumptions can attack its own members with a collective attack (see Example 17). For this type of self-attack no counterpart exists in AFs. The behavior we expect here is that a rule with a self-attacking set of assumptions in its body can be ignored. We have to be careful, though, for we do not want to ignore a rule that is involved in the attack of the paradoxical set on itself nor can we ignore a rule that defends any of the involved assumptions. For instance, in Example17 the attack on a should be maintained, while the attack on c can be ignored. This leads to the following notion of paradoxical rules.

Definition 20. *A rule* r *of the form* $r = (a \leftarrow S)$ *is paradoxical iff for every* $s \in S$ *there is an* $S' \subseteq S$ *s.t.*

- $S' \vdash c(s)$
- $head(r) \neq c(s')$ *for any* $s' \in S'$
- *there exists a derivation tree* t *such that* $S' \vdash_t c(s)$ *and* r *does not occur in* t

(PR) Paradoxical Rules: Removing a paradoxical rule r does not alter the models of D, i. e. $adm^w(D) = adm^w(D')$ where $D' = (\mathcal{L}, \mathcal{R} \setminus \{r\}, \mathcal{A}, c)$.

The treatment of problematic assumptions should not be limited to self-attackers. As Example 18 shows, the presence of odd cycles of attacks can lead to the exclusion of unrelated assumptions. Weak admissibility should allow for assumptions, which are not actively involved in deadlocks of this type, to be accepted. To formalize this we define the counterpart of an unattacked set of an AF in ABA.

Definition 21. *Let* $D = (\mathcal{L}, \mathcal{R}, \mathcal{A}, c)$ *be an ABAF. A set of rules* Q *is* independent *iff for any* $r \in \mathcal{R} \setminus Q$ *it holds that if* $head(r) = c(a)$*, then* $a \notin \bigcup_{q \in Q} body(q)$*.*

That is, no attack on the assumptions used in Q can be derived from the rules in $\mathcal{R} \setminus Q$. Hence an independent set of rules induces a part of an ABAF that is not attacked by any argument outside of it. The ABAF $D|_Q = (\mathcal{L}, Q, \mathcal{A}, c)$ is called the *restriction* of D to Q.

Now suppose $D|_Q$ does not contain any weakly admissible assumption. Then we expect that $D|_Q$ is no threat to the remaining part of the ABAF since the entire sub-framework is paradoxical. Now, a rule $r \notin Q$ with $body(r) \subseteq \bigcup_{q \in Q} body(q)$ relies on these paradoxical assumptions in Q in order to fire, while it cannot impact the (non-accepted) assumptions in Q. We expect such rules to be redundant.

(I) Independence: Let Q be an independent set of rules in an ABAF $D = (\mathcal{L}, \mathcal{R}, \mathcal{A}, c)$. If $adm^w(D|_Q) = \{\emptyset\}$, then removing a rule $r \notin Q$ with $body(r) \subseteq \bigcup_{q \in Q} body(q)$ does not alter the models of D, i. e. $adm^w(D) = adm^w(D')$ where $D' = (\mathcal{L}, \mathcal{R} \setminus \{r\}, \mathcal{A}, c)$.

Moreover, we strive to define a well-behaved semantics family. Consequently, we expect the usual relations:

(SR) Semantics Relation: For any ABAF $D = (\mathcal{L}, \mathcal{R}, \mathcal{A}, c)$

1. $adm^w(D), com^w(D), pref^w(D), grd^w(D) \neq \emptyset$,
2. $com^w(D) \subseteq adm^w(D)$
3. $pref^w(D) = \{S \subseteq \mathcal{A} \mid S \text{ is maximal in } com^w(D)\}$

To ensure the described behavior, adapting weak admissibility for ABAFs is a promising approach. However, as the examples show, one cannot use the weakly admissible semantics as-is on the AF associated with an ABAF. The remainder of the paper therefore investigates a more sophisticated adaption of weak admissibility for ABA. We first elevate the notion of weak admissibility to abstract SETAFs in the next section and then apply it to a SETAF-instantiation of ABAFs in Section 5.

4 SETAFs and Weak Admissibility

In recent studies, SETAFs have demonstrated multiple times that they are much closer to the behavior of ABA than the classical AF instantiation, and have thus more potential to be the suitable tool for our purpose (König, Rapberger, and Ulbricht 2022; Caminada et al. 2024; Dimopoulos et al. 2024). In this section, we introduce weak admissibility for SETAFs, guided by the motivation for the proposal in the AF case (Baumann, Brewka, and Ulbricht 2020b). Then in Section 5, due to the close correspondence between ABA and SETAFs, we revisit weak admissibility for ABA, with SETAFs as the means to instantiate the given ABAF.

4.1 Basic Definitions

The core concept underlying weak admissibility for AFs is the E-reduct (Baumann, Brewka, and Ulbricht 2020a). Towards weak admissibility for SETAFs, we first briefly recall the SETAF version of the E-reduct (Dvořák et al. 2024).

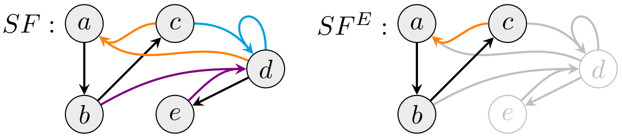
Definition 22. *Given a SETAF* $SF = (A, R)$ *and* $E \subseteq A$*, the* E-reduct *of* SF *is the SETAF* $SF^E = (A', R')$*, with*

$$\begin{aligned} A' &= A \setminus E_R^{\oplus} \\ R' &= \{(T \setminus E, h) \mid (T, h) \in R,\ T \cap E_R^+ = \emptyset, \\ &\quad T \not\subseteq E,\ h \in A'\} \end{aligned}$$

Dvořák et al. (2024) argue that the SETAF-reduct indeed generalizes the AF-reduct, in that for each SETAF $SF = (A, R)$ s.t. for each $(T, h) \in R$ it holds $|T| = 1$ (i.e., SF effectively is equivalent to an AF) the two reduct notions coincide. Moreover, the SETAF-reduct captures the same intuition by setting the arguments in E to true, the arguments in E^+ to false, and leaving the remaining ones as they are.

We recall the characteristic feature that the E-reduct w.r.t. a SETAF SF will sometimes contain "parts" of original attacks, and reiterate the underlying intuition.

Example 23. *Consider the SETAF $SF = (A, R)$ (left) and the reduct SF^E w.r.t. $E = \{d\}$ (right).*



While the attack $(\{b, e\}, d)$ is removed entirely since $e \in E_R^+$, for the attack $(\{c, d\}, a)$ we retain the attack $(\{c\}, a)$. Intuitively, since the reduct w.r.t. E simulates the remaining framework after accepting E, we can remove each attack (T, h) where $T \cap E^+ \neq \emptyset$, as the remaining arguments are defended against these attacks. Regarding d's incoming attack $(\{c, d\}, d)$, since d is accepted (in E), additionally accepting c leads to a rejection of the attacked argument (in this case, d itself) whereas defeating c renders d acceptable.

The E-reduct gives us the tools to generalize weak admissibility, originally due to Baumann, Brewka, and Ulbricht (2020b), to SETAFs. Note that the definition is recursive, but well-defined as in each recursion step the reduct contains fewer arguments and we deal only with finite SETAFs.

Definition 24. *Let $SF = (A, R)$ be a SETAF, let $E \subseteq A$ be a set of arguments, and $SF^E = (A', R')$ its E-reduct. Then E is called* weakly admissible *in SF ($E \in adm^w(SF)$) iff*

1. *$E \in cf(SF)$, and*
2. *for each $(T, h) \in R$ with $h \in E$, and $T \cap E_R^+ = \emptyset$ it holds $\nexists E' \in adm^w(SF^E)$ s.t. $T \cap A' \subseteq E'$.*

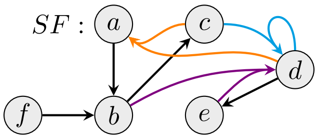
Let us head directly to some illustrative examples.

Example 25. *Recall the SETAF SF from Example 23. We show that $E = \{d\}$ is weakly admissible. We have already computed the reduct SF^E, so let us check the recursive definition: Since the attack $(\{c, d\}, d)$ is collective, $E \in cf(SF)$ holds. Now we consider the two in-coming attacks:*

- *For $(\{b, e\}, d)$ we have that $T = \{b, e\}$ does not satisfy $T \cap E_R^+ = \emptyset$ so this attack can be disregarded (intuitively, E defends itself against $(\{b, e\}, d)$).*
- *For $(\{c, d\}, d)$ we have that $T = \{c, d\}$ satisfies $T \cap E_R^+ = \emptyset$. We therefore need to check whether there is some $E' \in adm^w(SF^E)$ containing $T \cap A' = \{c\}$. However, it can be checked that an isolated odd-length cycle has no weakly admissible argument, so no such E' exists.*

We conclude $E \in adm^w(SF)$.

Example 26. *Now let us modify the previous example by disrupting the odd cycle:*



We note that $E = \{c, f\} \in adm(SF)$. Consequently, in the reduct SF^E, there is no attack towards E (an admissible set counters every attack). Therefore, $E \in adm^w(SF)$ as well; indeed, this observation generalizes as we see in Proposition 30 below. Utilizing this observation also leads to $E' = \{d\} \notin adm^w(SF)$ since this time, the paradoxical odd-length cycle is disrupted and c, attacking d in its reduct, can be accepted (together with f).

In order to handle more involved examples as well, we made several design choices in generalizing adm^w. Let us next examine these choices by contrasting weak admissibility on AFs and SETAFs, and argue why the alternative choices do not properly capture the intuition of weak admissibility. While the first conditions of Definition 8 and Definition 24 coincide, we see the following differences for the second conditions:

(a) Instead of comparing the set T to the set of accepted arguments over all weakly admissible extensions in SF^E (i.e., $\bigcup adm^w(SF^E)$), we require T *as a whole* not to appear in each weakly admissible extension of SF^E,

(b) we only consider attacks where $T \cap E_R^+ = \emptyset$, and

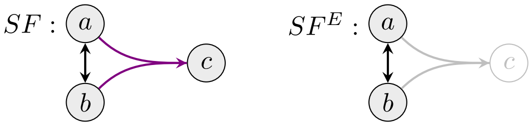
(c) we specify $T \cap A'$ for the comparison to E'.

Towards (a), first note that condition (2) of Definition 8 for AF weak admissibility can be equivalently reformulated as

2. for each $(t, h) \in R$ with $h \in E$ it holds $\nexists E' \in adm^w(F^E)$ s.t. $t \in E'$.

It becomes apparent that our definition of weak admissibility for SETAFs indeed generalizes its AF-counterpart in this regard. To argue for the need of this reformulation we provide the following illustrating example.

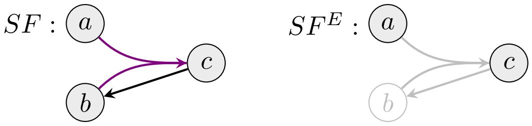
Example 27. *We illustrate the need of difference (a) from above: consider the following SETAF SF with $E = \{c\}$.*



Since we cannot accept the conflicting set $\{a, b\}$, c cannot be defeated by the attack. Hence, we would expect the set $\{c\}$ to be weakly admissible in SF. However, we have $adm^w(SF^E) = \{\{a\}, \{b\}\}$, i.e., $\{a, b\} \subseteq \bigcup adm^w(SF^E)$. Consequently, we insist that a single extension E' containing both arguments must exist; and we indeed observe that there is no $E' \in adm^w(SF^E)$ s.t. $\{a, b\} \subseteq E'$, as desired.

The difference (b) is explained entirely by the workings of the SETAF-reduct, as the next example illustrates.

Example 28. *Consider the SETAF SF with $E = \{c\}$.*



Clearly, $\{c\}$ is weakly admissible (and, in fact, also classically admissible). We know from Example 23 why we delete the entire attack $(\{a,b\},c)$ when we calculate the reduct w.r.t. $\{c\}$ and do not retain a partial attack $(\{a\},c)$: this is because $b \in \{c\}^+_R$, which means c is defended against the attack. Hence, we also do not need to consider $(\{a\},c)$ for weak admissibility of $\{c\}$ in this case. Finally, note that in the AF case we always trivially apply this restriction, as arguments a with $a \in E^+_R$ are excluded from the E-reduct.

Similarly, (c) is entirely due to the SETAF reduct. In this case we specify the intuitive requirement that for checking the "relevant" arguments in the reduct, all of the considered arguments actually appear in the reduct. In particular, no argument outside A' could appear in any $E' \in adm^w(SF^E)$.

4.2 Basic Properties

Now we have discussed and justified the definition of weak admissibility for SETAFs, and at the same time we have argued that indeed our notion properly generalizes its AF-counterpart. Let us next establish that the core properties of the semantics carry over from AFs to SETAFs. First of all, the SETAF version of weak admissibility coincides with the classical one whenever SF is an AF.

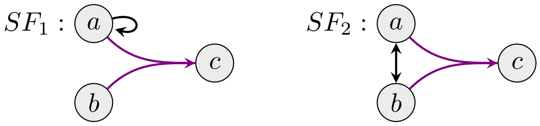
Proposition 29. *Let $SF = (A,R)$ be a SETAF s.t. $|T| = 1$ for each $(T,h) \in R$. Let F be the AF induced by SF. Then $adm^w(SF)$ (via Definition 24) and $adm^w(F)$ (via Definition 8) coincide.*

Moreover, adm^w generalizes classical admissibility.

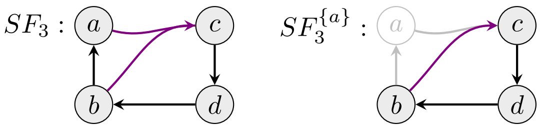
Proposition 30. *For all SETAFs $SF = (A,R)$ it holds $adm(SF) \subseteq adm^w(SF)$.*

Let us now investigate the feature of weak admissibility that self-attacking arguments can safely be removed from an argumentation framework without altering the weakly admissible sets (see Baumann, Brewka, and Ulbricht 2020b). We will see that in our setting, we can remove even more irrelevant parts of the framework. Let us investigate the desired property in more detail via the following example.

Example 31. *Consider the following SETAFs SF_1 and SF_2.*



We are again interested in $E = \{c\}$. In both SF_1 and SF_2 we cannot have the conflicting set $\{a,b\}$ in any weakly admissible extension. Hence, as expected we get that $E \in adm^w(SF_1)$ and $E \in adm^w(SF_2)$ (cf. Example 27). Even though SF_2 contains no self-attacks, the two cases are very similar: $(\{a,b\},c)$ cannot have any effect on the weakly admissible extensions, since the arguments in its tail $\{a,b\}$ attack each other. Attacks with conflicting tails are referred to as "inactive" attacks (Dvořák, Rapberger, and Woltran 2020). However, we cannot remove all inactive attacks:



While in SF_3 the set $\{a\}$ is weakly admissible (since $SF_3^{\{a\}}$ contains an odd-cycle), removing the attack $(\{a,b\},c)$ leads to $\{a\}$ not being weakly admissible (since then $\{b,c\}$ is (weakly) admissible in the reduct, attacking a).

We show that attacks with a certain type of conflicting tail (cf. SF_2 in Example 31) can safely be removed w.r.t. weak admissibility without changing the set of extensions.

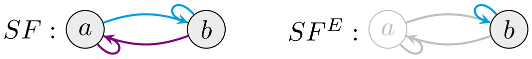
Proposition 32. *Let $SF = (A,R)$ be a SETAF, and let $(T,h) \in R$ with $T \neq \emptyset$ be an attack s.t. for every $t \in T$ there exists a set $T' \subseteq T \setminus \{h\}$ s.t. $(T',t) \in R$. Then $adm^w(SF) = adm^w(SF')$, where $SF' = (A, R \setminus \{(T,h)\})$.*

Due to Proposition 29, this result also applies to AFs, where it amounts to the removal of out-going attacks of self-attackers. Indeed, self-attackers can also be removed: we set $SF^\circ = SF\downarrow_{A'}$ where $A' = A \setminus \{a \mid (\{a\},a) \in R\}$. We can show that the removal of A' does not alter $adm^w(SF)$.

Theorem 33. *For any SETAF SF it holds $adm^w(SF) = adm^w(SF^\circ)$.*

In view of Proposition 32, one might wonder whether attacks of the form $(T \cup \{a\}, a)$ (with $T \neq \emptyset$) can also be removed. The intuitive reason why this is not the case is that the head h of the attack (T,h) can play a role in its own defense. Consequently, only if the "sub-conflict" occurs in $T \setminus h$, the property still applies. The following example illustrates this idea.

Example 34. *For the SETAF SF depicted below we get that $adm^w(SF) = \{\emptyset, \{a\}, \{b\}\}$. Consider $E = \{a\}$.*



Obviously, removing either argument changes the weakly admissible sets. Moreover, removing both attacks introduces the weakly admissible set $\{a,b\}$, while removing only one attack $(\{a,b\},a)$ gives us the weakly admissible sets $\{\emptyset,\{a\}\}$ (the other case is symmetrical).

Another central property is the compliance of weakly admissible extensions E with those in their reduct SF^E. This is formalized in the so-called *modularization property*.

Proposition 35 (Modularization)**.** *Let $SF = (A,R)$ be a SETAF and let $E, E' \subseteq A$ be disjoint. (i) If $E \in adm^w(SF)$ and $E' \in adm^w(SF^E)$, then $E \cup E' \in adm^w(SF)$. (ii) If $E \in adm^w(SF)$ and $E \cup E' \in adm^w(SF)$, then $E' \in adm^w(SF^E)$.*

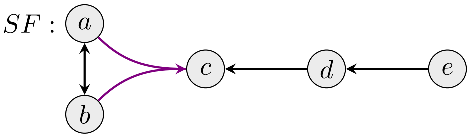
A direct consequence of modularization for weakly admissible semantics is that Dung's fundamental lemma holds.

Proposition 36. *Let SF be a SETAF. If $E \in adm^w(SF)$ and E defends $a \in A$, then $E \cup \{a\} \in adm^w(SF)$.*

4.3 Weak Defense

Having established the basic notion of weak admissibility, let us now turn our concept into a proper semantics family by also studying a suitable version of *weak defense*. That is, given $E \subseteq A$, under which conditions do we deem some argument $a \in A$ as "weakly defended" by E? To familiarize the reader with potential issues, let us consider an example.

Example 37. *Consider the following SETAF SF:*



Our goal is to find a weak defense notion where e weakly defends c. To this end consider the two attacks (d, c) and $(\{a, b, \}, c)$ directed towards it. i) The first one is refuted by e and thus should cause no harm. ii) The second one, $(\{a, b, \}, c)$, stems from a collective attack that is never valid in any reasonable viewpoint: a and b attack each other and can therefore never be jointly accepted.

Consequently, for each attack (T, h) there are two different reasons for considering h as "weakly defended" against it: i) (T, h) is counter-attacked or ii) T is not a "serious threat" for h. This motivates the following definition.

Definition 38. *Let $SF = (A, R)$ be a SETAF and let $E, X \subseteq A$. We say E* weakly defends *X (abbr. E w-defends X) if for each $(T, h) \in R$ with $h \in X$, one of the following two conditions hold:*

- *E attacks T, or*
- *the following two conditions hold simultaneously:*
  1. *there is no $E' \in adm^w(SF^E)$ with $T \subseteq E \cup E'$,*
  2. *there is some X' s.t. $X \subseteq X' \in adm^w(SF)$.*

Let us break down Definition 38. Given $E \subseteq A$, we want to know whether it defends some argument a. If for each $(T, a) \in R$ we have that E attacks T, then a is even classically defended. In this case, the first condition fires and thus, E weakly defends a. An example for the second condition is illustrated next.

Example 39. *Recall Example 37. Now $E = \{e\}$ indeed weakly defends $X = \{c\}$: we need to consider two attackers (d, c) and $(\{a, b\}, c)$. Regarding the former, the first condition in Definition 38 is satisfied, so we move on to $(\{a, b\}, c)$. The set E does not counter-attack $\{a, b\}$ so we have to check for the three conditions in the second item of Definition 38: 1. In SF^E which coincides with SF_2 from Example 31, the set $\{a, b\}$ is not weakly admissible; 2. $X = X'$ itself is weakly admissible in SF.*

The notion of weak defense for AFs has been refined in several ways (see e.g. (Dauphin, Rienstra, and van der Torre 2021) for thorough discussion on this matter). Within the scope of this work, we adapt a characterization from (Baumann, Brewka, and Ulbricht 2022) to SETAFs.

Proposition 40. *Let $SF = (A, R)$ be a SETAF and let $E \in adm^w(SF)$. Then, for any $D, X \subseteq A$ s.t. $E \subseteq X$ and $X = E \dot{\cup} D$ we have that E w-defends X iff*

1. *for any $(T, h) \in R$ with $h \in X$, there is no $E' \in adm^w(SF^E)$ with $T \subseteq E \cup E'$, and*
2. *there is some D' s.t. $D \subseteq D'$ with $D' \in adm^w(SF^E)$.*

This allows us to define our weak-semantics family.

Definition 41. *Let SF be a SETAF and let $E \in adm^w(F)$:*

- *E is* weakly complete, *$E \in com^w(SF)$, iff for each $X \supseteq E$ s.t. E w-defends X, we have $E = X$,*
- *E is* weakly preferred, *$E \in pref^w(SF)$, iff E is maximal w.r.t. $\subseteq$ in $adm^w(SF)$,*
- *E is* weakly grounded, *$E \in grd^w(SF)$, iff E is minimal w.r.t. $\subseteq$ in $com^w(SF)$.*

Example 42. *Recall the SETAF from Example 37. First of all, the empty set weakly defends e. Then, as we have discussed already, e weakly defends c, Since no further arguments are weakly defended, we get $\{c, e\} \in com^w(SF)$. Clearly, $\{c, e\} \in grd^w(SF)$ follows. Regarding weakly preferred semantics, note that e. g. $\{a, c, e\}$ is maximal in $adm^w(SF)$, so $pref^w(SF) = \{\{a, c, e\}, \{b, c, e\}\}$.*

We mention some central properties that hold for our novel weak semantics. First of all, these weak semantics faithfully generalize the AF case, as formalized below.

Proposition 43. *Let $SF = (A, R)$ be a SETAF s.t. $|T| = 1$ for each $(T, h) \in R$. Let F be the AF induced by SF. Then $\sigma^w(SF)$ (via Definition 41) and $\sigma^w(F)$ (via Definition 10) coincide for any $\sigma \in \{com, pref, grd\}$.*

As in the AF case, com^w extensions always exist.

Proposition 44. *For all SETAFs SF, $com^w(SF) \neq \emptyset$.*

Moreover, weakly preferred semantics can alternatively be defined as maximal in $com^w(SF)$ instead of $adm^w(SF)$.

Proposition 45. *For all SETAFs SF, $E \in pref^w(SF)$ iff E is $\subseteq$-maximal in $com^w(SF)$.*

We mention, however, that grd^w is not necessarily unique: this is inherited from the AF weak admissibility. As a final remark, we note that Theorem 33 can be extended to the remaining semantics as well.

Theorem 46. *Let SF be a SETAF. It holds $\sigma^w(SF) = \sigma^w(SF^\circ)$ for any $\sigma \in \{com, pref, grd\}$.*

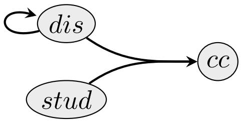
5 ABA, SETAFs and Weak Admissibility

Equipped with our thorough study on SETAF weak admissibility in Section 4, we are now ready to move on to ABA. ABAFs and SETAFs are closely related; we consider the following construction (König, Rapberger, and Ulbricht 2022; Caminada et al. 2024).

Definition 47. *Let $D = (\mathcal{L}, \mathcal{R}, \mathcal{A}, c)$ be an ABAF. We define the corresponding SETAF $SF_D = (A, R)$ by*

$$A = \mathcal{A}$$
$$R = \{(T, h) \mid h \in \mathcal{A}, T \subseteq \mathcal{A} \text{ and } T \vdash_{\mathtt{t}} c(h)\}$$

Example 48. *Recall our introductory Example 15 about climate change where $\mathcal{A} = \{stud, dis, cc\}$, $\mathcal{L} = \mathcal{A} \cup \{a_c \mid a \in \mathcal{A}\}$, contraries $c(a) = a_c$ for each assumption and rules $\mathcal{R} = \{(dis_c \leftarrow dis), (cc_c \leftarrow dis, stud)\}$. The instantiated SETAF SF_D is given as follows.*



We have that $adm^w(SF_D) = \{\emptyset, \{stud\}, \{cc\}, \{stud, cc\}\}$.

For the classical semantics, the following semantics correspondence is known.

Proposition 49. *Let D be an ABAF. For any semantics $\sigma \in \{adm, com, grd, pref\}$, it holds that $\sigma(D) = \sigma(SF_D)$*

We take this semantics correspondence as our starting point, and obtain the weak semantics for ABA as follows.

Definition 50. *Let D be an ABAF and SF_D be the instantiated SETAF. For any $\sigma^w \in \{adm^w, com^w, grd^w, pref^w\}$ we let $\sigma^w(D) = \sigma^w(SF_D)$.*

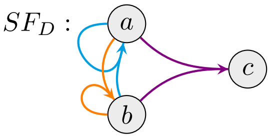
Hence, the models of an ABAF under weak admissibility are directly given by the corresponding weakly admissible semantics for the instantiated SETAF.

Example 51. *Continuing Example 48, the weakly admissible extensions of D can be read from the associated SETAF, i. e. $adm^w(D) = \{\emptyset, \{stud\}, \{cc\}, \{stud, cc\}\}$. Moreover, we e.g. have that $pref^w(D) = \{\{stud, cc\}\}$ which is the desired outcome (our agent accepts the scientific studies and the fact that climate change is real).*

This first example already shows that attacks from self-conflicting sets of assumptions are now treated satisfactory, i. e. they are ineffective. Let us examine the issues raised in Section 3. First of all, the SETAF for the ABAF D from Example 16 is just a single self-attacking argument, $SF_D = (\{a\}, \{(a, a)\})$, so $\{a\}$ is no longer weakly admissible as per Definition 50. The conflict was successfully directed to the paradoxical assumption itself. Building on this simple case, we furthermore demonstrate that self-attacking sets are now evaluated correctly.

Example 52. *Recall Example 17 with the ABAF $D = (\mathcal{L}, \mathcal{R}, \mathcal{A}, c)$ where $\mathcal{L} = \{a, b, c, a_c, b_c, c_c\}$, $\mathcal{A} = \{a, b\}$, contraries $c(a) = a_c$ as well as $c(b) = b_c$, $c(c) = c_c$, and the following rules $\mathcal{R}$. We compute the instantiated SETAF:*

$$\begin{array}{lcl} \mathcal{R}\text{:} & & \\ a_c & \leftarrow & a, b \\ b_c & \leftarrow & a, b \\ c_c & \leftarrow & a, b \end{array}$$



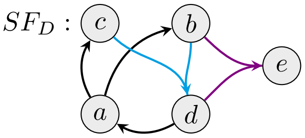
In the reduct $SF_D^{\{c\}}$ the joint attack from $\{a, b\}$ to a remains, so $\{a, b\}$ is not conflict-free. As a and b jointly attack c while there is no weakly admissible extension of $SF_D^{\{c\}}$ containing both, the attack to c can be ignored. Hence $\{c\} \in adm^w(SF_D)$. By Definition 50 the assumption c has thus become acceptable under weak admissibility for ABA, i. e. $\{c\} \in adm^w(D)$ which is the desired behavior.

Not only self-attackers, but also odd cycles of attacks are handled as intended under the SETAF-instantiation.

Example 53. *Recall Example 18 with the ABAF $D = (\mathcal{L}, \mathcal{R}, \mathcal{A}, c)$ where $\mathcal{L} = \{a, b, c, d, e, a_c, b_c, c_c, d_c, e_c\}$, $\mathcal{A} = \{a, b, c, d, e\}$, the usual contraries, and rules $\mathcal{R}$:*

$$a_c \leftarrow d \qquad b_c \leftarrow a \qquad c_c \leftarrow a \qquad d_c \leftarrow b, c \qquad e_c \leftarrow b, d$$

The instantiated SETAF is depicted below.



We have $adm^w(D) = \{\emptyset, \{e\}\}$, of which $\{e\}$ is the only complete, preferred and grounded extension. As we argued in Example 18, we deem $\{e\}$ acceptable which is captured by our SETAF instantiation.

Moving on from realizing the behavior we expect on our examples, let us now consider our desiderata. First of all, the relations between the semantics specified in desideratum **(SR)** translate directly from the SETAF setting to ABA, along with other properties we identified in Section 4. Most notably, we inherit the modularization property for ABA from the SETAF-setting, which guarantees that the union of a weakly admissible set of assumptions E together with any set of assumptions E', that is weakly admissible in the reduct SF_D^E, is weakly admissible, a non-trivial result under non-classical notions of admissibility that has proved powerful in multiple settings in AFs. Modularization is indispensable for the derivation of structural properties like the desiderata **(F)** and **(I)**.

Another, straightforward advantage of the proposed SETAF-based weak semantics for ABA is that in contrast to the AF-instantiation assumptions can become self-attackers themselves in the SETAF-instantiation, which makes negating their impact on other assumptions much easier and clearer. It comes as no surprise that the desiderata regarding paradoxical assumptions **(P)** and rule sets **(PR)** are satisfied under the proposed approach. We conclude our introduction of a SETAF-based weak admissibility for ABA by showing that our semantics satisfy the desiderata specified in Section 3.

Theorem 54. *The weakly admissible semantics for ABAF satisfies **(L)**, **(F)**, **(P)**, **(PR)** and **(I)**. The weakly complete and weakly preferred semantics satisfy **(SR)**.*

6 Summary und Related Work

In this paper, we studied paradoxical assumption sets in ABA and means to prevent them from blocking the acceptance of reasonable viewpoints. A similar problem received a lot of attention in abstract argumentation, namely the handling of self-attackers, direct and indirect. Several solutions have been proposed (Bodanza and Tohme 2009; Dondio and Longo 2019; Baumann, Brewka, and Ulbricht 2020b; Dvorák et al. 2022; Thimm 2023). Striving to tackle the issue for ABA, we decided to focus on one approach and introduced the weak admissibility semantics family to ABA which we believed to be the most promising candidates. The properties of weakly admissible semantics have been extensively studied for abstract argumentation (Dauphin, Rienstra, and van der Torre 2020; Dvorák, Ulbricht, and Woltran 2021; Baumann, Brewka, and Ulbricht 2022; Blümel and Ulbricht 2022b), which serves as a solid starting point for realizing the desired handling of paradoxical assumptions in ABA. That being said, we consider adding further weak semantics for ABA an important objective of future work.

We started by formalizing desiderata to capture the behavior we expect from a reasonable liberalization of classical admissibility for ABA. By doing so we provide a wider base for introducing and discussing more different weak semantics for ABA in the future, e. g. the undecidedness blocking semantics. Most of the principles, like independence, were motivated by our previous investigations of principle satisfaction by weak abstract argumentation semantics (Blümel and Ulbricht 2022a). We then observed that applying adm^w to the instantiated AF F_D does not provide us with intuitive acceptance conditions. However, SETAFs as proposed by Nielsen and Parsons (2006) are much closer to ABAFs compared to AFs, as they are capable of modeling *collective* attacks, a feature which is intrinsic to ABA. Driven by this observation, we proposed weak admissibility for SETAFs, as a faithful generalization of the weak semantics for Dung's AFs. We demonstrated that our proposal preserves many desirable properties known to hold for the weak AF semantics. We then applied these semantics to ABA by means of a SETAF instantiation and showed that they are well-behaved.

Our strategy to define weak admissibility for ABA was based on an instantiated abstract argumentation graph. This is similar in spirit to e.g. ASPIC$^+$ (Modgil and Prakken 2014) where the semantics are given in terms of the underlying AF. However, ABAFs are equipped with native semantics on their own, and thus do not rely on the construction of F_D to compute (classically) admissible extensions. This provides another viewpoint on the accepted assumptions besides the graph-theoretic interpretation, and also boosts the performance of ABA reasoners (Lehtonen, Wallner, and Järvisalo 2021; Lehtonen et al. 2023). Thus, both from a conceptual as well as computational point of view, it would be interesting to study how to define weak admissibility on ABAFs directly, without relying on abstract argumentation.

The conducted export of weak admissibility from Dung's classical AFs to two further argumentation formalisms, namely SETAFs and ABA, opens up a multitude of directions for future work. An important next step is a systematic investigation of the properties of weakly admissible semantics wrt. structured argumentation, e. g. whether it satisfies the rationality postulates (Caminada and Amgoud 2007). Several recent developments in the field of abstract argumentation are based on the notion of the reduct (Bengel and Thimm 2022; Dauphin, Rienstra, and van der Torre 2021; Blümel and Thimm 2023). This paper lays the groundwork for translating their results into SETAFs and ABA, in particular, the fact that the modularization property can be maintained is a promising result. Our work also contributes to a better understanding of inconsistent sets of assumptions in ABAFs by discussing both problematic cases and providing a first working solution. Equipped with this knowledge, one can now explore different options for handling the pointed out issues.

Acknowledgments

This research has been supported by the Vienna Science and Technology Fund (WWTF) through project ICT19-065, the Austrian Science Fund (FWF) through project P32830, and by the Federal Ministry of Education and Research of Germany and by Sächsische Staatsministerium für Wissenschaft, Kultur und Tourismus in the programme Center of Excellence for AI-research "Center for Scalable Data Analytics and Artificial IntelligenceDresden/Leipzig", project identification number: ScaDS.AI.

References

Baroni, P.; Gabbay, D.; Giacomin, M.; and van der Torre, L., eds. 2018. *Handbook of Formal Argumentation*. College Publications.

Baumann, R.; Brewka, G.; and Ulbricht, M. 2020a. Comparing weak admissibility semantics to their dung-style counterparts - reduct, modularization, and strong equivalence in abstract argumentation. In *Proc. KR 2020*, 79–88.

Baumann, R.; Brewka, G.; and Ulbricht, M. 2020b. Revisiting the foundations of abstract argumentation - semantics based on weak admissibility and weak defense. In *Proc. AAAI 2020*, 2742–2749. AAAI Press.

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.

Bengel, L., and Thimm, M. 2022. Serialisable semantics for abstract argumentation. In Toni, F.; Polberg, S.; Booth, R.; Caminada, M.; and Kido, H., eds., *Proc. COMMA 2022*, volume 353 of *Frontiers in Artificial Intelligence and Applications*, 80–91. IOS Press.

Bikakis, A.; Cohen, A.; Dvořák, W.; Flouris, G.; and Parsons, S. 2021. Joint attacks and accrual in argumentation frameworks. *FLAP* 8(6):1437–1501.

Blümel, L., and Thimm, M. 2023. Approximating weakly preferred semantics in abstract argumentation through vacuous reduct semantics. In Marquis, P.; Son, T. C.; and Kern-Isberner, G., eds., *Proc. KR 2023*, 107–116.

Blümel, L., and Ulbricht, M. 2022a. Defining defense and defeat in abstract argumentation from scratch - A generalizing approach. In *Proc. KR 2022*.

Blümel, L., and Ulbricht, M. 2022b. Defining defense and defeat in abstract argumentation from scratch - A generalizing approach. In Kern-Isberner, G.; Lakemeyer, G.; and Meyer, T., eds., *Proceedings of the 19th International Conference on Principles of Knowledge Representation and Reasoning, KR 2022, Haifa, Israel, July 31 - August 5, 2022*.

Bodanza, G. A., and Tohme, F. A. 2009. Two approaches to the problems of self-attacking arguments and general odd-length cycles of attack. *Journal of Applied Logic* 7(4):403 – 420. Special Issue: Formal Models of Belief Change in Rational Agents.

Caminada, M., and Amgoud, L. 2007. On the evaluation of argumentation formalisms. *Artif. Intell.* 171(5-6):286–310.

Caminada, M.; König, M.; Rapberger, A.; and Ulbricht, M. 2024. Attack semantics and collective attacks revisited. *Argument & Computation*. Pre press.

Charwat, G.; Dvořák, W.; Gaggl, S. A.; Wallner, J. P.; and Woltran, S. 2015. Methods for solving reasoning problems in abstract argumentation - A survey. *Artif. Intell.* 220:28–63.

Cyras, K.; Fan, X.; Schulz, C.; and Toni, F. 2018. Assumption-based argumentation: Disputes, explanations, preferences. In Baroni, P.; Gabbay, D.; Giacomin, M.; and van der Torre, L., eds., *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.

Dauphin, J.; Rienstra, T.; and van der Torre, L. 2020. A principle-based analysis of weakly admissible semantics. In Prakken, H.; Bistarelli, S.; Santini, F.; and Taticchi, C., eds., *Computational Models of Argument - Proceedings of COMMA 2020, Perugia, Italy, September 4-11, 2020*, volume 326 of *Frontiers in Artificial Intelligence and Applications*, 167–178. IOS Press.

Dauphin, J.; Rienstra, T.; and van der Torre, L. 2021. New weak admissibility semantics for abstract argumentation. In *Proc. CLAR 2021*, volume 13040 of *Lecture Notes in Computer Science*, 112–126. Springer.

Dimopoulos, Y.; Dvořák, W.; König, M.; Rapberger, A.; Ulbricht, M.; and Woltran, S. 2024. Redefining ABA+ semantics via abstract set-to-set attacks. In *Proc. AAAI 2024*, 10493–10500. AAAI Press.

Dondio, P., and Longo, L. 2019. Beyond reasonable doubt: A proposal for undecidedness blocking in abstract argumentation. *Intelligenza Artificiale* 13(2):123–135.

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):321–358.

Dunne, P. E. 2007. Computational properties of argument systems satisfying graph-theoretic constraints. *Artif. Intell.* 171(10-15):701–729.

Dvorák, W.; Rienstra, T.; van der Torre, L.; and Woltran, S. 2022. Non-admissibility in abstract argumentation. In *Computational Models of Argument - Proceedings of COMMA 2022, Cardiff, Wales, UK, 14-16 September 2022*, volume 353 of *Frontiers in Artificial Intelligence and Applications*, 128–139. IOS Press.

Dvorák, W.; Ulbricht, M.; and Woltran, S. 2021. Recursion in abstract argumentation is hard - on the complexity of semantics based on weak admissibility. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, 6288–6295. AAAI Press.

Dvořák, W., and Dunne, P. E. 2018. Computational problems in formal argumentation and their complexity. In *Handbook of Formal Argumentation*. College Publications. chapter 14, 631–687. Also appears in IfCoLog Journal of Logics and their Applications 4(8):2557–2622.

Dvořák, W.; König, M.; Ulbricht, M.; and Woltran, S. 2024. Principles and their computational consequences for argumentation frameworks with collective attacks. *J. Artif. Intell. Res.* 79:69–136.

Dvořák, W.; Greßler, A.; and Woltran, S. 2018. Evaluating SETAFs via answer-set programming. In *Proc. SAFA 2018*, volume 2171 of *CEUR Workshop Proceedings*, 10–21. CEUR-WS.org.

Dvořák, W.; König, M.; and Woltran, S. 2022a. Deletion-backdoors for argumentation frameworks with collective attacks. In *Proc. SAFA 2022*, volume 3236 of *CEUR Workshop Proceedings*, 98–110. CEUR-WS.org.

Dvořák, W.; König, M.; and Woltran, S. 2022b. Treewidth for argumentation frameworks with collective attacks. In *Proc. COMMA 2022*, 140–151. IOS Press.

Dvořák, W.; Rapberger, A.; and Woltran, S. 2020. On the different types of collective attacks in abstract argumentation: equivalence results for SETAFs. *J. Log. Comput.* 30(5):1063–1107.

Fan, X. 2018. A temporal planning example with assumption-based argumentation. In *Proc. PRIMA 2018*, 362–370.

Flouris, G., and Bikakis, A. 2019. A comprehensive study of argumentation frameworks with sets of attacking arguments. *Int. J. Approx. Reason.* 109:55–86.

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

König, M.; Rapberger, A.; and Ulbricht, M. 2022. Just a matter of perspective. In *Proc. COMMA 2022*, 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 *Proceedings of the 20th International Conference on Principles of Knowledge Representation and Reasoning, KR 2023, Rhodes, Greece, September 2-8, 2023*, 440–450.

Lehtonen, T.; Wallner, J. P.; and Järvisalo, M. 2021. Harnessing incremental answer set solving for reasoning in assumption-based argumentation. *Theory Pract. Log. Program.* 21(6):717–734.

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 *Proc. ArgMAS 2006*, volume 4766 of *Lecture Notes in Computer Science*, 54–73. Springer.

Thimm, M. 2023. On undisputed sets in abstract argumentation. In Williams, B.; Chen, Y.; and Neville, J., eds., *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, 6550–6557. AAAI Press.