Computing Stable Conclusions under the Weakest-Link Principle in the ASPIC+ Argumentation Formalism

Tuomo Lehtonen[1] , **Johannes P. Wallner**[2] , **Matti Järvisalo**[1]
[1]University of Helsinki, Helsinki, Finland
[2]Graz University of Technology, Graz, Austria
{tuomo.lehtonen, matti.jarvisalo}@helsinki.fi, wallner@ist.tugraz.at

Abstract

Rephrasing argumentation semantics in terms of subsets of defeasible elements allows for gaining new insights for reasoning about acceptance in established fragments of the central structured argumentation formalism of ASPIC$^+$. We provide a non-trivial generalization of these recent results, capturing preferences in ASPIC$^+$. In particular, considering preferences under the weakest-link principle, we show that the stable semantics can be phrased in terms of subsets of defeasible elements. We employ the rephrasing for establishing both complexity results and practical algorithms for reasoning about acceptance in this variant of ASPIC$^+$. Justified by completeness for the second level of the polynomial hierarchy, we develop an iterative answer set solving based approach to reasoning about acceptance under the so-called elitist lifting in ASPIC$^+$ frameworks. Our implementation of the approach scales well in practice.

1 Introduction

Argumentation is an established area of knowledge representation and reasoning research, with the fundamental aim of drawing conclusions from internally inconsistent or incomplete knowledge bases (Baroni et al. 2018; Gabbay et al. 2021; Atkinson et al. 2017). Arguments most often have an intrinsic structure made explicit through derivations from more basic structures. Computational models for structured argumentation (Bondarenko et al. 1997; Cyras et al. 2018; García and Simari 2014; García and Simari 2018; Besnard and Hunter 2008; Besnard and Hunter 2018)—ASPIC$^+$ (Modgil and Prakken 2018; Modgil and Prakken 2013) among the most prominent formalisms—enable making the internal structure of arguments explicit. In its general form, ASPIC$^+$ allows for arguments that combine strict inference rules—capturing deductively valid inferences—and defeasible inference rules—capturing presumptive inference—as well as accounting for preferential information. On one hand, this generality enables application in various settings, including legal reasoning (Prakken et al. 2015; Prakken 2012), ontology-based data access (Yun and Croitoru 2016), intelligence analysis (Toniolo et al. 2015), and information extraction for online crime reports (Schraagen et al. 2018). On the other hand, developing insights on the complexity and algorithmic aspects of reasoning in ASPIC$^+$ is still called for and remains a challenge.

Recently, an answer set programming (ASP) (Niemelä 1999; Brewka et al. 2015) approach to reasoning in ASPIC$^+$ without preferences was proposed (Lehtonen, Wallner, and Järvisalo 2020). Based on phrasing argumentation semantics in terms of subsets of defeasible elements in ASPIC$^+$, the approach avoids a potentially exponential translation to Dung's abstract argumentation frameworks (Dung 1995) employed earlier for realizing reasoning in ASPIC$^+$ (Snaith and Reed 2012; Thimm 2017) altogether, and instead employs an ASP solver on a direct ASP encoding on the level of ASPIC$^+$ for the reasoning task at hand.

In this work we study possibilities of generalizing these recent insights to preferential reasoning in ASPIC$^+$. In particular, we consider an instantiation of ASPIC$^+$ composed of atomic sentences, including axioms and ordinary premises, and allowing asymmetric negation. Towards a computational approach supporting preferences, we formally extend the foundations of rephrasing semantics developed earlier (Lehtonen, Wallner, and Järvisalo 2020) to cover preferences under the central weakest-link principle and elitist lifting (Modgil and Prakken 2018), focusing on the stable semantics as the first goal.

We provide both new complexity results as well as algorithms for credulous and skeptical reasoning, together with an experimental evaluation of an implementation of the algorithms. The non-trivial extension we develop of the earlier rephrasing to incorporate preferential reasoning is central for these contributions. In terms of complexity results, we show Σ_2^P and Π_2^P completeness for credulous and skeptical reasoning, respectively, for ASPIC$^+$ frameworks satisfying prominent rationality criteria (Modgil and Prakken 2018). Furthermore, we establish that reasoning has milder complexity (NP- and coNP-complete) for the case without defeasible rules, thereby being of the same complexity as the case without preferences (Lehtonen, Wallner, and Järvisalo 2020). Our rephrasing is vital here, since an explicated abstract framework yields structures not bounded polynomially, and these complexity classes restrict algorithms (among other aspects) to polynomial space consumption. The complexity results clearly indicate that inclusion of preferences increase complexity of reasoning in ASPIC$^+$. Moreover, our rephrasing highlights that the jump in complexity is due to a particular type of attack ("contradictory rebut") when combined with preferential reasoning.

In addition to the complexity results, in light of completeness of acceptance for the second level of the polynomial hierarchy, we employ the rephrasing in developing ASP-based counterexample-guided abstraction refinement algorithms for credulous and skeptical reasoning in ASPIC$^+$ with preferences, making use of incremental ASP solving techniques, and show that an implementation of the approach shows promising scalability up to hundreds of sentences.

After recalling necessary preliminaries of ASPIC$^+$ in Section 2, we overview our complexity results in Section 3. The extended rephrasing with preferences is presented in two parts: in Section 4 defeats on assumption are presented and Section 5 shows the rephrasing of stable semantics. In Section 6 the novel algorithms are presented, and experiments are presented in Section 7. We conclude after discussing related works in Section 8. An extended version of the paper with formal proofs is available via the authors' webpages.

2 ASPIC$^+$ Framework

We recall background on ASPIC$^+$ (Modgil and Prakken 2018; Modgil and Prakken 2013; Prakken 2010). We assume a set (language) $\mathcal{L}$ composed of atoms x. We start with contrariness.

Definition 1. *Let a contrary function be* $^-: \mathcal{L} \to 2^{\mathcal{L}}$. *We say that* $a \in \mathcal{L}$ *is a contrary of* $b \in \mathcal{L}$ *if* $a \in \overline{b}$ *and* $b \notin \overline{a}$. *We say that* a *is a contradictory of* b *if* $a \in \overline{b}$ *and* $b \in \overline{a}$.

That is, contraries represent an asymmetric relation, while contradictories are symmetric. When a and b are contradictory to each other we sometimes write $-a = b$ (or $a = -b$). An atom may be a contradictory to several atoms.

A central part of an ASPIC$^+$ framework is a knowledge base $\mathcal{K} \subseteq \mathcal{L}$ comprised of a defeasible part (ordinary premises $\mathcal{K}_p$) and a non-defeasible part (axioms $\mathcal{K}_n$).

Definition 2. *A knowledge base is a set* $\mathcal{K}_n \cup \mathcal{K}_p = \mathcal{K} \subseteq \mathcal{L}$, *with disjoint sets* $\mathcal{K}_n$ *(axioms) and* $\mathcal{K}_p$ *(ordinary premises).*

Another part of ASPIC$^+$ is a set of rules over $\mathcal{L}$, denoted by $\mathcal{R}$. This set is composed of defeasible rules $a_1, \ldots, a_n \Rightarrow b$ and strict rules $a_1, \ldots, a_n \to b$. We denote the set of defeasible rules by $\mathcal{R}_d$ and the set of strict rules by $\mathcal{R}_s$. When we do not distinguish between strict or defeasible rules, we write $a_1, \ldots, a_n \rightsquigarrow b$. A partial function $n : \mathcal{R}_d \to \mathcal{L}$ gives names to defeasible rules. For a rule $r = a_1, \ldots, a_n \rightsquigarrow b$, we denote its head by $head(r) = b$ and its body by $body(r) = \{a_1, \ldots, a_n\}$.

For preferences, we consider preorders (i.e., reflexive and transitive binary relations) $\leq = \leq_p \cup \leq_d$, composed of a preorder on ordinary premises $\leq_p$ and a preorder on defeasible rules $\leq_d$.

Definition 3. *An argumentation theory (AT) is a tuple* $(\mathcal{L}, \mathcal{R}, n, ^-, \mathcal{K}, \leq)$, *with a knowledge base* $\mathcal{K} \subseteq \mathcal{L}$, *rules* $\mathcal{R} = \mathcal{R}_d \cup \mathcal{R}_s$ *over* $\mathcal{L}$, *a contrary function* $^-: \mathcal{L} \to 2^{\mathcal{L}}$, *a partial function* $n : \mathcal{R}_d \to \mathcal{L}$, *and a preorder* $\leq = \leq_p \cup \leq_d$.

Each part of an AT is assumed to be finite. Arguments are constructed from parts of an AT. An argument represents a "derivation tree" starting from elements in the knowledge base and uses rules to derive a conclusion.

Definition 4. *Given an AT* $T = (\mathcal{L}, \mathcal{R}, n, ^-, \mathcal{K}, \leq)$, *the set of arguments in* T *is inductively defined as follows.*

- *If* $x \in \mathcal{K}$, *then* $A = x$ *is an argument with* $\mathtt{Conc}(A) = x$.
- *If* $A_1, \ldots, A_n$ *are arguments,* $x_i = \mathtt{Conc}(A_i)$ *for* $1 \leq i \leq n$, *and* $(x_1, \ldots, x_n \rightsquigarrow x) \in \mathcal{R}$, *then* $A = A_1, \ldots, A_n \rightsquigarrow x$ *is an argument with* $\mathtt{Conc}(A) = x$.

There are no other arguments.

We make use of the following shorthands.

Definition 5. *Let* $T = (\mathcal{L}, \mathcal{R}, n, ^-, \mathcal{K}, \leq)$ *be an AT, and* A *an argument in* T.

- *If* $A = x \in \mathcal{K}$ *then* $\mathtt{Sub}(A) = \{A\}$ *and* $\mathtt{Rules}(A) = \emptyset$.
- *If* $A = A_1, \ldots, A_n \rightsquigarrow x$, *then*
  $\mathtt{Sub}(A) = \{A\} \cup \bigcup_{i=1}^{n} \mathtt{Sub}(A_i)$,
  $\mathtt{TopRule}(A) = (\mathtt{Conc}(A_1), \ldots, \mathtt{Conc}(A_n) \rightsquigarrow x)$, *and*
  $\mathtt{Rules}(A) = \{\mathtt{TopRule}(A)\} \cup \bigcup_{i=1}^{n} \mathtt{Rules}(A_i)$.

Further, $\mathtt{Prem}(A) = \mathtt{Sub}(A) \cap \mathcal{K}$, $\mathtt{Prem}_\mathtt{d}(A) = \mathtt{Prem}(A) \cap \mathcal{K}_p$, *and* $\mathtt{DefRules}(A) = \mathtt{Rules}(A) \cap \mathcal{R}_d$.

That is, we define shorthands for the subarguments ($\mathtt{Sub}$) of an argument, the rules and defeasible rules in the argument ($\mathtt{Rules}$ and $\mathtt{DefRules}$), the topmost rule ($\mathtt{TopRule}$), the premises of the argument within the knowledge base ($\mathtt{Prem}$), and the ordinary premises ($\mathtt{Prem}_\mathtt{d}$). Further, $\mathtt{defPart}(A) = \mathtt{Prem}_\mathtt{d}(A) \cup \mathtt{DefRules}(A)$. If $A \in \mathcal{K}$, then $\mathtt{TopRule}(A)$ is undefined. We extend the shorthands for a set of arguments $\mathcal{A}$ as $\mathtt{Conc}(\mathcal{A}) = \{\mathtt{Conc}(A) \mid A \in \mathcal{A}\}$ and $\mathtt{TopRule}(\mathcal{A}) = \{\mathtt{TopRule}(A) \mid A \in \mathcal{A}\}$. For each shorthand $f \in \{\mathtt{Sub}, \mathtt{Rules}, \mathtt{DefRules}, \mathtt{Prem}, \mathtt{Prem}_\mathtt{d}\}$ returning a set, we define $f(\mathcal{A}) = \bigcup_{A \in \mathcal{A}} f(A)$. An argument A is an immediate subargument of $B = A_1, \ldots, A_n \rightsquigarrow x$ if $A \in \{A_1, \ldots, A_n\}$. We allow only finite structures as arguments (i.e., arguments which are "trees" of finite size), and consider as arguments those arguments A for which $\mathtt{Sub}(A)$ is finite (disallowing infinite chaining of rules, e.g., via $x \rightsquigarrow x$), as also defined by Modgil and Prakken (2018).

Conflicts among arguments are defined via attacks between arguments.

Definition 6. *Given an AT* $T = (\mathcal{L}, \mathcal{R}, n, ^-, \mathcal{K}, \leq)$ *and two arguments* A *and* B *in* T, *argument* A *attacks argument* B *iff* A *undercuts, rebuts, or undermines* B, *where*

- A *undercuts* B *(on* B'*) iff* $\mathtt{Conc}(A) \in \overline{n(r)}$ *for some* $B' \in \mathtt{Sub}(B)$ *such that* $\mathtt{TopRule}(B') = r$ *is defeasible;*
- A *rebuts* B *(on* B'*) iff* $\mathtt{Conc}(A) \in \overline{x}$ *for some* $B' = B_1, \ldots, B_n \Rightarrow x \in \mathtt{Sub}(B)$*; and*
- A *undermines* B *(on* x*) iff* $\mathtt{Conc}(A) \in \overline{x}$ *and* $x \in \mathtt{Prem}_\mathtt{d}(B)$.

That is, an argument attacks another argument on the defeasible parts of the latter. Ordinary premises can be attacked by undermining, and defeasible rules can be attacked by rebutting the conclusion or undercutting the rule itself. For rebuts and undermining one distinguishes further if $\mathtt{Conc}(A)$ and x are contraries or contradictories: in the former case we say that A contrary undermines (rebuts) B and in the latter that A contradictory undermines (rebuts) B.

The preorders $\leq_p$ and $\leq_d$ on the defeasible parts are lifted to strict partial orders (i.e., to irreflexive, asymmetric, and transitive binary relations) $\triangleleft_p$ and $\triangleleft_d$, via lifting operators. We focus on the elitist lifting (Modgil and Prakken 2018). As usual, the strict part of $\leq_p$ ($\leq_d$) is defined by $x <_p y$ iff $x \leq_p y$ and $y \not\leq_p x$ (same for $\leq_d$).

Definition 7. *Let $\leq$ be a preorder on a set X. Define $\triangleleft^{\mathit{Eli}}$ for two non-empty Y, $Z \subseteq X$ by $Y \triangleleft^{\mathit{Eli}} Z$ iff $\exists a \in Y$ s.t. $\forall b \in Z$ we have $a < b$. Moreover, $\emptyset \ntriangleleft^{\mathit{Eli}} Y$ and $Z \triangleleft^{\mathit{Eli}} \emptyset$ for each non-empty Z.*

That is, Z is preferred to Y if there is at least one element in Y that is strictly less preferred to each element in Z. The empty set is a special case, being strictly preferred to each non-empty set, and cannot be less preferred than any set.

For two preorders $\leq_p$ and $\leq_d$ and their liftings $\triangleleft^{\mathit{Eli}} = \triangleleft_p^{\mathit{Eli}} \cup \triangleleft_d^{\mathit{Eli}}$, one defines a strict partial order $\prec$ on arguments, denoting the preference (ranking) on arguments using certain principles. We omit the superscript *Eli* in the remainder of the paper. We focus here on the weakest-link principle, by which one considers all defeasible elements of arguments in the comparison. An argument B is strictly preferred to A ($A \prec B$) whenever

- if $\mathtt{DefRules}(A) = \mathtt{DefRules}(B) = \emptyset$, then $\mathtt{Prem_d}(A) \triangleleft \mathtt{Prem_d}(B)$;
- if $\mathtt{Prem_d}(A) = \mathtt{Prem_d}(B) = \emptyset$, then $\mathtt{DefRules}(A) \triangleleft \mathtt{DefRules}(B)$;
- else $\mathtt{Prem_d}(A) \triangleleft \mathtt{Prem_d}(B)$ and $\mathtt{DefRules}(A) \triangleleft \mathtt{DefRules}(B)$.

Defeats between arguments, i.e., successful attacks, are defined based on argument preferences as follows.

Definition 8. *Given an AT $T = (\mathcal{L}, \mathcal{R}, n, ^-, \mathcal{K}, \leq)$ and two arguments A and B in T, argument A defeats argument B iff A successfully undercuts, rebuts, or undermines B, where*

- *A successfully undercuts B if A undercuts B;*
- *A successfully rebuts B (on B') iff A contrary rebuts B, or A rebuts B on B' and $A \not\prec B'$; and*
- *A successfully undermines B (on x) iff A contrary undermines B, or A undermines B on x and $A \not\prec x$.*

In other words, an undercut always succeeds, as do contrary rebuts and undermining attacks. For intuition, an undercut attacks the very applicability of a rule, and such an attack takes precedence over preferences between arguments. Similarly, as contraries signal an asymmetric incompatibility of two atoms, the presence of a contrary of an atom x results in x being attacked regardless of preferences. Modgil and Prakken (2013; 2018) provide a discussion on this. For the contradictory variants of rebut and undermining (i.e., when conclusion of A and the ordinary premise or conclusion of a defeasible rule in B are contradictories of each other), the preference order $\prec$ on arguments decides whether the attack succeeds: if A is strictly less preferred to the attacked subargument, the attack fails, otherwise it is a defeat.

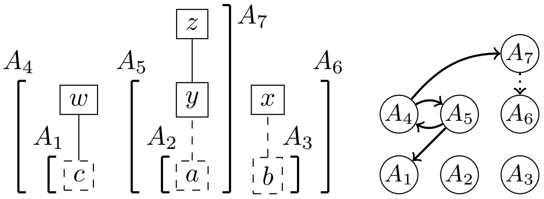

Figure 1: Example AT and corresponding AF

Example 1. *Let $T = (\mathcal{L}, \mathcal{R}, n, ^-, \mathcal{K}, \leq)$ be an AT with $\mathcal{L} = \{a, b, c, w, x, y, z\}$, $\mathcal{K}_p = \{a, b, c\}$, $\mathcal{K}_n = \emptyset$, $\overline{x} = \{z\}$, $\overline{z} = \{x\}$, $\overline{c} = \{y\}$, $\overline{n(r_1)} = \{w\}$, $\mathcal{R}_d = \{(r_1 : a \Rightarrow y), (r_2 : b \Rightarrow x)\}$, and $\mathcal{R}_s = \{(y \rightarrow z), (c \rightarrow w)\}$. We write names of defeasible rules by r_i before the rule. Moreover, let $a \leq_p b$ and $r_1 \leq_d r_2$. The AT is shown in Figure 1(left) with the arguments it gives rise to. Defeasible premises or inference is marked with dashed lines. The arguments and their defeat relation are shown on the right side of this figure. It holds that A_4 undercuts both A_5 and A_7 (on A_5), A_5 contrary undermines both A_1 and A_4 (on A_1), and A_7 contradictory rebuts A_6. Since x and z are contradictory to each other, but A_7 concludes z via a strict rule, it is the case that A_6 does not contradictory rebut A_7. Except for the last attack, all are successful and thus defeats, since $A_7 \prec A_6$ ($\mathtt{Prem_d}(A_7) = \{a\} \triangleleft_p \{b\} = \mathtt{Prem_d}(A_6)$ due to $a <_p b$ and $\mathtt{DefRules}(A_7) = \{r_1\} \triangleleft_d \{r_2\} = \mathtt{DefRules}(A_6)$ due to $r_1 <_d r_2$). The unsuccessful contradictory rebut is denoted as a dotted arrow.*

In ASPIC$^+$, there are important conditions which ensure satisfaction of rationality postulates, which might fail in the very general case (Modgil and Prakken 2018; Caminada 2018). It turns out that these conditions are useful for our computational approach, as well. These conditions constrain the set of strict rules in an AT, and we make use of three such conditions here, borrowing from Modgil and Prakken (2018).

A set of strict rules $\mathcal{R}_s$ is said to be closed under transposition if for each $a_1, \ldots, a_n \rightarrow b \in \mathcal{R}_s$ it holds that for each i, $1 \leq i \leq n$, we have $a_1, \ldots, a_{i-1}, b', a_{i+1}, \ldots, a_n \rightarrow a'_i \in \mathcal{R}_s$ for all contradictories $-b' = b$ and $-a'_i = a_i$, and at least one contradictory of each must exist. An AT T is strict-consistent if there are no arguments A, B with $\mathtt{defPart}(A) = \mathtt{defPart}(B) = \emptyset$ s.t. $\mathtt{Conc}(A)$ is a contrary of $\mathtt{Conc}(B)$, or $\mathtt{Conc}(A)$ and $\mathtt{Conc}(B)$ are contradictory to each other.

Definition 9. *We say an AT T is well-formed if*

- *$\mathcal{R}_s$ is closed under transposition,*
- *T is strict-consistent, and*
- *if x is a contrary of y then $y \notin \mathcal{K}_n$ and $y \neq head(r)$ for all $r \in \mathcal{R}_s$.*

For intuition, these conditions aim to avoid certain inconsistencies that may arise, with strict-consistency likely the most immediate: if an AT is not strict-consistent then there are a arguments composed only of axioms and strict rules concluding contrary or contradictory atoms. For details we refer the reader to the work of Modgil and Prakken (2018). We remark that we are using a subset of the conditions of "well-defined" ATs as defined by Modgil and Prakken (2018).

Semantics of ATs are defined via a translation to (abstract) argumentation frameworks (AFs) (Dung 1995). An AF is a pair $F = (\mathcal{A}, \mathcal{D})$ of a set of (abstract) arguments $\mathcal{A}$ and defeats $\mathcal{D} \subseteq \mathcal{A} \times \mathcal{A}$ between arguments. If $(A, B) \in \mathcal{D}$ we say that A defeats B. Similarly, $\mathcal{S} \subseteq \mathcal{A}$ defeats $B \in \mathcal{A}$ if there is an $A \in \mathcal{S}$ with $(A, B) \in \mathcal{D}$. We say that $\mathcal{S}$ defends an argument A if for each $B \in \mathcal{A}$ such that $(B, A) \in \mathcal{D}$, there is a $C \in \mathcal{S}$ such that $(C, B) \in \mathcal{D}$. We consider the AF semantics of conflict-free and admissible sets, and complete and stable extensions, with the corresponding functions $\sigma \in \{cf, adm, com, stb\}$. A semantics $\sigma(F) \subseteq 2^{\mathcal{A}}$ returns a set of extensions. An extension under a semantics σ is a σ-extension for short.

Definition 10. *Given an AF $F = (\mathcal{A}, \mathcal{D})$, a set $\mathcal{E} \subseteq \mathcal{A}$ is conflict-free (in F) if there are no A, B in $\mathcal{E}$ such that $(A, B) \in \mathcal{D}$. The set of all conflict-free sets of F is denoted by $cf(F)$. For an $\mathcal{E} \in cf(F)$, we have*

- *$\mathcal{E} \in adm(F)$ iff each $A \in \mathcal{E}$ is defended by $\mathcal{E}$;*
- *$\mathcal{E} \in com(F)$ iff $\mathcal{E} \in adm(F)$ and each A defended by $\mathcal{E}$ is in $\mathcal{E}$;*
- *$\mathcal{E} \in stb(F)$ iff $\mathcal{E}$ defeats each argument in $\mathcal{A} \setminus \mathcal{E}$.*

ATs can be translated to AFs as follows.

Definition 11. *Let $T = (\mathcal{L}, \mathcal{R}, n, ^-, \mathcal{K}, \leq)$ be an AT. An AF $F = (\mathcal{A}, \mathcal{D})$ corresponds to T if $\mathcal{A}$ is the set of all arguments in T and $\mathcal{D}$ the defeat relation based on T.*

Reasoning on ATs consists of checking whether a queried atom is warranted, by asking (credulously) whether there is a σ-extension having an argument concluding the atom, or (skeptically) whether all σ-extensions have such an argument.

Definition 12. *Given an AT $T = (\mathcal{L}, \mathcal{R}, n, ^-, \mathcal{K}, \leq)$ and an AF F corresponding to T, we say that $x \in \mathcal{L}$ is*

- *skeptically justified in T under semantics σ if in each $\mathcal{E} \in \sigma(F)$ there is an $A \in \mathcal{E}$ with $\texttt{Conc}(A) = x$;*
- *credulously justified in T under semantics σ if there is an $\mathcal{E} \in \sigma(F)$ with an $A \in \mathcal{E}$ s.t. $\texttt{Conc}(A) = x$.*

Example 2. *Continuing Example 1, the AF corresponding to the AT is shown in Figure 1 (right). There are two stable extensions: $\mathcal{E}_1 = \{A_1, A_2, A_3, A_4, A_6\}$ and $\mathcal{E}_2 = \{A_2, A_3, A_5, A_6, A_7\}$. Since $\texttt{Conc}(A_4) = w$, $A_4 \in \mathcal{E}_1$, and there is no argument in $\mathcal{E}_2$ with conclusion w, it holds that w is credulously but not skeptically justified under stable semantics.*

A useful property is that each complete extension of an AF corresponding to an AT is closed under subarguments, i.e., if $\mathcal{E}$ is complete and $A \in \mathcal{E}$, then all $\texttt{Sub}(A)$ are in $\mathcal{E}$, as well. Note that stable extensions are also complete extensions.

Proposition 1 (Modgil and Prakken 2013)**.** *Let $T = (\mathcal{L}, \mathcal{R}, n, ^-, \mathcal{K}, \leq)$ be an AT, $F = (\mathcal{A}, \mathcal{D})$ the AF corresponding to T, and $\mathcal{E} \in com(F)$. It holds that $\mathcal{E}$ is closed under subarguments.*

Building on earlier work (Lehtonen, Wallner, and Järvisalo 2020), we utilize the concept of so-called assumptions (P, D) for a given AT $T = (\mathcal{L}, \mathcal{R}, n, ^-, \mathcal{K}, \leq)$, which represent parts of the defeasible elements: $P \subseteq \mathcal{K}_p$ and $D \subseteq \mathcal{R}_d$. We define $(P, D) \sqsubseteq (P', D')$ if $P \subseteq P'$ and $D \subseteq D'$.

Given a set of rules $\mathcal{R}$ and a set of atoms $L \subseteq \mathcal{L}$, we say that $x \in \mathcal{L}$ is derivable from L via $\mathcal{R}$, denoted by $L \vdash_{\mathcal{R}} x$, if (i) $x \in L$ or (ii) there is a sequence of rules $(r_1, \ldots, r_n)$ from $\mathcal{R}$ s.t. $head(r_n) = x$ and for each rule r_i it holds that each atom in the body of r_i is derived from rules earlier in the sequence or is in L, i.e., $body(r_i) \subseteq L \cup \bigcup_{j<i} head(r_j)$.

We extend the derivability notion to assumptions in a straightforward way: given an AT $T = (\mathcal{L}, \mathcal{R}, n, ^-, \mathcal{K}, \leq)$ and an assumption (P, D) in T, we say that from (P, D) one can derive (in T) an atom $x \in \mathcal{L}$, denoted by $(P, D) \vdash_T x$, if $P \cup \mathcal{K}_n \vdash_{D \cup \mathcal{R}_s} x$, i.e., x is derivable from the defeasible elements in the assumption and all non-defeasible parts in the AT. The deductive closure of an assumption in T is then defined as $Th_T(P, D) = \{x \in \mathcal{L} \mid (P, D) \vdash_T x\}$. We say that a rule r is applicable by an assumption (P, D) if $body(r) \subseteq Th_T(P, D)$, i.e., all elements of the body of r can be derived using the assumption. For an assumption (P, D) in T and an argument A in T we say that A is based on (P, D) if A uses only defeasible elements from this assumption, i.e., A is based on (P, D) in T if $\texttt{DefRules}(A) \subseteq D$ and $\texttt{Prem}_\texttt{d}(A) \subseteq P$.

Assumptions and arguments have a direct connection.

Proposition 2 (Lehtonen, Wallner, and Järvisalo 2020)**.** *Let $T = (\mathcal{L}, \mathcal{R}, n, ^-, \mathcal{K}, \leq)$ be an AT and (P, D) an assumption in T. There is an argument A based on (P, D) in T with $\texttt{Conc}(A) = x$ iff $(P, D) \vdash_T x$.*

3 Complexity of Reasoning

In this section, we provide an overview of our complexity results for preferential reasoning under stable semantics in ASPIC$^+$. The rephrasing of stable semantics essential for the formal proofs of the complexity results (as well as for the algorithms presented in Section 6) is postponed until Sections 4–5.

While an AF corresponding to a given AT is, in general, not bounded polynomially in size w.r.t. the given AT, we showed (Lehtonen, Wallner, and Järvisalo 2020) that one can define criteria on assumptions (P, D) so that there is a direct correspondence between assumptions and extensions of arguments, without explicating the extension (and without constructing the corresponding AF), with assumptions being bounded polynomially regarding the input AT. Such an approach is essential to show complexity results for classes in the polynomial hierarchy (which are restricted to polynomial space). Concretely, assumptions, and criteria on assumptions, allow for a design of algorithms that operate in polynomial space, and fall into a class of the polynomial hierarchy, in the form of a non-deterministic construction of a (P, D) assumption and subsequent verification of the conditions imposed by the semantics.

Preferential reasoning is important to ASPIC$^+$, but obtaining criteria on assumptions reflecting preferences is, as we will show, non-immediate, and, moreover, increases computational complexity. Recall that preferences in ASPIC$^+$ lead to a modified attack structure (the defeat relation), which means that preferences change conflicts in the corresponding AF, an object we have to avoid explicitly constructing in order to obtain tight complexity results.

We show that preferential reasoning, under stable semantics, the weakest-link principle, and elitist lifting in ASPIC$^+$, remains on the same level of complexity in case no defeasible rules are present as in the case no preferences are present. This result, in fact, aligns reasoning under stable semantics in ASPIC$^+$ and assumption-based argumentation (ABA) with certain kinds of preferences (called ABA$^+$) (Čyras and Toni 2016) from the view of complexity (Lehtonen, Wallner, and Järvisalo 2021).

Theorem 3. *For ATs without defeasible rules and stable semantics, credulous justification is NP-complete and skeptical justification is coNP-complete.*

While hardness for these cases follows in a rather direct fashion from existing results (e.g., see hardness on AFs (Dvořák and Dunne 2018) or ASPIC$^+$ without preferences (Lehtonen, Wallner, and Järvisalo 2020)), membership results are based on our rephrasing that characterizes stable semantics in terms of conditions on assumption sets, formally presented in the next subsections.

However, in contrast, when including defeasible rules we show that reasoning jumps to the second level of the polynomial hierarchy, under prominent instantiations of ASPIC$^+$ satisfying rationality postulates, concretely what we refer to as well-formed ATs. These results clearly set ASPIC$^+$ and ABA$^+$ apart, in terms of computational complexity of reasoning. Hardness holds even in the case of no strict rules.

Theorem 4. *For well-formed ATs and stable semantics, credulous justification is Σ_2^P-complete and skeptical justification is Π_2^P-complete. Hardness holds even when there are no strict rules.*

For intuition on why well-formedness, i.e., conditions usually used for showing rationality postulates, are useful for showing complexity results, these conditions also lead to properties that allow to align stable extensions and stable assumptions in a more direct way. In the following section we present a rephrasing of defeats to the viewpoint of assumptions, where we clearly distinguish between two types of defeats with one being the underlying reason for the complexity jump, and subsequently highlight useful properties following from assuming well-formed ATs.

4 Defeats on Assumptions

In this section we investigate defeats defined on assumptions (P, D), and their relation to defeats on arguments, as a precursor to rephrasing semantics on assumptions.

It will be useful to classify defeats into two categories: one contains successful contradictory rebuts and the other category, which we call individual defeats, contains all other types of defeats.

Definition 13. *Let $T = (\mathcal{L}, \mathcal{R}, n, ^-, \mathcal{K}, \leq)$ be an AT, and A and B two arguments in T. We say that A individually defeats B if A*

- *successfully undercuts B,*
- *contrary rebuts B, or*
- *successfully undermines B.*

The reason for considering these two classes will become clear when considering their corresponding definitions on assumptions and their respective complexity. We define individual defeats on assumptions as follows.

Definition 14. *Let $T = (\mathcal{L}, \mathcal{R}, n, ^-, \mathcal{K}, \leq)$ be an AT, and (P, D) be an assumption in T. We say that (P, D) individually defeats an $x \in \mathcal{K}_p \cup \mathcal{R}_d$ if*

- *$x = p \in \mathcal{K}_p$ and*
  - *there is a contrary of p in $Th_T(P, D)$ (contrary undermine) or*
  - *there is a contradictory of p in $Th_T(P', D)$ with $P' = \{p' \in P \mid p' \not< p\}$ (contradictory undermine),*
- *$x = r \in \mathcal{R}_d$ and*
  - *$\overline{n(r)} \cap Th_T(P, D) \neq \emptyset$ (undercut),*
  - *there is a contrary of $head(r)$ in $Th_T(P, D)$ (contrary rebut).*

These conditions reflect individual defeats on arguments. Assumption (P, D) undermines an ordinary premise if one can derive a contrary of the premise or a contradictory using not less preferred premises. Undercuts and contrary rebuts on assumptions directly reflect undercuts and contrary rebuts on arguments (which do not require preference handling).

Next we show that an assumption (P, D) individually defeating x represents a set of arguments $\mathcal{S}$ for which the following holds: for each argument B containing x (as an ordinary premise or a defeasible rule) there is an argument $A \in \mathcal{S}$ individually defeating B. This motivates the name: the defeat targets a single defeasible element and does not require (more) context information on the defeated argument.

Proposition 5. *Let $T = (\mathcal{L}, \mathcal{R}, n, ^-, \mathcal{K}, \leq)$ be an AT, (P, D) be an assumption in T, and $x \in \mathcal{K}_p \cup \mathcal{R}_d$. There is an argument A based on (P, D) s.t. A individually defeats an argument B in T whenever $x \in \mathtt{Prem_d}(B) \cup \mathtt{DefRules}(B)$ iff (P, D) individually defeats x.*

What remains is the case of contradictory rebut. It turns out that phrasing contradictory rebuts on assumptions leads to a more complex condition.

Definition 15. *Let $T = (\mathcal{L}, \mathcal{R}, n, ^-, \mathcal{K}, \leq)$ be an AT, and (P, D) and (P', D') be assumptions in T. Further, let $W = Th_T(P'', D) \cup Th_T(P, D'')$ with*

- *$P'' = \{p \in P \mid \exists p' \in P', p \not< p'\}$ and*
- *$D'' = \{r \in D \mid \exists r' \in D', r \not< r'\}$.*

We say (P, D) contradictory rebuts (P', D') on $r' \in D'$ if

- *P' is non-empty and there is a contradictory of $head(r')$ in W, or*
- *P' is empty and there is a contradictory of $head(r')$ in $Th_T(P, D'')$.*

Sets P'' and D'' are chosen in a way to accommodate the preference induced by the elitist lifting (for each element there is one not more preferred). The case distinction reflects the weakest-link principle (if P' is empty only defeasible rules are taken into account; D' is non-empty).

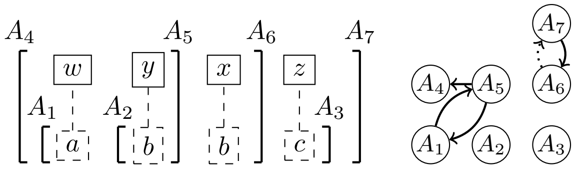

Figure 2: AT with $\mathcal{R}_s = \emptyset$ and corresponding AF

For individual defeats it holds that if an assumption individually defeats a defeasible element x, then this assumption (and its corresponding arguments) defeats all assumptions (arguments) individually containing the defeasible element x. In contrast, contradictory defeats are defined on assumptions contradictory rebutting a *specific* assumption (and as we will see, making no claim on sub-assumptions, for instance), which is why checking contradictory rebuts is computationally harder than checking individual attacks.

Example 3. *Let $T = (\mathcal{L}, \mathcal{R}, n, ^{-}, \mathcal{K}, \leq)$ be an AT with $\mathcal{L} = \{a, b, c, w, x, y, z\}$, $\mathcal{K}_p = \{a, b, c\}$, $\mathcal{K}_n = \emptyset$, $\overline{x} = \{z\}$, $\overline{z} = \{x\}$, $\overline{a} = \{y\}$, $\overline{y} = \{a\}$, $\mathcal{R}_s = \emptyset$, and $\mathcal{R}_d = \{(r_1 : b \Rightarrow x), (r_2 : b \Rightarrow y), (r_3 : c \Rightarrow z), (r_4 : a \Rightarrow w)\}$. Further, let $a \leq_p b$, $b \leq_p c$, and $r_1 \leq_d r_3$. The AT and corresponding AF are shown in Figure 2. Argument A_5 successfully contradictory undermines A_1 (and A_4 on A_1) and A_1 successfully contradictory rebuts A_5. The latter holds because $\texttt{Prem}_\texttt{d}(A_1) = \{a\} \lhd \{b\} = \texttt{Prem}_\texttt{d}(A_5)$ due to $a <_p b$, but $\texttt{DefRules}(A_1) = \emptyset$ and $\emptyset \not\lhd \{r_2\} = \texttt{DefRules}(A_5)$. Argument A_7 successfully contradictory rebuts A_6, but A_6 does not successfully contradictory rebut A_7, due to $A_6 \prec A_7$ ($b <_p c$ and $r_1 <_d r_3$). Consider $P = \{b, c\}$ and $D = \{r_1, r_2\}$. It holds that (P, D) individually defeats a (because we can derive the contradictory of a from (P, D) and neither b nor c is less preferred to a). For $P' = \{a, c\}$ and $D' = \{r_3, r_4\}$ it holds that (P, D) contradictory rebuts (P', D') on r_3, since with $P'' = \{b, c\} = P$ and $D'' = \{r_1, r_2\} = D$ we can derive a contradictory of $head(r_3) = z$ (namely x) by Definition 15. However, (P, D) does not contradictory rebut $(P^*, D^*) = (\{c\}, \{r_3\})$ on r_3 (the ingredients for A_7) since, again by Definition 15, neither from $(\{c\}, D)$ nor from $(P, \{r_2\})$ one can derive x (one requires b and defeasible rule r_1).*

That is, an assumption might contradictory rebut an assumption (P, D), but not $(P', D') \sqsubseteq (P, D)$. Next we state that if (P, D) contrary rebuts (P', D'), then there is an argument based on (P, D) successfully contradictory rebutting arguments containing exactly (P', D') as their defeasible part.

Proposition 6. *Let $T = (\mathcal{L}, \mathcal{R}, n, ^{-}, \mathcal{K}, \leq)$ be an AT, and (P, D) and (P', D') be assumptions in T. If (P, D) contradictory rebuts (P', D') on r then all arguments B in T with*

- $\texttt{Prem}_\texttt{d}(B) = P'$,
- $\texttt{DefRules}(B) = D'$, *and*
- $\texttt{TopRule}(B) = r$.

are successfully contradictory rebutted on B by an argument A based on (P, D).

5 Rephrasing Stability on Assumptions

In this section we make use of assumptions and defeats to define stable semantics on assumptions such that "stable assumptions" have a useful correspondence to stable extensions. In contrast to the case without preferences (Lehtonen, Wallner, and Järvisalo 2020), where a direct correspondence was possible, it turns out that defeasible rules, and even more, the *combination* of strict and defeasible rules, together with preferential reasoning, is a particular challenge for computation. General translation results (Modgil and Prakken 2018) that reduce an ASPIC$^+$ AT to an AT without strict rules or without defeasible rules are not directly usable in our case.[1]

We first show auxiliary results. First, a useful property is that if an argument attacks (but not necessarily defeats) another argument, this attack is present as a defeat, in a certain well-behaved manner, in case the underlying AT is well-formed (in particular closed under transposition). In case no strict rules exist at all, a rather strong condition holds: there is a defeat among subarguments of the two arguments involved in an attack. In case there are strict rules, which are closed under transposition, it can be the case that the defeat is "propagated" to a superargument of one of the two arguments (more precisely a superargument of a subargument of one of the two), but nevertheless in a controlled manner, as stated in the following lemma.

Lemma 7. *Let T be an AT with $\mathcal{R}_s$ closed under transposition, and A and B two arguments in T with $\texttt{TopRule}(A) \in \mathcal{R}_s$. If A unsuccessfully contradictory rebuts or unsuccessfully contradictory undermines B, then there is an argument C in T which defeats A and for every $C' \in \texttt{Sub}(C)$ with $\texttt{TopRule}(C')$ being defeasible or $C' \in \mathcal{K}_p$ it holds that $C' \in \texttt{Sub}(A) \cup \texttt{Sub}(B)$.*

Based on the lemma, the next proposition formalizes that attacks in well-formed ATs, do not, in a sense, disappear.

Proposition 8. *Let $T = (\mathcal{L}, \mathcal{R}, n, ^{-}, \mathcal{K}, \leq)$ be an AT, and argument A in T attacks argument B in T.*

- *If $\mathcal{R}_s = \emptyset$, then A defeats B or some $B' \in \texttt{Sub}(B)$ defeats A.*
- *If $\mathcal{R}_s$ is closed under transposition, then there is an argument C based on $(\texttt{Prem}_\texttt{d}(A) \cup \texttt{Prem}_\texttt{d}(B), \texttt{DefRules}(A) \cup \texttt{DefRules}(B))$ s.t. C defeats A or B.*

Example 4. *In the AT from Example 1, A_7 contradictory rebuts A_6, but A_7 does not defeat A_6. No subargument of A_6 defeats A_7. This exemplifies the fact that in presence of both strict and defeasible rules the first item of Proposition 8 does not hold. A simple example with only strict rules is obtained by considering two ordinary premises a and b, with $a <_p b$ and a strict rule $a \rightarrow x$ with x and b being contradictory. Then there are no defeats (the argument concluding x is strictly weaker than argument b); however, there is an attack based on contradictory undermining. Crucially the sets of strict rules are not closed under transposition.*

[1] While a translation of strict rules to defeasible ones has been proposed (Li and Parsons 2015), it turned out not to be valid (Li 2019) and hence not applicable for our case.

Another useful ingredient is a property related to subargument closure (see Proposition 1): if $\mathcal{E}$ is complete in an AT, then for each argument A whose defeasible elements are in $\mathcal{E}$ (i.e., $\texttt{defPart}(A) \subseteq \texttt{defPart}(\mathcal{E})$) it holds that $A \in \mathcal{E}$. In contrast to closure under subarguments, this kind of closure property does not hold in general (see extended version for an example); however, it does hold in case there are either no strict or no defeasible rules for complete semantics, and for well-formed ATs for stable semantics. Motivation for this property is that a correspondence of a (P, D) and an extension can be defined in a way that the extension contains all arguments that can be constructed using defeasible elements in (P, D) and strict components (strict rules, axioms) in an AT. Formally, for a semantics σ a set of ATs is closed under defeasible elements if for each $\mathcal{E} \in \sigma(F)$, for the AF F corresponding to T, it holds that all arguments A with $\texttt{defPart}(A) \subseteq \texttt{defPart}(\mathcal{E})$ are in $\mathcal{E}$.

Proposition 9. *The set of ATs with $\mathcal{R}_s = \emptyset$ or $\mathcal{R}_d = \emptyset$ is closed under defeasible elements for complete semantics and the set of ATs with $\mathcal{R}_s$ closed under transposition is closed under defeasible elements for stable semantics.*

We are now in a position to define stability on assumptions. We define w-stable assumptions for the case of well-formed ATs, and s-stable assumption for when there are only strict rules.

Definition 16. *Let $T = (\mathcal{L}, \mathcal{R}, n, ^-, \mathcal{K}, \leq)$ be a well-formed AT. Assumption (P, D) of T is w-stable if all D are applicable by (P, D) and*

1. *$\nexists x, y \in Th_T(P, D)$ such that x is a contrary of y, x and y are contradictory, or $x \in \{\overline{n(r)} \mid r \in D\}$,*
2. *(P, D) individually defeats all $p \in \mathcal{K}_p \setminus P$, and*
3. *(P, D) contradictory rebuts all (P', D') on some r' where*
   - *each rule in D' is applicable by (P', D'),*
   - *$D' \nsubseteq D$, and*
   - *$P' \subseteq P$ and $D' \subseteq D_U$, where D_U are the rules in $\mathcal{R}_d$ that are not individually defeated by (P, D).*

The first condition states that one cannot derive from this assumption x and y that are contrary or contradictory to each other, and that one cannot derive contraries/contradictories of names of rules in D. This ensures conflict-freeness. Individual defeats and contradictory rebuts among arguments based on a w-stable (P, D) are prevented (otherwise contrary or contradictory x and y are derivable, or contrary/contradictory of a rule name is derivable). The second condition is direct: if violated there is an argument (an ordinary premise) not defeated. The last condition takes care of undercuts and rebuts "outside" the set to be stable: D_U "removes" all possible arguments that are undercut or contrary rebutted (by individual defeat), and if there is an argument still possible, then this argument has P' and D' (exactly) as its defeasible elements, and it has to be the case that (P, D) contradictory rebuts (P', D').

Example 5. *Considering the AT from Example 3, it holds that (P, D) from the example is not a w-stable assumption (a corresponding stable extension containing all arguments based on (P, D) also does not exist): condition 1 is satisfied directly, and condition 2, as well. However, condition 3 is not satisfied: while (P, D) does contradictory rebut (P', D') from the example, it holds that (P, D) does not contradictory rebut (P^*, D^*). A w-stable assumption is $(\{a, b, c\}, \{r_3, r_4\})$.*

We show the correspondence between w-stable assumptions and stable extensions.

Theorem 10. *Let $T = (\mathcal{L}, \mathcal{R}, n, ^-, \mathcal{K}, \leq)$ be a well-formed AT, and $F = (\mathcal{A}, \mathcal{D})$ the corresponding AF to T.*

- *If (P, D) is a w-stable assumption in T, then $\mathcal{E} = \{A \mid$* A based on (P, D) in $T\}$ *is a stable extension of F.*
- *If $\mathcal{E}$ is a stable extension of F, then (P, D) is a w-stable assumption of T with $P = \texttt{Prem}_\texttt{d}(\mathcal{E})$ and $D = \texttt{DefRules}(\mathcal{E})$.*

In the theorem, we see that closure under defeasible elements is utilized directly (e.g., via the construction of extensions from an assumption). Well-formedness is most visible in proving the correspondence. For an intuition, the proof of Proposition 8 (and Lemma 7) makes use of transposition: otherwise there might be a conflict derived on an assumption, but not present in a corresponding AF, since, e.g., if A attacks B, $A \prec B$, and the top rule of A is strict. Without transposition such a conflict might not materialize as a defeat in the AT (with transposition one may "chain" transposed rules to obtain a defeat).

We move on to the case of no defeasible rules. Here only individual defeats are present since rebuts require defeasible rules.

Definition 17. *Let $T = (\mathcal{L}, \mathcal{R}, n, ^-, \mathcal{K}, \leq)$ be an AT with $\mathcal{R}_d = \emptyset$. We say that $(P, \emptyset)$, $P \subseteq \mathcal{K}_p$, is s-stable in T if*

- *$(P, \emptyset)$ does not individually defeat any $p \in P$, and*
- *$(P, \emptyset)$ individually defeats all $p \in \mathcal{K}_p \setminus P$.*

If clear from the context, we write (P) instead of $(P, \emptyset)$.

Proposition 11. *Let $T = (\mathcal{L}, \mathcal{R}, n, ^-, \mathcal{K}, \leq)$ be an AT with $\mathcal{R}_d = \emptyset$, and $F = (\mathcal{A}, \mathcal{D})$ the corresponding AF to T.*

- *If (P) is an s-stable assumption in T, then $\mathcal{E} = \{A \mid$* A based on $(P, \emptyset)$ in $T\}$ *is a stable extension of F.*
- *If $\mathcal{E}$ is a stable extension of F, then (P) is an s-stable assumption of T with $P = \texttt{Prem}_\texttt{d}(\mathcal{E})$.*

The preceding results directly lead to the result that checking credulous or skeptical justification is possible via inspecting stable assumptions instead of extensions.

Proposition 12. *Given an AT T which is well-formed (or which has empty $\mathcal{R}_d$), and an $x \in \mathcal{L}$, it holds that*

- *x is credulously justified in T under stable semantics iff there is a w-stable (s-stable) assumption (P, D) in T with $x \in Th_T(P, D)$, and*
- *x is skeptically justified in T under stable semantics iff in all w-stable (s-stable) assumptions (P, D) in T we have $x \in Th_T(P, D)$.*

We end this section with complexity results for verifying whether an assumption is s- or w-stable, which form the basis for the membership results of Theorem 3 and Theorem 4.

Algorithm 1 Credulous justification

```
Require: Well-formed AT $T$ and queried atom $s \in \mathcal{L}$
Ensure: return YES if $s$ is credulously justified in $T$ under
    stable semantics, NO otherwise
1: $\pi \leftarrow \pi_{1,2}(T) \cup \{\leftarrow \texttt{not derived}(s)\}$
2: while $\pi$ has an answer set $M$ do
3:     if $\pi \cup \pi_{\neg 3}(M)$ has no answer sets then return YES
4:     else $\pi \leftarrow \pi \cup \pi_r(M)$
5: return NO
```

Theorem 13. *Verifying whether an assumption is s-stable is in P, and verifying whether an assumption is w-stable is coNP-complete.*

6 Algorithm for Stable Semantics

Our approach to credulous acceptance under stable semantics is outlined as Algorithm 1. (The ASP encodings used are detailed in the extended version available online.) One ASP solver provides candidate solutions corresponding to assumptions (P, D) that derive the queried atom, satisfy Items 1–2 of Definition 16 and the applicability of every $r \in D$ (Lines 1–2). Another ASP solver checks for counterexamples to the solution candidate (Line 3). A counterexample is an assumption (P', D') such that (P, D) does not contradictory rebut it while the other conditions of Item 3 of Definition 16 hold. If there is no counterexample, the candidate is w-stable and thus the query is credulously justified (Line 3).

We employ two encodings, $\pi_{1,2}(T)$ for candidate generation and $\pi_{\neg 3}$ for counterexample finding, with the following properties: (P, D) derives s and satisfies the first two conditions in T iff there is an answer set M of $\pi_{1,2}(T) \cup \{\leftarrow \texttt{not derived(s)}\}$ with $P \cup D = \{p \in (\mathcal{K}_p \cup \mathcal{R}_d) \mid \textbf{in}(p) \in M\}$. The $\{\leftarrow \texttt{not derived(s)}\}$ is a constraint ruling out answers where s is not derivable from the guessed assumption set. The answer set M also indicates the premises and rules that (P, D) does not defeat individually. For verifying if the answer set corresponds to a w-stable assumption, it holds that (P', D') is a counterexample in T iff there is an answer set M' of $\pi_{\neg 3}(M)$ with $P' \cup D' = \{p \in (\mathcal{K}_p \cup \mathcal{R}_d) \mid \textbf{in}(p) \in M'\}$. If a counterexample is found, the current π is refined, via π_r, to exclude each $(P'', D'') \sqsubseteq (P, D)$ (Line 5); the encoding $\pi_r(M)$ enforces that some defeasible element that is not in the candidate in M needs to be in any future candidate. Excluding all subassumptions is valid, since if (P, D) does not contradictory rebut a counterexample (P', D'), then no subassumption of (P, D) can contradictory rebut (P', D') either (follows from Definition 15).

The algorithm for skeptical reasoning (Algorithm 2) follows with relatively minor changes from Algorithm 1. In short, it searches for a counterexample, namely, for a stable extension that does not contain the query s. For this, the constraint in Line 1 is changed to rule out answers where s is derivable from the guessed assumption set. Then the algorithm returns NO if it finds a suitable counterexample and YES otherwise.

Algorithm 2 Skeptical justification

```
Require: Well-formed AT $T$ and queried atom $s \in \mathcal{L}$
Ensure: return YES if $s$ is skeptically justified in $T$ under
    stable semantics, NO otherwise
1: $\pi \leftarrow \pi_{1,2}(T) \cup \{\leftarrow \texttt{derived(s)}\}$
2: while $\pi$ has an answer set $M$ do
3:     if $\pi \cup \pi_{\neg 3}(M)$ has no answer sets then return NO
4:     else $\pi \leftarrow \pi \cup \pi_r(M)$
5: return YES
```

7 Experiments

We implemented (available at https://bitbucket.org/coreo-group/aspforaspic/) the ASP-based CEGAR algorithms using the incremental Python interface of Clingo v5.5.1 (Gebser et al. 2019) under default parameters. The experiments were run on 2.60-GHz Intel Xeon E5-2670 8-core 64-GB machines with CentOS 7 under a per-instance 600-second time and 16-GB memory limit.

We generated ATs with $N = 100, 200, ...800$ atoms (i.e., members of $\mathcal{L}$ excluding the names for defeasible rules) as follows, selecting one queried non-premise atom per framework: all atoms aside from axioms and 10% of defeasible rules were assigned a contradictory or asymmetric contrary (each with equal probability); 5% of all atoms are axioms and 20% of atoms are premises. We varied the number of rules deriving each atom (rpa) and the sizes of rule bodies (rs): for each non-premise atom, the number of rules deriving the atom was chosen at random from $[1, 5]$ or $[1, 10]$, as was the number of atoms in the body of each rule body. When the head has a contradictory, a rule deriving it was chosen to be strict with a 10% probability and the atoms in the rule body were selected from the sentences that have a contradictory (closure under transposition requires each atom present in a strict rule to have a contradictory). Closure under transposition was enforced by creating the required additional strict rules. We generated preference relations over premises and defeasible rules for each framework by choosing for both a random permutation $(x_i)_{0<i\leq n}$ of the elements and for each $j < i$ setting x_i to be preferred to x_j with probability 30%. For each N and different combinations of rpa and rs we generated 10 frameworks.

Tables 1 and 2 give the number of timeouts and mean runtimes (in parentheses, with timeouts included as 600 s) of our approach for each N and choice of rs, rpa for credulous and skeptical reasoning, respectively. The empirical hardness of the instances depend on the parameters and reasoning tasks. Skeptical justification is empirically harder than credulous on all of the parameter families tested. The instances with many smaller rules seems the hardest for both reasoning tasks while the instances with fewer and smaller rules are comparatively easier.

In terms of systems for direct runtime comparison, there are few options currently. The Tweety library (Thimm 2017) offers one possible point of comparison. However, Tweety employs a translation to Dung AFs, and as its first step explicitly generates the arguments from a given AT. We observed that already the argument construction step fails (due

	#timeouts (mean runtime (s))					
N	rs=5, rpa=5		rs=5, rpa=10		rs=10, rpa=10	
100	0	(1)	0	(16)	0	(6)
200	0	(8)	4	(321)	0	(49)
300	1	(85)	6	(428)	0	(182)
400	0	(64)	4	(491)	3	(474)
500	4	(316)	10	(600)	10	(600)
600	0	(222)	10	(600)	10	(600)
700	1	(405)	10	(600)	10	(600)
800	3	(556)	10	(600)	10	(600)

Table 1: Timeouts and runtimes on credulous reasoning.

	#timeouts (mean runtime (s))					
N	rs=5, rpa=5		rs=5, rpa=10		rs=10, rpa=10	
100	0	(2)	0	(74)	0	(7)
200	0	(51)	8	(509)	0	(114)
300	1	(190)	10	(600)	1	(279)
400	7	(445)	9	(582)	5	(531)
500	9	(567)	10	(600)	10	(600)
600	8	(524)	10	(600)	10	(600)
700	9	(580)	10	(600)	10	(600)
800	10	(600)	10	(600)	10	(600)

Table 2: Timeouts and runtimes on skeptical reasoning.

to time or memory out) for all but seven of the benchmark instances (five of the succeeding instances had $N = 100$ and two $N = 200$). This further emphasizes the benefits of our AT-level ASP-based approach compared to approaches resorting to AF translation.

8 Related Work

For overviews on computational approaches to structured argumentation see, e.g., surveys by Dvořák and Dunne (2018) and Cerutti et al. (2018). Regarding rephrasing (sets of) arguments—or argument structures—in different forms, in assumption-based argumentation (Bondarenko et al. 1997; Cyras et al. 2018) semantics have been defined in terms of extensions of arguments and on subsets of assumptions, and recently extended with preferences, e.g., in ABA$^+$ (Čyras and Toni 2016). In ASPIC$^+$, preferred subtheories (Brewka 1989) and specific instantiations of ASPIC$^+$ have a semantic correspondence (Modgil and Prakken 2013). There are also connections between repairs of ontological knowledge bases and extensions of AFs instantiated from knowledge bases (Croitoru and Vesic 2013). By studying outcomes of rule-based systems and their instantiated arguments and attacks, certain subparts of the rule-base are connected to semantics of the resulting AF (Amgoud and Besnard 2019). Structured argumentation frameworks, including a fragment of ASPIC$^+$ (Heyninck and Straßer 2021) by making use of properties implying satisfaction of rationality postulates, have been connected to maximal consistent subset reasoning (see, e.g., the recent survey by Arieli, Borg, and Heyninck (2019)). In contrast, we consider ASPIC$^+$ with atomic strict and defeasible rules, contraries and contradictories, weakest-link principle, elitist lifting, and all forms of attacks and defeats defined on these based on (Modgil and Prakken 2018) (undermining, undercut, and rebut, with their contrary and contradictory versions), and present a rephrasing of stability in terms of defeasible elements of the given AT which is designed for computational purposes. We show correspondences for the fragments of (i) AT without defeasible rules and (ii) well-formed ATs. Conditions underlying well-formed ATs have been used earlier for showing satisfaction of properties regarding rationality. We show that these are also viable for computational purposes.

Computational complexity for structured argumentation includes several results, e.g., for assumption-based argumentation (Dimopoulos, Nebel, and Toni 2002; Cyras, Heinrich, and Toni 2021; Karamlou, Cyras, and Toni 2019; Lehtonen, Wallner, and Järvisalo 2021), for deductive argumentation, e.g., (Wooldridge, Dunne, and Parsons 2006; Hirsch and Gorogiannis 2010), and for DeLP, e.g., (Alfano et al. 2021). For ASPIC$^+$, complexity results were obtained for the case without preferences, upon which we build (Lehtonen, Wallner, and Järvisalo 2020). We present novel complexity results for ASPIC$^+$ with preferences.

Cerutti et al. (2018) provide a survey on systems for structured argumentation. Specific to ASPIC$^+$, the systems Tweety (Thimm 2017) and TOAST (Snaith and Reed 2012) implement reasoning for specific tasks. We were unable to obtain the source code for TOAST (Snaith and Reed 2012) from the authors for a further potential direct comparison. Further systems implementing reasoning in contexts drawing inspiration from ASPIC$^+$ (as considered here) include EPR (Visser 2008) and Arg2P (Calegari et al. 2022).

9 Conclusions

As a non-trivial extension of a recently proposed approach to reasoning in ASPIC$^+$, we established both complexity results and formal foundations for an incremental ASP approach to stable conclusions under the weakest-link principle. In terms of complexity, as the main contributions we established completeness for the second level of the polynomial hierarchy for both credulous and skeptical acceptance, thereby witnessing a jump in complexity due to inclusion of preferences. Pertaining to this complexity class, we proposed a counterexample-guided abstraction refinement style approach using incremental ASP solving for the task, scaling up to hundreds of atoms in practice. The approach circumvents a potential exponential blow-up intrinsic to approaches translating ASPIC$^+$ reasoning to abstract argumentation via formally rephrasing the semantics in terms of subsets of defeasible elements.

Our work opens up directions for further work, including extending the approach to further variants of preferential reasoning in ASPIC$^+$ and structured argumentation (Beirlaen et al. 2018; Young, Modgil, and Rodrigues 2016; Dyrkolbotn, Pedersen, and Broersen 2018; Heyninck and Straßer 2019), in terms of both complexity analysis and algorithms, as well as investigating computational properties of further fragments of ASPIC$^+$.

Acknowledgements

This work has been financially supported in part by Academy of Finland (grant 322869), University of Helsinki Doctoral Programme in Computer Science DoCS, and by the Austrian Science Fund (FWF) P35632. The authors wish to thank the Finnish Computing Competence Infrastructure (FCCI) for supporting this project with computational and data storage resources.

References

Alfano, G.; Greco, S.; Parisi, F.; Simari, G. I.; and Simari, G. R. 2021. Incremental computation for structured argumentation over dynamic DeLP knowledge bases. *Artif. Intell.* 300:103553.

Amgoud, L., and Besnard, P. 2019. A formal characterization of the outcomes of rule-based argumentation systems. *Knowledge and Information Systems* 61(1):543–588.

Arieli, O.; Borg, A.; and Heyninck, J. 2019. A review of the relations between logical argumentation and reasoning with maximal consistency. *Ann. Math. Artif. Intell.* 87(3):187–226.

Atkinson, K.; Baroni, P.; Giacomin, M.; Hunter, A.; Prakken, H.; Reed, C.; Simari, G. R.; Thimm, M.; and Villata, S. 2017. Towards artificial argumentation. *AI Magazine* 38(3):25–36.

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

Beirlaen, M.; Heyninck, J.; Pardo, P.; and Straßer, C. 2018. Argument strength in formal argumentation. *IfCoLog Journal of Logics and their Applications* 5(3):629–676.

Besnard, P., and Hunter, A. 2008. *Elements of Argumentation*. MIT Press.

Besnard, P., and Hunter, A. 2018. A review of argumentation based on deductive arguments. In Baroni et al. (2018). chapter 9, 437–484.

Bondarenko, A.; Dung, P. M.; Kowalski, R. A.; and Toni, F. 1997. An abstract, argumentation-theoretic approach to default reasoning. *Artif. Intell.* 93:63–101.

Brewka, G.; Delgrande, J. P.; Romero, J.; and Schaub, T. 2015. asprin: Customizing answer set preferences without a headache. In *AAAI*, 1467–1474. AAAI Press.

Brewka, G. 1989. Preferred subtheories: An extended logical framework for default reasoning. In *IJCAI*, 1043–1048. Morgan Kaufmann.

Calegari, R.; Omicini, A.; Pisano, G.; and Sartor, G. 2022. Arg2P: An argumentation framework for explainable intelligent systems. *J. Log. Comput.* 32(2):369–401.

Caminada, M. 2018. Rationality postulates: Applying argumentation theory for non-monotonic reasoning. In Baroni et al. (2018). chapter 15, 771–795.

Cerutti, F.; Gaggl, S. A.; Thimm, M.; and Wallner, J. P. 2018. Foundations of implementations for formal argumentation. In Baroni et al. (2018). chapter 15, 688–767.

Croitoru, M., and Vesic, S. 2013. What can argumentation do for inconsistent ontology query answering? In *SUM*, volume 8078 of *LNCS*, 15–29. Springer.

Čyras, K., and Toni, F. 2016. ABA+: Assumption-based argumentation with preferences. In *KR*, 553–556. AAAI Press.

Cyras, K.; Fan, X.; Schulz, C.; and Toni, F. 2018. Assumption-based argumentation: Disputes, explanations, preferences. In Baroni et al. (2018). chapter 7, 365–408.

Cyras, K.; Heinrich, Q.; and Toni, F. 2021. Computational complexity of flat and generic assumption-based argumentation, with and without probabilities. *Artif. Intell.* 293:103449.

Dimopoulos, Y.; Nebel, B.; and Toni, F. 2002. On the computational complexity of assumption-based argumentation for default reasoning. *Artif. Intell.* 141(1/2):57–78.

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.

Dvořák, W., and Dunne, P. E. 2018. Computational problems in formal argumentation and their complexity. In Baroni et al. (2018). chapter 13, 631–688.

Dyrkolbotn, S. K.; Pedersen, T.; and Broersen, J. M. 2018. On elitist lifting and consistency in structured argumentation. *IfCoLog Journal of Logics and their Applications* 5(3):709–746.

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. 2014. Defeasible logic programming: Delp-servers, contextual queries, and explanations for answers. *Argument Comput.* 5(1):63–88.

García, A. J., and Simari, G. R. 2018. Argumentation based on logic programming. In Baroni et al. (2018). chapter 8, 409–435.

Gebser, M.; Kaminski, R.; Kaufmann, B.; and Schaub, T. 2019. Multi-shot ASP solving with clingo. *Theory Pract. Log. Program.* 19(1):27–82.

Heyninck, J., and Straßer, C. 2019. A fully rational argumentation system for preordered defeasible rules. In *AAMAS*, 1704–1712. IFAAMAS.

Heyninck, J., and Straßer, C. 2021. Rationality and maximal consistent sets for a fragment of ASPIC+ without undercut. *Argument & Compututation* 12(1):3–47.

Hirsch, R., and Gorogiannis, N. 2010. The complexity of the warranted formula problem in propositional argumentation. *J. Log. Comput.* 20(2):481–499.

Karamlou, A.; Cyras, K.; and Toni, F. 2019. Complexity results and algorithms for bipolar argumentation. In *AAMAS*, 1713–1721. IFAAMAS.

Lehtonen, T.; Wallner, J. P.; and Järvisalo, M. 2020. An answer set programming approach to argumentative reasoning in the ASPIC+ framework. In *KR*, 636–646. IJCAI.org.

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.

Li, Z., and Parsons, S. 2015. On argumentation with purely defeasible rules. In *SUM*, volume 9310 of *LNCS*, 330–343. Springer.

Li, Z. 2019. *A Purely Defeasible Argumentation Framework*. Ph.D. Dissertation, City University of New York.

Modgil, S., and Prakken, H. 2013. A general account of argumentation with preferences. *Artif. Intell.* 195:361–397.

Modgil, S., and Prakken, H. 2018. Abstract rule-based argumentation. In Baroni et al. (2018). chapter 6, 287–364.

Niemelä, I. 1999. Logic programs with stable model semantics as a constraint programming paradigm. *Ann. Math. Artif. Intell.* 25(3-4):241–273.

Prakken, H.; Wyner, A. Z.; Bench-Capon, T. J. M.; and Atkinson, K. 2015. A formalization of argumentation schemes for legal case-based reasoning in ASPIC+. *J. Log. Comput.* 25(5):1141–1166.

Prakken, H. 2010. An abstract framework for argumentation with structured arguments. *Argument & Computation* 1(2):93–124.

Prakken, H. 2012. Reconstructing Popov v. Hayashi in a framework for argumentation with structured arguments and Dungean semantics. *Artif. Intell. Law.* 20(1):57–82.

Schraagen, M.; Odekerken, D.; Testerink, B.; and Bex, F. 2018. Argumentation-driven information extraction for online crime reports. In *CIKM*, volume 2482 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Snaith, M., and Reed, C. 2012. TOAST: online ASPIC^{+} implementation. In *COMMA*, volume 245 of *FAIA*, 509–510. IOS Press.

Thimm, M. 2017. The Tweety library collection for logical aspects of artificial intelligence and knowledge representation. *Künstliche Intelligenz* 31(1):93–97.

Toniolo, A.; Norman, T. J.; Etuk, A.; Cerutti, F.; Ouyang, W. R.; Srivastava, M. B.; Oren, N.; Dropps, T.; Allen, J. A.; and Sullivan, P. 2015. Supporting reasoning with different types of evidence in intelligence analysis. In *AAMAS*, 781–789. ACM.

Visser, W. 2008. Implementation of argument-based practical reasoning. Master's thesis, Utrecht University.

Wooldridge, M. J.; Dunne, P. E.; and Parsons, S. 2006. On the complexity of linking deductive and abstract argument systems. In *AAAI*, 299–304. AAAI Press.

Young, A. P.; Modgil, S.; and Rodrigues, O. 2016. Prioritised default logic as rational argumentation. In *AAMAS*, 626–634. ACM.

Yun, B., and Croitoru, M. 2016. An argumentation workflow for reasoning in ontology based data access. In *COMMA*, volume 287 of *FAIA*, 61–68. IOS Press.

Supplement to: Computing Stable Conclusions under the Weakest-Link Principle in the ASPIC+ Argumentation Formalism

Tuomo Lehtonen[1] , **Johannes P. Wallner**[2] , **Matti Järvisalo**[1]
[1]University of Helsinki, Helsinki, Finland
[2]Graz University of Technology, Graz, Austria
{tuomo.lehtonen, matti.jarvisalo}@helsinki.fi, wallner@ist.tugraz.at

Formal Proofs

Proof of Proposition 5. We first remark that if an argument A individually defeats an argument B on B'' with $B'' = x$ or $\mathtt{TopRule}(B'') = x$, then A individually defeats all arguments B' with $x \in \mathtt{defPart}(B')$. To see this, consider the cases: if A successfully undercuts B on B'' with $\mathtt{TopRule}(B'') = r$, then A successfully undercuts any B' if $r \in \mathtt{DefRules}(B')$ (then there is a subargument of B' with top rule being r), if A contrary rebuts B on B'' with $\mathtt{TopRule}(B'') = r$, then A contrary rebuts any B' if $r \in \mathtt{DefRules}(B')$, and if A successfully undermines B on p, then A successfully undermines any B' with $p \in \mathtt{Prem}_{\mathtt{d}}(B')$ (since $\mathtt{Prem}_{\mathtt{d}}(A) \ntriangleleft \{p\}$ holds for contradictory undermining).

Assume that argument A based on (P, D) individually defeats B with $\mathtt{Conc}(A) = y$. Then $y \in Th_T(P, D)$ (Proposition 2). If A undercuts B, on B'' with $\mathtt{TopRule}(B'') = r$, then $y \in \overline{n(r)}$ and (P, D) individually defeats r. If A contrary rebuts B, then $\mathtt{Conc}(A)$ is a contrary of the head of a defeasible rule r in B. Then (P, D) individually defeats r. If A undermines B on $p \in \mathcal{K}_p$, then either $\mathtt{Conc}(A)$ is a contrary of p or $\mathtt{Conc}(A)$ is a contradictory of p and $\mathtt{Prem}_{\mathtt{d}}(A) \ntriangleleft \{p\}$ (argument p does not have any rules). It holds that $y \in Th_T(\mathtt{Prem}_{\mathtt{d}}(A), D)$ (with possibly $\mathtt{Prem}_{\mathtt{d}}(A) \subseteq P$). Since $\mathtt{Prem}_{\mathtt{d}}(A) \ntriangleleft \{p\}$, it holds that $\mathtt{Prem}_{\mathtt{d}}(A) \subseteq \{p' \in P \mid p' \nprec p\}$ ($\mathtt{Prem}_{\mathtt{d}}(A)$ cannot have a single ordinary premise less preferred to p, otherwise $\mathtt{Prem}_{\mathtt{d}}(A) \triangleleft \emptyset$ and in turn $A \prec p$, a contradiction to A successfully contradictory undermining p). Then (P, D) individually defeats p.

Assume now that (P, D) individually defeats x. For each element in $y \in Th_T(P, D)$ there is an argument A based on (P, D) with $\mathtt{Conc}(A) = y$ (Proposition 2). If $x = p \in \mathcal{K}_p$ then there are two cases. If there is a contrary $y \in Th_T(P, D)$ of p then A defeats all arguments B on p if $p \in \mathtt{Prem}_{\mathtt{d}}(B)$. If there is a contradictory $y \in Th(P', D)$ of p with $P' = \{p' \in P \mid p' \nprec p\}$ then there is an argument A' based on (P', D) concluding y and A' defeats all arguments B on p if $p \in \mathtt{Prem}_{\mathtt{d}}(B)$: $A' \nprec p$, due to $\mathtt{Prem}_{\mathtt{d}}(A') = P' \ntriangleleft \{p\}$ (there is no ordinary premise in A' that is less preferred to p). The remaining cases are also similar w.r.t. the other direction: if (P, D) undercuts an $r \in \mathcal{R}_d$, then A derives a $y \in \overline{n(r)}$ (then A successfully undercuts all arguments B containing r), if (P, D) contrary rebuts r, then A derives a contrary of the head of r, and defeats any argument B containing r. □

Proof of Proposition 6. Assume (P, D) contradictory rebuts (P', D') on r and B is defined as in the statement. Consider first the case that P' is non-empty. It holds that W (from Definition 15) contains an x which is a contradictory of $head(r)$. By definition, it follows that $x \in Th_T(P'', D)$ or $x \in Th_T(P, D'')$, with P'' and D'' as in Definition 15. In both cases, there exists an argument A based on either (P'', D) or (P, D'') concluding x (Proposition 2). Argument A then rebuts B on B, since the top rule of B is defeasible and its head is a contradictory of x. We claim that $A \nprec B$. It holds that both P' and D' are non-empty (former by presumption and latter due to having a defeasible rule r). We have $P'' \ntriangleleft P'$ and $D'' \ntriangleleft D'$, by construction (for each element in P'' and D'' there is one in P' and D' not more preferred). It holds that B contains both P' and D' fully (by assumption), and A contains either a subset of P'' or D'' as ordinary premises or defeasible rules. Generally, if $X \nprec Y$, then $X' \nprec Y$ if $X' \subseteq X$: otherwise there exists an $x \in X'$ s.t. x is less preferred to all $y \in Y$, but then $x \in X$, a contradiction. Thus, $\mathtt{Prem}_{\mathtt{d}}(A) \ntriangleleft \mathtt{Prem}_{\mathtt{d}}(B)$ or $\mathtt{DefRules}(A) \ntriangleleft \mathtt{DefRules}(B)$. If both $\mathtt{Prem}_{\mathtt{d}}(B)$ and $\mathtt{DefRules}(B)$ are non-empty, by definition of the weakest link principle, we infer that $A \nprec B$. This implies that A successfully contradictory rebuts B.

For the case that $P' = \emptyset$, we first (again) conclude that D' is non-empty (r is defeasible and in $\mathtt{DefRules}(B)$). The remaining reasoning as analogous to the case with P' non-empty, with the following differences. Argument A is based on (P, D''), by definition (case P' empty). Argument A contradictory rebuts B on B (as above). We again claim that $A \nprec B$. Differently to before, $\mathtt{Prem}_{\mathtt{d}}(A) \triangleleft \mathtt{Prem}_{\mathtt{d}}(B) = \emptyset$ in case $\mathtt{Prem}_{\mathtt{d}}(A) \neq \emptyset$ and otherwise $\mathtt{Prem}_{\mathtt{d}}(A) = \emptyset$ and $\mathtt{Prem}_{\mathtt{d}}(B) = \emptyset$ are incomparable w.r.t. $\triangleleft$. In the former case $A \nprec B$ iff $\mathtt{DefRules}(A) \ntriangleleft \mathtt{DefRules}(B)$ and in the latter case both arguments have no ordinary premises, and again $A \nprec B$ iff $\mathtt{DefRules}(A) \ntriangleleft \mathtt{DefRules}(B)$, by the weakest link principle. It holds that $\mathtt{DefRules}(A) \ntriangleleft \mathtt{DefRules}(B)$, since $\mathtt{DefRules}(A) \subseteq D''$ and $D'' \ntriangleleft \mathtt{DefRules}(B) = D'$ (by definition). This implies that A successfully contradictory rebuts B. □

Lemma 1. *Let T be a well-formed AT and A and B two arguments in T. If the conclusions of A and B are contrary or contradictory, then there are two arguments A' and B' both based on* $(\texttt{Prem}_\texttt{d}(A) \cup \texttt{Prem}_\texttt{d}(B), \texttt{DefRules}(A) \cup \texttt{DefRules}(B))$ *s.t. A' attacks B'.*

Proof. If $\texttt{Conc}(A)$ is a contrary of $\texttt{Conc}(B)$, then $\texttt{TopRule}(B)$ is defeasible or $B \in \mathcal{K}_p$, by assumption of T being well-formed. Then A contrary rebuts B or A undermines B, in both cases A attacks B. Consider the case that $\texttt{Conc}(A)$ and $\texttt{Conc}(B)$ are contradictory to each other. Since T is well-formed, T is strict-consistent. Thus, $\texttt{defPart}(A) \neq \emptyset$ or $\texttt{defPart}(B) \neq \emptyset$ (otherwise strict-consistency is violated). Say $\texttt{defPart}(A) \neq \emptyset$ (other case analogous). Consider the following algorithm with initially $B' = B$ and $A' = A$:

1. If $A' \in \mathcal{K}_p$ then B' attacks A' on A' and terminate.
2. Let $A' = A'_1, \dots, A'_n \rightsquigarrow \texttt{Conc}(A')$, and let $r = \texttt{TopRule}(A') = \texttt{Conc}(A'_1), \dots, \texttt{Conc}(A'_i), \dots, \texttt{Conc}(A'_n) \rightsquigarrow \texttt{Conc}(A')$.
3. If $r \in \mathcal{R}_d$, then B' attacks A' on A' and terminate.
4. If $r \in \mathcal{R}_s$, then select $A'_i \in body(r)$ with $\texttt{defPart}(A'_i) \neq \emptyset$, and let $r' = \texttt{Conc}(A'_1), \dots, \texttt{Conc}(A'_{i-1}), \texttt{Conc}(B'), \texttt{Conc}(A'_{i+1}), \dots, \texttt{Conc}(A'_n) \rightarrow -\texttt{Conc}(A'_i)$ for $\texttt{Conc}(A'_i)$ and $-\texttt{Conc}(A'_i)$ contradictories to each other.
5. Update A and B' as follows: $B' = A'_1, \dots, A'_{i-1}, B', A'_{i+1}, \dots, A_n \rightarrow -\texttt{Conc}(A'_i)$ and $A' = A'_i$. Repeat first step.

This algorithm terminates, since arguments are finite, and at each step a proper subargument of the previous iteration is selected. The algorithm terminates with a superargument B' of B s.t. all defeasible elements of B' are in $\texttt{defPart}(A) \cup \texttt{defPart}(B)$. Then B' attacks a subargument of A. □

Proof of Lemma 7. We first show the case for contradictory rebut (the case for contradictory undermining is similar and treated afterwards).

Assume that A contradictorily rebuts B on B', and $A \prec B'$, i.e., A does not successfully contradictorily rebut B on B'. By assumption that $\texttt{TopRule}(A) = r \in \mathcal{R}_s$ with $r = a_1, \dots, a_n \rightarrow b'$ and b' and $\texttt{Conc}(B')$ are contradictories to each other (A rebuts B' on B'), it follows that the immediate subarguments of A are $A_1, \dots, A_n$ with $\texttt{Conc}(A_i) = a_i$ for $1 \leq i \leq n$. Moreover, since A rebuts B', it must be that $\texttt{TopRule}(B')$ is defeasible. Because $A \prec B'$, we have $\texttt{DefRules}(A) \lhd \texttt{DefRules}(B')$, and since $\emptyset \ntriangleleft X$ for any set $X \subseteq \mathcal{R}_d$, it must hold that there is a defeasible rule in $\texttt{DefRules}(A)$, and subsequently in (at least) one A_i. Further, since $\texttt{DefRules}(A) \lhd \texttt{DefRules}(B')$, there is a rule $r' \in \texttt{DefRules}(A)$ s.t. $r' < r''$ for any $r'' \in \texttt{DefRules}(B')$ (by definition of elitist lifting). Since $\lhd$ is a strict partial order, it holds that there is an $r^* \in \texttt{DefRules}(A)$ s.t. (a) there is no rule $r''' \in \texttt{DefRules}(A)$ with $r''' < r^*$, and (b) $r^* = r'$ or $r^* < r'$. That is, r^* is strictly less preferred to all defeasible rules in $\texttt{DefRules}(B)$, and r^* is $<$-minimal among all defeasible rules in A. It holds that r^* must occur in some subargument of A, i.e., $r^* \in \texttt{DefRules}(A_i)$ for some $1 \leq i \leq n$.

We prove an auxiliary claim.

Claim: Given an $A' \in \texttt{Sub}(A)$ with $A' = A_1, \dots, A_n \rightarrow \texttt{Conc}(A')$, and a superargument C of B' (i.e., $B' \in \texttt{Sub}(C)$) s.t.

- $\texttt{Conc}(A')$ and $\texttt{Conc}(C)$ are contradictories of each other and
- there is an immediate subargument A_i of A' with $r \in \texttt{DefRules}(A_i)$ s.t. there is no rule $r' \in \texttt{DefRules}(C) \cup \texttt{DefRules}(A_1) \cup \dots \cup \texttt{DefRules}(A_{i-1}) \cup \texttt{DefRules}(A_{i+1}) \cup \dots \cup \texttt{DefRules}(A_n)$ with $r' < r$.

Then it holds that $C' = A_1, \dots, A_{i-1}, C, A_{i+1}, \dots, A_n \rightarrow x$ is an argument in T, with x and $\texttt{Conc}(A_i)$ contradictories to each other. Moreover,

- C' defeats A_i, or
- $\texttt{TopRule}(A_i) \in \mathcal{R}_s$, and there is an immediate subargument A'_j of $A_i = A'_1, \dots, A'_j, \dots, A'_m$ with $r_j \in \texttt{DefRules}(A'_j)$ s.t. there is no rule $r'' \in \texttt{DefRules}(C') \cup \texttt{DefRules}(A'_1) \cup \dots \cup \texttt{DefRules}(A'_{j-1}) \cup \texttt{DefRules}(A'_{j+1}) \cup \dots \cup \texttt{DefRules}(A'_n)$ with $r'' < r_j$.

We prove this claim. Assume that A' and C are given as stated. It holds that there is a strict rule $\texttt{Conc}(A_1), \dots, \texttt{Conc}(A_n) \rightarrow \texttt{Conc}(A')$, by assumption that A' is an argument in T. Since $\mathcal{R}_s$ is closed under transposition, and $\texttt{Conc}(A')$ and $\texttt{Conc}(C)$ are contradictories to each other, there is a strict rule $\texttt{Conc}(A_1), \dots, \texttt{Conc}(A_{i-1}), \texttt{Conc}(C), \texttt{Conc}(A_{i+1}), \dots, \texttt{Conc}(A_n) \rightarrow -\texttt{Conc}(A_i)$ with $\texttt{Conc}(A_i)$ and $-\texttt{Conc}(A_i)$ being contradictories to each other. Then there is an argument $C' = A_1, \dots, A_{i-1}, C, A_{i+1}, \dots, A_n$ in T. By assumption, for a rule $r \in \texttt{DefRules}(A_i)$ there is no rule $r' \in \texttt{DefRules}(C)$ s.t. $r' < r$. This implies that $\texttt{DefRules}(C') \ntriangleleft \texttt{DefRules}(A_i)$ and subsequently by definition of elitist lifting, $C \not\prec A_i$. If $\texttt{TopRule}(A_i)$ is defeasible, then C contradictorily rebuts A_i, and by $C \not\prec A_i$, C successfully defeats A_i. If $\texttt{TopRule}(A_i)$ is strict, then $A_i = A'_1, \dots, A'_m \rightarrow \texttt{Conc}(A_i)$. It holds that $r \in \texttt{DefRules}(A_i)$, and, thus, $r \in \texttt{DefRules}(A'_j)$ for some $1 \leq j \leq m$. By definition, $<$ is a preorder. Then there is an immediate subargument A'_k with a rule $r'' \in \texttt{DefRules}(A'_k)$ s.t. $r'' = r$ or $r'' < r$, and there is no $r''' \in \texttt{DefRules}(A_i)$ with $r''' < r''$ (a minimal rule according to $<$). For this subargument A'_k the conditions of the claim hold.

Note that A and B' from above satisfy the conditions of the claim, as shown above. Consider the following algorithm, with $X = A$ and $Y = B'$, initially:

1. compute the immediate subargument X' of X satisfying the conditions of the claim.
2. construct Y', as per claim.
3. terminate if Y' defeats X', otherwise start again with $Y = Y'$ and $X = X'$.

This algorithm terminates, since always subarguments are constructed, and arguments are finite, by definition. The algorithm has to terminate by a defeat, since there are always defeasible rules in the subarguments of X (cannot be an ordinary premise or an "empty" argument, or an argument without defeasible rules). It holds that, after termination, the argument Y' satisfies the conditions of the lemma: every subargument with a top rule being defeasible is, by construction in the claim, a subargument of either A or B.

We move to the case of contradictory undermining, i.e., A contradictorily undermines B on $B' \in \mathcal{K}_p$, but $A \prec B$. By definition, we have $\mathtt{DefRules}(A) = \mathtt{DefRules}(B') = \emptyset$ (otherwise $A \prec B$ would be impossible). This implies that there is a $p \in \mathtt{Prem_d}(A)$ with $p \prec B'$. Find some $p' \in \mathtt{Prem_d}(A)$ that is $<$-minimal and $p' < p$ or $p' = p$. By analogous reasoning as above, one can iteratively construct a superargument C' of B' s.t. C' undermines A (on p') by transposing rules of A (all rules in A are strict). Moreover, $\mathtt{Prem_d}(C') \subseteq \mathtt{Prem_d}(A) \cup \{B\}$ and $\mathtt{DefRules}(C') = \emptyset$. Also, $C' \not\prec A$, since there cannot be an ordinary premise in C' ranked lower than p' (which is $<$-minimal among all ordinary premises in $\mathtt{Prem_d}(A) \cup \{B\}$, and, thus, also in $\mathtt{Prem_d}(C')$). □

Proof of Proposition 8. If A attacks B on B', and either $A \not\prec B'$ or the attack is via undercut, contrary undermine, or contrary rebut, it immediately follows that A defeats B on B'. Assume then that $A \prec B'$ and A (unsuccessfully) undermines or rebuts B' via a contradictory.

Assume A unsuccessfully contradictory undermines B on $B' = p \in \mathtt{Prem_d}(B)$, i.e. $\mathtt{Conc}(A)$ is a contradictory of p. Since there are no strict rules, if A has no defeasible rules, then either A is an axiom or an ordinary premise. The former contradicts $A \prec B'$. In the latter case $B' = p$ contradictory undermines A, as $\mathtt{Conc}(A) = -p$ and $A \prec p$, implying $p \not\prec A$ (recall that $\prec$ is asymmetric). If A has defeasible rules, then $B' = p$ defeats A via contradictory rebut since $-p = \mathtt{Conc}(A)$, and it holds that $p \not\prec A$.

Assume A unsuccessfully contradictory rebuts B on B', i.e. $\mathtt{Conc}(A) = -\mathtt{Conc}(B')$. With similar reasoning as above, if A has no defeasible rules, then A is an ordinary premise and B' successfully contradictory undermines A (A cannot be an axiom, as stated above). If A has defeasible rules, then $\mathtt{TopRule}(A)$ is defeasible, and again $\mathtt{Conc}(A)$ and $\mathtt{Conc}(B')$ are contradictories. Then since $A \prec B'$, we have $B' \not\prec A$ and thus B' successfully contradictory rebuts A.

We now prove the second item. As before, if A attacks B on B', and either $A \not\prec B'$ or the attack is an undercut, a contrary undermining, or a contrary rebut, then A defeats B. Assume then that A contradictorily rebuts or contradictorily undermines B on B' and $A \prec B'$. If this attack is successful, the claim follows immediately. Assume that the attack is unsuccessful. By Lemma 7, it follows that there is an argument C s.t. all defeasible elements of C are part of A and B. Concretely, argument C is s.t. (i) C defeats A, and $\mathtt{Prem_d}(C) \subseteq \mathtt{Prem_d}(A) \cup \mathtt{Prem_d}(B)$ and $\mathtt{DefRules}(C) \subseteq \mathtt{DefRules}(A) \cup \mathtt{DefRules}(B)$. The claim follows. □

Proof of Proposition 9. Assume $\mathcal{E}$ is complete in F, and there is an argument A s.t. $A \notin \mathcal{E}$ and $\mathtt{defPart}(A) \subseteq \mathtt{defPart}(\mathcal{E})$. Then there is an argument B s.t. B defeats A and no argument in $\mathcal{E}$ defeats B (due to $\mathcal{E}$ being complete). If B successfully undermines A on $p \in \mathcal{K}_p$ then either (i) B contrary undermines A or (ii) B contradictory undermines A and $\mathtt{Prem_d}(B) \not\lhd \{p\}$ (it holds that $\mathtt{DefRules}(p) = \emptyset$, implying that $B \not\prec p$ is based on comparing $\mathtt{Prem_d}(B)$ and $\{p\}$; if B has no defeasible rules this is in the definition of the weakest link principle, and if B has defeasible rules then $\mathtt{DefRules}(B) \lhd \mathtt{DefRules}(p) = \emptyset$ and the comparison between ordinary premises of B and p decides the argument ranking). By assumption, there is an argument $C \in \mathcal{E}$ with $p \in \mathtt{Prem_d}(C)$. In case (i), B contrary undermines C. In case (ii) B contradictory undermines C. If $\mathcal{R}_d = \emptyset$, there can be no other types of defeats (attacks), and this implies that in the case of no defeasible rules B defeats an argument in $\mathcal{E}$, implying that some argument in $\mathcal{E}$ defeats B (due to $\mathcal{E}$ being complete), a contradiction.

We continue with $\mathcal{R}_s = \emptyset$. Then B also does not successfully undermine A (the same contradiction is implied). Then B successfully rebuts or successfully undercuts A. If B successfully undercuts A on A', then $\mathtt{TopRule}(A') = r$ and there is an argument $C \in \mathcal{E}$ with $r \in \mathtt{DefRules}(C)$. Then B successfully undercuts C, implying that $\mathcal{E}$ defeats B, a contradiction. If B successfully rebuts A on A', then $\mathtt{TopRule}(A') = r$ and there is an argument $C \in \mathcal{E}$ with $r \in \mathtt{DefRules}(C)$. If B contrary rebuts A on A', then B contrary rebuts C, implying again a contradiction. The final case is that B contradictory rebuts A on A'. It follows that B rebuts C on C' with $\mathtt{TopRule}(C') = r$ (since $r \in \mathtt{DefRules}(C)$). Argument B does not defeat C (otherwise B would defeat an argument in $\mathcal{E}$, and $\mathcal{E}$ would defeat B due to being complete), which implies that $B \prec C'$. Since $\mathcal{R}_s = \emptyset$, it follows that $\mathtt{TopRule}(B) = r'$ is defeasible or B has no defeasible rules, and the conclusions of B and C' are contradictories of each other. If B has no defeasible rules, B is either an ordinary premise or an axiom. The latter contradicts $B \prec C'$ (axioms cannot be less preferred). If B is an ordinary premise, C' (contradictory) undermines B. If B contains defeasible rules, it holds that C' (contradictory) rebuts B on B. Since $B \prec C'$ it follows that $C' \not\prec B$ ($\prec$ is asymmetric). This implies that C' defeats B. By Proposition 1, it holds that $C' \in \mathcal{E}$ because $C \in \mathcal{E}$ (closure under subarguments). This implies that $\mathcal{E}$ defeats B, a contradiction.

Suppose that $\mathcal{E}$ is a stable extension of the AF $F = (\mathcal{A}, \mathcal{D})$ corresponding to a given AT whose strict rules are closed under transposition, and that A is an argument in the AT, $A \notin \mathcal{E}$, and $\mathtt{defPart}(A) \subseteq \mathtt{defPart}(\mathcal{E})$ (i.e., this stable extensions shows that the AT is not closed under defeasible elements for stable semantics). By definition, $\exists B \in \mathcal{E}$ s.t. B defeats A on A'. First consider the case that this is due to an attack via B (i) undercutting, (ii) contrary rebutting, or (iii) contrary undermining A'. Then there is an argument $C \in \mathcal{E}$ with (i and ii) $\mathtt{TopRule}(A') \in \mathtt{DefRules}(C)$, or (iii) $A' \in \mathcal{K}_p$ and $A' \in \mathtt{Prem_d}(C)$, by construction of A

and $\mathcal{E}$. But then B attacks C and B defeats C: the preference among these arguments is not relevant, and B attacks all arguments containing $\texttt{TopRule}(A')$ (case i or ii) or if ordinary premise A' is part of the argument (case iii). The two remaining cases are that B contradictorily rebuts or contradictorily undermines A on A'. Similarly as before, there is an argument $C \in \mathcal{E}$ with $\texttt{TopRule}(A') \in \texttt{DefRules}(C)$ or $A' \in \texttt{Prem}_\texttt{d}(C)$ in the former and latter cases, respectively. Since $\mathcal{E}$ is stable, it means that B does not defeat C. By Lemma 7, it holds that there is an argument C' with $C \in \texttt{Sub}(C')$ s.t. C' defeats B and each subargument C'' of C' with a top defeasible rule or C'' being an ordinary premise, is a subargument of either B or C. Since C' defeats B, it holds that $C' \notin \mathcal{E}$, and, by definition, there is a $D \in \mathcal{E}$ s.t. D defeats C' on some $X \in \texttt{Sub}(C')$. By definition of attacks, X is either an ordinary premise or has a top rule being defeasible. Thus, X is a subargument of either B or C. By Proposition 1 (see, e.g., (Modgil and Prakken 2013)) it holds that $X \in \mathcal{E}$, a contradiction, since $D \in \mathcal{E}$ defeats $X \in \mathcal{E}$ and $\mathcal{E}$ is stable (conflict-free). We conclude that $\mathcal{E}$ is closed under defeasible elements. □

Proposition 14. *Let T be an AT. It holds that defeats are monotone w.r.t. the subargument relation, i.e., for arguments A and B of T it holds that if A defeats B, then A defeats any superargument B' of B.*

Proof. Assume A defeats B. If A successfully undercuts, undermines, or rebuts B, then A undercuts, undermines, or rebuts B on $B'' \in \texttt{Sub}(B)$, and in case of contradictory undermining or rebutting, we additionally have $A \not\prec B''$. Since $B \in \texttt{Sub}(B')$, we have $B'' \in \texttt{Sub}(B)$. This implies that A successfully undercuts, undermines, or rebuts B' on B'', and in turn A defeats B'. □

Proof of Theorem 10. Within this proof, we shorten "w-stable" to "stable" (i.e., all stable assumptions mean w-stable assumptions). For the first item, assume that (P, D) is a stable assumption and let $\mathcal{E} = \{A \mid A$ based on (P, D) in $T\}$. Suppose for contradiction that $\mathcal{E}$ is not stable. Due to the first item of Definition 16, $\mathcal{E}$ is conflict-free. Suppose the contrary, then there are $A, B \in \mathcal{E}$ s.t. A defeats B. This implies that (i) A undermines B, (ii) A undercuts B, or (iii) A rebuts B. In case (i) $\texttt{Conc}(A)$ and $p \in \texttt{Prem}_\texttt{d}(B)$ are contrary or contradictory, in case (ii) $\texttt{Conc}(A)$ is contrary or contradictory to an $n(r)$ with $r \in \texttt{DefRules}(B)$, in case (iii) $\texttt{Conc}(A)$ is contrary or contradictory to $\mathit{head}(r)$ for an $r \in \texttt{DefRules}(B)$. In all three cases a violation of condition 1 of Definition 16 is present. We conclude that $\mathcal{E}$ is conflict-free, and there is an argument A of T such that $A \notin \mathcal{E}$ and $\mathcal{E}$ does not defeat A. If there is an $x \in \texttt{defPart}(A)$ that (P, D) individually defeats, there is by Proposition 5 an argument B based on (P, D) that individually defeats A. By presumption, $B \in \mathcal{E}$, contradicting $\mathcal{E}$ not defeating A. Thus $\texttt{DefRules}(A) \subseteq D_U$ where D_U are the rules in $\mathcal{R}_d$ that (P, D) does not individually defeat. Due to the second item of Definition 16, the set of premises not defeated by (P, D) equals P, and thus $\texttt{Prem}_\texttt{d}(A) \subseteq P$. Moreover, as $A \notin \mathcal{E}$, it must be that $\texttt{DefRules}(A) \nsubseteq D$. Now we find that $(\texttt{Prem}_\texttt{d}(A), \texttt{DefRules}(A))$ is one such (P', D') that by Definition 16 is contradictory rebutted by (P, D) on some $r \in \texttt{DefRules}(A)$.

We prove auxiliary results first. If B is an argument in T and (P, D) contradictory rebuts $(\texttt{Prem}_\texttt{d}(B), \texttt{DefRules}(B))$ on some $r \in \texttt{DefRules}(B)$, then $r \notin D$. Suppose the contrary, i.e., $r \in D$. All rules in D are applicable by (P, D), implying that there is an argument C based on (P, D) with $r \in \texttt{DefRules}(C)$. It holds that $C \in \mathcal{E}$ ($\mathcal{E}$ contains all arguments based on (P, D) by construction). By Definition 15 and Proposition 2, if (P, D) contradictory rebuts $(\texttt{Prem}_\texttt{d}(B), \texttt{DefRules}(B))$ on r, then there is an argument X based on (P, D) s.t. X concludes the contradictory of $\mathit{head}(r)$ (in Definition 15 if a contradictory is derived from W or from (P'', D), then the same contradictory is derivable from (P, D) as derivability is monotone, and then there is an argument based on (P, D) concluding this contradictory). Both arguments X and C are based on (P, D), but then item 1 of Definition 16 is violated: two contradictory atoms are derivable from (P, D).

Claim: If (P, D) contradictory rebuts $(\texttt{Prem}_\texttt{d}(A_i), \texttt{DefRules}(A_i))$ on $r_i \in \texttt{DefRules}(A_i)$ with $r_i \notin D$ and argument A_i s.t. $\texttt{Prem}_\texttt{d}(A_i) \subseteq P$, $\texttt{DefRules}(A_i) \subseteq D_U$, $\texttt{DefRules}(A_i) \nsubseteq D$ then it holds that

- (P, D) contradictory rebuts $(\texttt{Prem}_\texttt{d}(A_i), \texttt{DefRules}(A_i))$ on $\texttt{TopRule}(A_i)$ or
- there is a proper subargument A_{i+1} of A_i (i.e., $A_{i+1} \neq A_i$ and $A_{i+1} \in \texttt{Sub}(A_i)$) s.t. (P, D) contradictory rebuts $(\texttt{Prem}_\texttt{d}(A_{i+1}), \texttt{DefRules}(A_{i+1}))$ on $r_{i+1} \notin D$ and $\texttt{TopRule}(A_{i+1}) = r_i$.

Assume that the first item does not hold, i.e., (P, D) does not contradictory rebut $(\texttt{Prem}_\texttt{d}(A_i), \texttt{DefRules}(A_i))$ on $\texttt{TopRule}(A_i)$. Then (P, D) contradictory rebuts $(\texttt{Prem}_\texttt{d}(A_i), \texttt{DefRules}(A_i))$ on $r_i \neq \texttt{TopRule}(A_i)$. Then there is a subargument A_{i+1} with $\texttt{TopRule}(A_{i+1}) = r_i$. It holds that $\texttt{Prem}_\texttt{d}(A_{i+1}) \subseteq \texttt{Prem}_\texttt{d}(A_i) \subseteq P$, $\texttt{DefRules}(A_{i+1}) \subseteq \texttt{DefRules}(A_i) \subseteq D_U$, and $r_i \notin D$. Then, by assumption of (P, D) being stable, it follows that (P, D) contradictory rebuts $(\texttt{Prem}_\texttt{d}(A_{i+1}), \texttt{DefRules}(A_{i+1}))$ on some r_{i+1}. It cannot be that $r_{i+1} \in D$, as shown above. Thus, $r_{i+1} \notin D$. This proves the claim. Note that $(\texttt{Prem}_\texttt{d}(A_{i+1}), \texttt{DefRules}(A_{i+1}))$ satisfies the condition of the claim. It holds that (P, D) contradictory rebuts $(\texttt{Prem}_\texttt{d}(A_1), \texttt{DefRules}(A_1))$ with $A_1 = A$ on an $r_1 \in \texttt{DefRules}(A_1)$, and the conditions of the claim hold. Consider the following algorithm: iterate through subarguments A_{i+1} of A_i (via the second item of the claim) until (P, D) contradictory rebuts $(\texttt{Prem}_\texttt{d}(A_n), \texttt{DefRules}(A_n))$ on $\texttt{TopRule}(A_n)$. This algorithm terminates, since $\texttt{Sub}(A)$ is finite (by definition of arguments). Then there is an argument based on (P, D) that successfully contradictory rebuts A_n, and, via Proposition 14, all superarguments of A_n, in particular $A_1 = A$. Since $\mathcal{E}$ contains all arguments based on (P, D), $\mathcal{E}$ defeats A, a contradiction. Therefore if (P, D) is a stable assumption, $\mathcal{E}$ is a stable extension.

For the second item, assume that $\mathcal{E}$ is a stable extension of F. Recall that $\mathcal{E}$ contains all arguments based on (P, D), by Proposition 9. For contradiction, suppose that (P, D) is not a stable assumption of T, where $P = \mathtt{Prem_d}(\mathcal{E})$ and $D = \mathtt{DefRules}(\mathcal{E})$. If (contrary to Item 1 of Definition 16) $\exists x, y \in Th_T(P, D)$ such that $x \in \overline{y}$ or $-y = x$, then there are arguments A and B, based on (P, D) (and in $\mathcal{E}$) s.t. $\mathtt{Conc}(A) = x$ and $\mathtt{Conc}(B) = y$. By Lemma 1, it holds that there are arguments A' and B', based on (P, D), s.t. A' attacks B', and both are in $\mathcal{E}$. By Proposition 8, A' defeats B' or some argument B'' based on (P, D) defeats A' or B'. By Proposition 9 $A', B', B'' \in \mathcal{E}$, contradicting $\mathcal{E}$ being stable. If x is contrary/contradictory of a name of a defeasible rule in D, then there is an argument based on (P, D) that concludes x and defeats (successfully undercuts) each argument containing the rule. There is an argument containing the rule in $\mathcal{E}$ (by construction of (P, D)). Suppose, on the other hand (contrary to Item 2 of Definition 16), $\exists p \in \mathcal{K}_p \setminus P$ such that (P, D) does not individually defeat p. Since (P, D) does not individually defeat p, by Proposition 5 there is a subargument of p that no argument based on (P, D) defeats, with p itself being the only possibility for such an argument. Thus there can be no argument in $\mathcal{E}$ that defeats p, and by presumption $p \notin \mathcal{E}$, which contradicts $\mathcal{E}$ being stable. Lastly suppose (contrary to Item 3 of Definition 16) that there is an assumption (P', D') such that each rule in D' is applicable by (P', D'), $D' \nsubseteq D$, $P' \subseteq P$, and $D' \subseteq D_U$, where D_U contains all defeasible rules that are not individually defeated by (P, D), but (P, D) does not contradictorily rebut (P', D') on any $r \in D'$. Consider any argument A based on (P', D') with $r = \mathtt{TopRule}(A)$ and $r \in D' \setminus D$ (exists since all rules in D' are applicable by (P', D')). First consider the case that P' is non-empty. It holds by definition that it is not possible to derive a contradictory of $head(r)$ for any $r \in D'$ from neither (P'', D) nor (P, D'') such that $P'' = \{p \in P \mid \exists p' \in \mathtt{Prem_d}(A), p \not< p'\}$ and $D'' = \{r \in D \mid \exists r' \in \mathtt{DefRules}(A), r \not< r'\}$. This implies that there is no argument B based on (P, D) with $\mathtt{Conc}(B)$ a contradictory of $head(r)$ such that $B \not\prec A$: if there is a B based on (P, D) concluding a contradictory of $head(r)$ it cannot be based on either (P, D'') or (P'', D), then both the premises and defeasible rules of B contain an element that is less preferred to all elements in the corresponding set in (P', D'), in which case $B \prec A$ by definition. The case with P' empty is analogous, except that then B contains a defeasible rule in $D \setminus D''$, and either an ordinary premise or no premise. In all these cases $B \prec A$ (A contains no ordinary premises, so if B contains one then the ordinary premise set is less preferred to B, then only comparison on defeasible rule set is relevant, by weakest link principle; in both cases the comparison between sets of defeasible rules decides the preference and B contains a defeasible rule strictly less preferred to all defeasible rules in A). $\mathcal{E}$ contains only arguments that are based on (P, D) by presumption, and thus $\mathcal{E}$ does not defeat A (no successful contradictory rebuts by reasoning above, and no individual defeats by construction of D_U and P), contradicting $\mathcal{E}$ being a stable extension. Therefore if $\mathcal{E}$ is a stable extension, (P, D) is a stable assumption. □

Proof of Proposition 11. Since $\mathcal{R}_d = \emptyset$, if any A in T defeats any B in T, it holds that A successfully contrary undermines B or A successfully contradictory undermines B. This implies that A individually defeats B (there are only individual defeats in T). Moreover, any defeat means that A undermines B on a $p \in \mathtt{Prem_d}(B)$. Let P be s-stable. Consider $\mathcal{E} = \{A \mid A \text{ based on } (P, \emptyset) \text{ in } T\}$. Suppose $\mathcal{E}$ is not stable. Then $\mathcal{E}$ is not conflict-free or there is an argument in T not in $\mathcal{E}$ and not defeated by $\mathcal{E}$. If $\mathcal{E}$ is not conflict-free, then there are two arguments $A, B \in \mathcal{E}$ s.t. A defeats B on some $p \in \mathtt{Prem_d}(B)$. Then A individually defeats B. By Proposition 5, $(P, \emptyset)$ individually $p \in P$, a contradiction. Suppose that there is an argument $A \notin \mathcal{E}$ in T that is not defeated by $\mathcal{E}$. Then $(P, \emptyset)$ does not individually defeat any $\mathtt{Prem_d}(A)$, which contains an ordinary premise not in P. We conclude that $\mathcal{E}$ is stable in F.

Assume that $\mathcal{E}$ is stable in F, and let $P = \mathtt{Prem_d}(\mathcal{E})$. Suppose P is not s-stable. If $(P, \emptyset)$ individually defeats an $p \in P$, then there is an argument A based on $(P, \emptyset)$ s.t. A individually defeats any argument B with $p \in \mathtt{Prem_d}(B)$. It holds that argument $p \in \mathcal{E}$, implied by both Proposition 9 and by Proposition 1. This contradicts $\mathcal{E}$ being conflict-free. Suppose $(P, \emptyset)$ does not individually defeat a $p \in \mathcal{K}_p \setminus P$. Then there is no argument based on $(P, \emptyset)$ that individually defeats argument p. It holds that $\mathcal{E}$ contains only arguments based on $(P, \emptyset)$. Moreover, $p \notin \mathcal{E}$, a contradiction. □

Proof of Proposition 12. The proof for the two cases of no strict or no defeasible rules is analogous, we show the case for $\mathcal{R}_s = \emptyset$, i.e., for w-stable assumptions. It holds that x is credulously justified in T under stable semantics iff there is a stable extension $\mathcal{E}$ with an argument concluding x iff (by Theorem 10) there is a w-stable assumption $(\mathtt{Prem_d}(\mathcal{E}), \mathtt{DefRules}(\mathcal{E}))$ with (by Proposition 2) $x \in Th_T(\mathtt{Prem_d}(\mathcal{E}), \mathtt{DefRules}(\mathcal{E}))$.

It holds that x is not skeptically justified in T under stable semantics iff there is a stable extension $\mathcal{E}$ of T without an argument concluding x iff there is a w-stable assumption $(\mathtt{Prem_d}(\mathcal{E}), \mathtt{DefRules}(\mathcal{E}))$ with $x \notin Th_T(\mathtt{Prem_d}(\mathcal{E}), \mathtt{DefRules}(\mathcal{E}))$. □

Proof of Theorem 3. Given an assumption (P, D), checking s-stability can be achieved by checking for each $p \in P$ and $p \in \mathcal{K}_p \setminus P$ whether the former are not individually defeated by (P, D) and the latter are individually defeated by (P, D). For credulous reasoning, first perform a non-deterministic construction of a (P, D) assumption, check s-stability, and compute $Th_T(P, D)$, and for skeptical reasoning consider the complementary problem and guess an assumption, check s-stability, and again compute the deductive closure. By Proposition 12, this computation decides the corresponding decision problems. For hardness, the reduction provided by (Lehtonen, Wallner, and Järvisalo 2020) (Proposition 7) applies here (constructed AT has no preferences and only ordinary premises). □

Proof of Theorem 13. For membership in coNP, consider the complementary problem: checking whether a given (P, D) assumption is not w-stable. Conditions 1. and 2. of Definition 16 (and applicability of rules) can be checked in polynomial time. Perform a non-deterministic construction of an assumption (P', D'). Pre-conditions of condition 3. can be checked in polynomial time. Checking whether (P, D) contradictory rebuts (P', D') on some $r \in D'$ can be done in polynomial time (the corresponding checks rely on computing one or two deductive closures). If (P, D) does not contradictory rebut (P', D') on some $r \in D'$, then the given assumption is not w-stable.

Let $\phi = c_1, \ldots, c_m$ be a Boolean formula in conjunctive normal form (CNF) with clauses $C = \{c_1, \ldots, c_m\}$ of the form $c_i = l_{i,1} \vee l_{i,2} \vee l_{i,3}$ over vocabulary $X = \{x_1, \ldots, x_n\}$. For a set X let $\neg X = \{\neg x \mid x \in X\}$. We note that "$\neg x$" if used in an AT is *one symbol* within this proof, and a negated literal in the Boolean formula. Construct AT $T = (\mathcal{L}, \mathcal{R}, n, ^-, \mathcal{K}, \leq)$ as follows.

$$\begin{aligned}
\mathcal{K}_p &= X \cup \neg X \cup \{a_i \mid x_i \in X\} \cup \{\neg a_i \mid x_i \in X\} \\
\mathcal{L} &= C \cup \{f, f'\} \cup \mathcal{K}_p \\
\mathcal{R}_d &= \{r_{i,j} : l_{i,j} \Rightarrow c_i \mid c_i \in C, 1 \leq j \leq 3\} \cup \\
&\quad \{r_{f,i} : a_i, \overline{a_i} \Rightarrow f' \mid x_i \in X\} \cup \\
&\quad \{r_f : c_1, \ldots, c_m \Rightarrow f\} \\
&\quad \overline{f} = \{f'\}, \overline{f'} = \{f\} \\
&\quad r_{f,i} < r_f, \forall 1 \leq i \leq n \\
&\quad r_{f,i} < r_{j,t}, \forall 1 \leq i \leq n, \forall 1 \leq j \leq m, \forall 1 \leq t \leq 3 \\
&\quad a_i < x_i, \neg a_i < \neg x_i, \forall 1 \leq i \leq n \\
&\quad a_i < x_j, a_i < \neg x_j, \forall 1 \leq i, j \leq n \text{ with } i \neq j \\
&\quad \neg a_i < x_j, \neg a_i < \neg x_j, \forall 1 \leq i, j \leq n \text{ with } i \neq j
\end{aligned}$$

We claim that the set of arguments $\mathcal{E} = \{A \mid A \text{ argument in } T, \mathtt{defPart}(A) \subseteq \mathcal{K}_p \cup (\mathcal{R}_d \setminus \{r_f\})\}$ is stable iff ϕ is unsatisfiable. That is, $(\mathcal{K}_p, \mathcal{R}_d \setminus \{r_f\})$ is w-stable iff ϕ is unsatisfiable. AT T can be constructed in polynomial time w.r.t. the size of ϕ.

First, observe that $\mathcal{E}$ contains all arguments of T except for those that include defeasible rule r_f. If an argument A includes rule r_f, then $\mathtt{TopRule}(A) = r_f$ (no further derivations possible). Moreover, A includes one rule $r_{i,j}$ for each clause (for each $1 \leq i \leq m$), and includes some subset of $X \cup \neg X$ as ordinary premises.

For an argument A containing rule r_f we say that A is inconsistent if there is an i s.t. $\{x_i, \neg x_i\} \subseteq \mathtt{Prem}_{\mathtt{d}}(A)$ (i.e., A is not a partial truth value assignment on X), otherwise we say that A is consistent. Assume that A is inconsistent, and there is an i s.t. $\{x_i, \neg x_i\} \subseteq \mathtt{Prem}_{\mathtt{d}}(A)$. Consider argument $B = a_i, \neg a_i \Rightarrow f'$, which is in $\mathcal{E}$. Argument B rebuts A, by construction. Since $\mathtt{DefRules}(B) = r_{f,i}$ and this rule is strictly less preferred to all defeasible rules in A (actually all other than $r_{f,j}$ for some j), it follows that $\mathtt{DefRules}(B) \lhd \mathtt{DefRules}(A)$. We have $\mathtt{Prem}_{\mathtt{d}}(B) = \{a_i, \neg a_i\}$. By construction, it holds that $a_i \not< \neg x_i$ and $\neg a_i \not< x_i$. This implies (under the Elitist lifting) that $\mathtt{Prem}_{\mathtt{d}}(B) \not\lhd \mathtt{Prem}_{\mathtt{d}}(A)$, and $B \not\prec A$. Thus, B defeats A. Now assume that A is consistent. Only arguments B with $\mathtt{TopRule}(B) = r_{f,i}$ attack (rebut) A (only f and f' are contradictories in the AT T). Consider an arbitrary such $B = a_j, \neg a_j \Rightarrow f'$. Since A is consistent, either x_j or $\neg x_j$ is not in $\mathtt{Prem}_{\mathtt{d}}(A)$. Say $\neg x_j$ is not in $\mathtt{Prem}_{\mathtt{d}}(A)$ (other case analogous). Then $a_j < z$ for all $z \in \mathtt{Prem}_{\mathtt{d}}(A)$, since $a_j < x_j$ and $a_j < z$ for all $z \in X \cup ((\neg X) \setminus \{\neg x_j\})$ (a_j is strictly less preferred to all ordinary premises in A except for $\neg x_j$ which does not occur in A). By the same reasoning as above $\mathtt{DefRules}(B) \lhd \mathtt{DefRules}(A)$, and moreover, $\mathtt{Prem}_{\mathtt{d}}(B) \lhd \mathtt{Prem}_{\mathtt{d}}(A)$. Thus, $B \prec A$, and B does not defeat A. Since B was arbitrary, no argument in $\mathcal{E}$ defeats A.

Assume that ϕ is unsatisfiable. Then there are no consistent arguments A concluding f. Otherwise, A represents a partial truth value assignment on X (via $\mathtt{Prem}_{\mathtt{d}}(A)$) s.t. each clause is satisfied (by including rules $r_{i,j}$ for each clause), implying that ϕ is satisfiable. Then $\mathcal{E}$ is stable: this set does not defeat itself (only attacks are outside to arguments concluding f), and by the reasoning above, all arguments outside are inconsistent, by assumption that ϕ is unsatisfiable. Assume that ϕ is satisfiable. Then there is an argument outside $\mathcal{E}$ that is consistent and concludes f. By the reasoning above $\mathcal{E}$ does not defeat this specific argument. □

Proof of Theorem 4. For membership in Σ_2^P for both credulous and skeptical justification, consider a non-deterministic construction of a (P, D) assumption. Check w-stability (in coNP) via an NP oracle. Finally, check whether $Th_T(P, D)$ includes the queried atom. By Proposition 12, the queried atom can be derived iff credulous acceptance holds.

We show hardness for credulous justification first. Let $\phi = c_1, \ldots, c_m$ be a Boolean formula in conjunctive normal form (CNF) with clauses $C = \{c_1, \ldots, c_m\}$ of the form $c_i = l_{i,1} \vee l_{i,2} \vee l_{i,3}$ over vocabulary $X \cup Y$ with $X = \{x_1, \ldots, x_s\}$ and $Y = \{y_1, \ldots, y_n\}$. For a set X let $\neg X = \{\neg x \mid x \in X\}$. We note that "$\neg x$" if used in an AT is *one symbol* within this proof, and a negated literal in the Boolean formula. Construct AT $T = (\mathcal{L}, \mathcal{R}, n, ^-, \mathcal{K}, \leq)$ as follows.

$$\begin{aligned}
\mathcal{K}_p &= X \cup \neg X \cup Y \cup \neg Y \cup \{a_i \mid x_i \in X\} \cup \{\neg a_i \mid x_i \in X\} \\
\mathcal{L} &= C \cup \{f, f'\} \cup \mathcal{K}_p \\
\mathcal{R}_d &= \{r_{i,j} : l_j \Rightarrow c_i \mid c \in C, 1 \leq j \leq 3\} \cup \\
&\quad \{r_{f,i} : a_i, \neg a_i, g_1, \ldots, g_s \Rightarrow f' \mid y_i \in X\} \cup \\
&\quad \{r_f : c_1, \ldots, c_m \Rightarrow f\} \\
&\quad \{r_{g,i,1} : x_i \Rightarrow g_i \mid x_i \in X\} \cup \{r_{g,i,2} : \neg x_i \Rightarrow g_i \mid x_i \in X\} \\
&\quad \overline{f} = \{f'\}, \overline{f'} = \{f\}, \overline{x} = \{\neg x\}, \overline{\neg x} = \{x\} \forall x \in X \\
&\quad r_{f,i} < r_f, \forall 1 \leq i \leq n \\
&\quad r_{f,i} < r_{j,t}, \forall 1 \leq i \leq n, \forall 1 \leq j \leq m, \forall 1 \leq t \leq 3 \\
&\quad r_{g,i,w} < r_f, \forall 1 \leq i \leq s, w \in \{1, 2\} \\
&\quad r_{g,i,w} < r_{j,t}, \\
&\qquad \forall 1 \leq i \leq n, \forall 1 \leq j \leq m, \forall 1 \leq t \leq 3, w \in \{1, 2\} \\
&\quad a_i < y_i, \neg a_i < \neg y_i, \forall 1 \leq i \leq n \\
&\quad a_i < y_j, a_i < \neg y_j, \forall 1 \leq i, j \leq n \text{ with } i \neq j \\
&\quad \neg a_i < y_j, \neg a_i < \neg y_j, \forall 1 \leq i, j \leq n \text{ with } i \neq j \\
&\quad a_i < z, \neg a_i < z, \text{ with } z \in X \cup \neg X
\end{aligned}$$

We claim that there is a stable extension with a conclusion f' iff there is partial truth value assignment τ_X (defined only on X) s.t. $\phi[\tau_X]$ is unsatisfiable, with $\phi[\tau_X]$ being the formula ϕ where each clause c_i removed if τ_X satisfies c_i and if a variable in X occurs in a clause, the corresponding literal is removed. That is, f' is credulously justified under stable semantics iff there is an assignment on X s.t. for every assignment on Y ϕ is refuted. AT T can be constructed in polynomial time w.r.t. the size of ϕ.

Assume that there is an assignment τ_X on X s.t. for every completion of τ_X to τ (including Y) ϕ is refuted. Construct (P, D) with

$$P = \mathcal{K}_p \setminus (\{x \in X \mid \tau(x) = 0\} \cup \{\neg x \in \neg X \mid \tau(x) = 1\}$$
$$D = \mathcal{R}_d \setminus (\{r_f\} \cup \{r_{g,i,1} \mid \tau(x_i) = 0\} \cup \{r_{g,i,2} \mid \tau(x_i) = 1\}))$$

That is, P contains all ordinary premises, except those that are assigned differently by τ_X (if $\tau(x_i) = 1$ then x_i is in P and $\neg x_i$ is not), and D contains all defeasible rules except r_f and depending on τ_X rule $r_{g,i,1}$ is excluded if $\tau(x_i) = 0$ and rule $r_{g,i,2}$ is excluded if $\tau(x_i) = 1$. Moreover, let $\mathcal{E} = \{A \mid A \text{ based on } (P, D)\}$. It holds that $f' \in Th_T(P, D)$ and $f' \in \texttt{Conc}(\mathcal{E})$. We show that $\mathcal{E}$ is stable in the corresponding AF $F = (\mathcal{A}, \mathcal{D})$ to T. AT T contains no possibilities for undercuts, and all rebuts or underminings are of the contradictory sort. Since rule r_f is not contained in D, it holds that $\mathcal{E}$ contains no argument concluding f. There is no rule concluding an ordinary premise in $X \cup \neg X$. Moreover, if $x_i \in P$ ($\neg x_i \in P$) then $\neg x_i \notin P$ ($x_i \notin P$). Then $\mathcal{E}$ is conflict-free: there is no possibility for rebuts or undercuts (from (P, D) one cannot derive a contradictory part of P or derived via D). It remains to show that if $A \notin \mathcal{E}$ it holds that $\mathcal{E}$ defeats A. First consider that A is an ordinary premise not in P. By construction, there is an argument in $\mathcal{E}$ defeating A. If A does not contain the rule r_f, then $A = A' \Rightarrow g_i$ for some i and subargument A' (by construction of D). Then $\mathcal{E}$ successfully undermines A on A' (which is an ordinary premise not in P). Thus, A contains rule r_f. As in the proof of Proposition 13, we say that an argument B is inconsistent if for some i, $1 \leq i \leq n$ it holds that $y_i, \neg y_i \in \texttt{Prem}_{\texttt{d}}(B)$. Consider two cases: there is an $x_i \in X$ s.t. $\tau_X(x_i) = 1$ s.t. $\neg x_i \in \texttt{Prem}_{\texttt{d}}(A)$ or $\tau_X(x_i) = 0$ s.t. $x_i \in \texttt{Prem}_{\texttt{d}}(A)$ (i.e., $\texttt{Prem}_{\texttt{d}}(A)$ follows the truth assignment τ_X or not), or this is not the case.

- Assume that there is an $x_i \in X$ s.t. $\tau_X(x_i) = 1$ s.t. $\neg x_i \in \texttt{Prem}_{\texttt{d}}(A)$ or $\tau_X(x_i) = 0$ s.t. $x_i \in \texttt{Prem}_{\texttt{d}}(A)$. Consider the case that $x_i \in X$ s.t. $\tau_X(x_i) = 1$s.t. $\neg x_i \in \texttt{Prem}_{\texttt{d}}(A)$ (other case analogous). Since $\tau(x_i) = 1$, it holds that $x_i \in P$ and $\neg x_i \notin P$ (by construction). Then there is an argument $B = x_i \in \mathcal{E}$. It holds that B successfully contradictory undermines A (x_i and $\neg x_i$ are incomparable w.r.t. $\leq$). Thus, $\mathcal{E}$ defeats A.
- Assume that there is no $x_i \in X$ s.t. $\tau_X(x_i) = 1$ s.t. $\neg x_i \in \texttt{Prem}_{\texttt{d}}(A)$ or $\tau_X(x_i) = 0$ s.t. $x_i \in \texttt{Prem}_{\texttt{d}}(A)$. Then A follows τ_X in the sense that if x_i is assigned true then A does not contain $\neg x_i$ and if x_i is assigned false then A does not contain x_i. It can be the case that A contains neither. Consider two sub cases.
  - Argument A is inconsistent. It must be that A is inconsistent because there is an i s.t. $y_i, \neg y_i \in \texttt{Prem}_{\texttt{d}}(A)$. Consider argument B with $\texttt{TopRule}(B) = r_{f,i}$, i.e., the top rule is $r_{f,i} : a_i, \neg a_i, g_1, \ldots, g_s \Rightarrow f'$. This argument exists, because all ordinary premises (or ordinary premises used to derive the g_j's) are part of P. It holds that $a_i \not< \neg y_i$ and $\neg a_i \not< y_i$. Thus, $\texttt{Prem}_{\texttt{d}}(B) \not\triangleleft \texttt{Prem}_{\texttt{d}}(A)$, and $B \not\prec A$. Then B defeats A.
  - Argument A is consistent. Then we arrive directly at a contradiction. Consider $\texttt{Prem}_{\texttt{d}}(A)$. It holds that $\texttt{Prem}_{\texttt{d}}(A)$ represents a partial truth value assignment on $X \cup Y$ s.t. ϕ is satisfied. That is, for the presumed assignment on X there is a completion to $X \cup Y$ s.t. ϕ is satisfied, contradicting the initial assumption.

Thus, $\mathcal{E}$ defeats A if the initial assumption holds. Then $\mathcal{E}$ is stable, and f' is credulously justified under stable semantics.

Assume now that f' is credulously justified under stable semantics. Then there is a stable extension $\mathcal{E}$ of F s.t. $f' \in \texttt{Conc}(\mathcal{E})$. Consider an argument A in $\mathcal{E}$ that concludes f'. Then $\texttt{TopRule}(A)$ is $r_{f,i}$ for some i. By construction, $\texttt{Prem}_{\texttt{d}}(A)$ contains for each $x_i \in X$ either x_i or $\neg x_i$. By Proposition 9, it holds that for each such x_i ($\neg x_i$) there is an argument x_i ($\neg x_i$) in $\mathcal{E}$ (and due to conflict-freeness, this does not hold for the complementary literals, i.e., if $x_i \in \texttt{Prem}_{\texttt{d}}(A)$ then $\neg x_i$ is not part of any argument in $\mathcal{E}$). Consider τ_X s.t. $\tau_X(x_i)$ assigned true if $x_i \in \texttt{Prem}_{\texttt{d}}(A)$ and false if $\neg x_i \in \texttt{Prem}_{\texttt{d}}(A)$. Consider any completion of τ_X to Y, leading to τ. If τ satisfies ϕ, then there is a *consistent* argument B s.t. $\texttt{Prem}_{\texttt{d}}(B) \subseteq X \cup \neg X \cup Y \cup \neg Y$ and $\texttt{Conc}(B) = f$ via top rule r_f. To see this, r_f derives f via subarguments deriving g_i and $x_i \in X$ or $x_i, \neg x_i$, and each clause is satisfied (by assumption that τ satisfies ϕ). Since argument B is consistent, there is no argument in $\mathcal{E}$ that defeats B. Suppose the contrary, i.e., there is an argument $C \in \mathcal{E}$ that defeats B. Then B does not undermine B (by reasoning above, the contradictories of x_i or $\neg x_i$ are not part of $\mathcal{E}$. Then C contradictory rebuts B, via $\texttt{TopRule}(C) = r_{f,i}$ for some i. It holds that either y_i or $\neg y_i$ is not in $\texttt{Prem}_{\texttt{d}}(B)$. Consider first the case that one of them is in $\texttt{Prem}_{\texttt{d}}(B)$. Then $a_i < y_i$ in the former case and $\neg a_i < \neg y_i$ in the latter case. Moreover, both a_i and $\neg a_i$ are strictly less preferred to all other ordinary premises in $\texttt{Prem}_{\texttt{d}}(B)$. Then $\texttt{Prem}_{\texttt{d}}(C) \triangleleft \texttt{Prem}_{\texttt{d}}(B)$. Moreover, $\texttt{DefRules}(C) \triangleleft \texttt{DefRules}(B)$, since C contains rules of type $r_{f,i}$ and $r_{g,i,w}$ and B contains rules of type r_f and $r_{i,j}$, the former are all less preferred to the latter. Then $C \prec B$ and C does not defeat B. Then $\mathcal{E}$ does not defeat B. It cannot be that $B \in \mathcal{E}$: Then B attacks (rebuts) some argument in $\mathcal{E}$, and via Proposition 8 and Proposition 9 we arrive at the contradiction that $\mathcal{E}$ is not conflict-free.

For skeptical justification, we reduce credulous justification to non-skeptical justification. Let $T = (\mathcal{L}, \mathcal{R}, n, ^-, \mathcal{K}, \leq)$ be an AT, and $s \in \mathcal{L}$. Construct AT T' by T' being the same as T, except for (i) an additional ordinary premise $p \notin \mathcal{L}$, and $\overline{p} = \{s\}$. Consider the corresponding AFs $F = (\mathcal{A}, \mathcal{D})$ to T and $F' = (\mathcal{A}', \mathcal{D}')$ to T'. It holds that $\mathcal{A} = \mathcal{A} \cup \{(p)\}$, since T' gives rise to the same arguments as T, and additionally argument $A = p$. Regarding defeats, all arguments with $\texttt{Conc}(s)$ attack, and defeat, argument A (contrary undermining). Recall that stable extensions are closed under subarguments (Proposition 1). If a set of arguments is stable in F and conclusion s is among the conclusions in this extension, then there is a stable extension in F' concluding s and defeating p. Vice versa, a stable extension concluding s in F' implies existence of a stable extension (in fact the same) in F. It holds that

$$\begin{aligned} & s \text{ is credulously justified in } T \\ & \text{iff } \exists \mathcal{E} \in \mathit{stb}(F), s \in \mathtt{Conc}(\mathcal{E}) \\ & \text{iff } \exists \mathcal{E}' \in \mathit{stb}(F'), s \in \mathtt{Conc}(\mathcal{E}') \\ (*)\ & \text{iff } \exists \mathcal{E}' \in \mathit{stb}(F'), p \notin \mathtt{Conc}(\mathcal{E}') \\ & \text{iff } p \text{ is not skeptically justified in } T'. \end{aligned}$$

The "iff" $(*)$ holds, since a stable extension $\mathcal{E}'$ in F' with conclusion s must necessary defeat argument $A = p$, and, since p is fresh in T', no other argument can conclude p. This implies that p is not concluded in $\mathcal{E}'$. Conversely, if there is a stable extension $\mathcal{E}'$ in F' not concluding p, then argument $A = p$ must be defeated by $\mathcal{E}'$ (with arguments concluding s being the only possibility of defeating p). This is a reduction from "yes" instances of the credulous justification problem to "no" instances of the skeptical justification problem, under stable semantics. Thus, skeptical justification under stable semantics is Π_2^P hard. □

Examples

Example 1. *Let* $T = (\mathcal{L}, \mathcal{R}, n, ^-, \mathcal{K}, \leq)$ *be an AT with* $\mathcal{L} = \{a, b, c, d, e, p, q, q', x, y, y', z\}$, $\mathcal{K}_p = \{a, b, c, d, e\}$, $\mathcal{K}_n = \emptyset$, $\overline{y} = \{y'\}$, $\overline{y'} = \{y\}$, $\overline{q} = \{q'\}$, $\overline{q'} = \{q\}$, $\overline{p} = \{p'\}$, $\overline{p'} = \{p\}$

$$\begin{aligned} \mathcal{R} = \{ & r_1 : q, p \to y', \\ & r_2 : a \to x, \\ & r_3 : b \to x, \\ & r_4 : x \Rightarrow y, \\ & r_5 : c \Rightarrow q, \\ & r_6 : d \Rightarrow p, \\ & r_7 : e \Rightarrow q', \\ & r_8 : y, p \to q' \\ & r_9 : f \Rightarrow p' \} \end{aligned}$$

Moreover, let $c \leq_p a$ *and* $r_5 \leq_d r_4$.

The resulting AT and AF are shown in Figure 1. The grounded extension contains, among others A_1, A_2, *and* A_{13}. *From the defeasible contents of these three, one can construct* A_{14}, *which is not part of the grounded extension.*

Encodings for the Algorithm

The encodings employed by our algorithm detailed in the main paper are detailed in Listings 1 and 2. (Note that the ASP codes are also available at `https://bitbucket.org/coreo-group/aspforaspic`.) The refinement $\pi_r(M)$ is defined as $\{\leftarrow\ out(x_1), ..., out(x_n).\ \mid\ \forall x_i : out(x_i) \in M\}$.

Listing 1 checks if there are assumptions in the given AT that satisfy conditions 1 and 2 from Definition 15, and if so, detects this solution candidate and the defeasible elements are are not individually defeated by the candidate. The queried atom is enforced to be derivable from the candidate (Line 1), and the transitivity of preferences is encoded (Lines 2-5). In Lines 6–7, the symmetry of the contradictory relation and asymmetry of contrary relation is enforced. Lines 8–12 encode a non-deterministic divide of the defeasible parts of the given AT into elements that are in the candidate assumption (P, D) and those that are out. Lines 13–16 compute atoms that are derivable from the candidate assumption. Line 17 enforces that all rules in D are applicable. The individual defeats except for contradictory undermine are computed in Lines 18–20 and for the purposes of checking conflict-freeness, Line 21 moreover computes for which rules the contradictory of the head of said rule is derivable from the candidate. For the purposes of checking contradictory undermining, Lines 22–26 derive all atoms that are derivable from (P', D), where $P' \subseteq P$ and P' is not less preferred than a given premise. Finally Lines 27 and 28 enforce conflict-freeness of the candidate, Lines 29–30 compute which defeasible elements are undefeated by the candidate, and Line 31 enforces that all premises not in the candidate must be defeated by the candidate.

Listing 2 gets as input the candidate under consideration (predicate in) and defeasible elements that are not individually defeated by the candidate (predicate undefeated), and checks if there is a counterexample to the candidate being stable, according to the last item of Definition 15. Lines 1–7 are shared from Listing 1. Lines 8–11 non-deterministically guess a subset of the defeasible parts that are not defeated (predicate suspect, this constituting a possible counterexample, (P', D') from the last item of Definition 15). Lines 12–16 compute atoms derivable from the suspects and enforce that all rules in the suspects must be applicable by (P', D'). Lines 17 and 18 identify the "not less preferred" premise and rule sets P'' and D'' as defined in Definition 14 for contradictory rebuts. It is checked in Lines 19–26 which atoms are derivable from these sets separately, and (successful) contradictory rebuts are checked in Lines 27 and 28. Lastly Line 30 checks if the suspect elements are a subset of the candidate, and Lines 31 and 32 rule out the suspect assumptions as a counterexample if the suspects are a subset of the candidate or if the suspect assumption is contradictory rebutted by the candidate, respectively.

References

Lehtonen, T.; Wallner, J. P.; and Järvisalo, M. 2020. An answer set programming approach to argumentative reasoning in the ASPIC+ framework. In *KR*, 636–646. IJCAI.org.

Modgil, S., and Prakken, H. 2013. A general account of argumentation with preferences. *Artif. Intell.* 195:361–397.

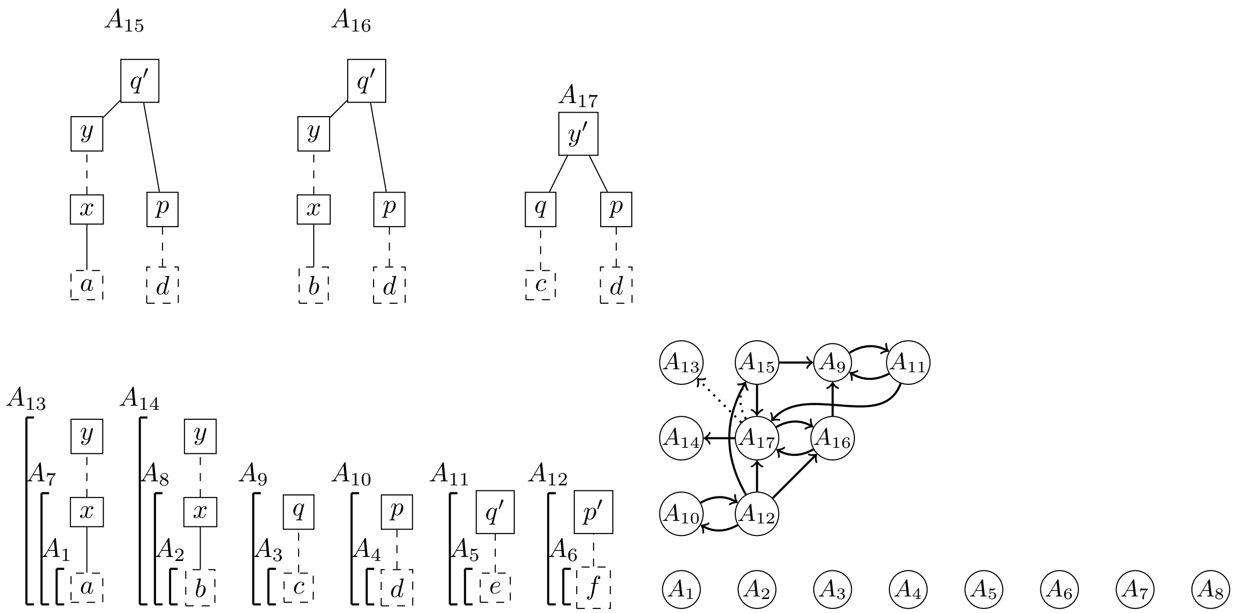

Figure 1: Example AT with strict and defeasible rules not satisfying property stated in Proposition 9, and corresponding AF

Listing 1: Module $\pi_{1,2}(T, query)$

```
preferred(X,Z) :- preferred(X,Y), preferred(Y,Z).
strictly_less_preferred(X,Y) :- preferred(Y,X),
     not preferred(X,Y).
no_less_preferred(X,Y) :- premise(X), premise(Y),
     not strictly_less_preferred(X,Y).
no_less_preferred(X,Y) :- head(X,_), head(Y,_),
     not strictly_less_preferred(X,Y).
contradicts(X,Y) :- contradicts(Y,X).
in(X) :- axiom(X).
in(X) :- premise(X), not out(X).
out(X) :- premise(X), not in(X).
in(R) :- head(R,_), not out(R).
out(R) :- head(R,_), not in(R).
derived(X) :- axiom(X).
derived(X) :- premise(X), in(X).
derived(X) :- head(R,X), used_by_in(R).
derived(X) :- strict_head(R,X), used_by_in(R).
used_by_in(R) :- in(R), head(R,_), derived(X) :
     body(R,X).
used_by_in(R) :- in(R), strict_head(R,_), derived(X)
     : strict_body(R,X).
:- in(R), not used_by_in(R), head(R,_).
defeated(X) :- derived(Y), contrary(X,Y), head(X,_).
defeated(X) :- derived(Y), contrary(X,Y), premise(X).
defeated(X) :- head(X,S), derived(Y), contrary(S,Y).
contradict_rebut_conflict(X) :- head(X,S),
     derived(Y), contradicts(S,Y).
pref_derived(X,Y) :- premise(X), in(X),
     no_less_preferred(X,Y), premise(Y).
pref_derived(X,Y) :- axiom(X), premise(Y).
pref_derived(X,Y) :- head(R,X),
     used_by_pref_premises(R,Y).
pref_derived(X,Y) :- strict_head(R,X),
     used_by_pref_premises(R,Y).
used_by_pref_premises(R,Y) :- head(R,_), premise(Y),
     in(R), pref_derived(X,Y) : body(R,X).
used_by_pref_premises(R,Y) :- strict_head(R,_),
     premise(Y), in(R), pref_derived(X,Y) :
     strict_body(R,X).
defeated(X) :- pref_derived(Y,X), contradicts(X,Y),
     premise(X).
:- in(X), defeated(X).
:- in(X), contradict_rebut_conflict(X).
:- derived(X), supported(Y), contrary(X,Y).
:- derived(X), supported(Y), contradicts(X,Y).
undefeated(X) :- premise(X), not defeated(X).
undefeated(R) :- head(R,_), not defeated(R).
:- premise(X), out(X), undefeated(X).
```

Listing 2: Module $\pi_{\neg 3}(M)$

```
preferred(X,Z) :- preferred(X,Y), preferred(Y,Z).
strictly_less_preferred(X,Y) :- preferred(Y,X),
     not preferred(X,Y).
not_less_preferred(X,Y) :- premise(X),
     not strictly_less_preferred(X,Y), premise(Y).
not_less_preferred(X,Y) :- head(X,_),
     not strictly_less_preferred(X,Y), head(Y,_).
contradicts(X,Y) :- contradicts(Y,X).
in(X) :- axiom(X).
suspect(X) :- axiom(X).
suspect(X) :- undefeated(X), not other(X).
other(X) :- premise(X), not suspect(X).
other(R) :- head(R,_), not suspect(R).
other(R) :- strict_head(R,_), not suspect(R).
derived_by_suspects(X) :- axiom(X).
derived_by_suspects(X) :- premise(X), suspect(X).
derived_by_suspects(X) :- head(R,X),
     used_by_suspects(R).
derived_by_suspects(X) :- strict_head(R,X),
     used_by_suspects(R).
used_by_suspects(R) :- suspect(R),
     derived_by_suspects(X) : body(R,X), head(R,_).
used_by_suspects(R) :- suspect(R),
     derived_by_suspects(X) : strict_body(R,X),
     strict_head(R,_).
:- suspect(R), not used_by_suspects(R), head(R,_).
pref_premise(X) :- premise(X), in(X),
     not_less_preferred(X,Y), premise(Y), suspect(Y).
pref_rule(R) :- head(R,_), in(R), not_less_preferred(
     R,Y), head(Y,_), suspect(Y).
derived_by_pref_prems(X) :- pref_premise(X).
derived_by_pref_prems(X) :- axiom(X).
derived_by_pref_prems(X) :- head(R,X),
     used_by_pref_premises(R).
derived_by_pref_prems(X) :- strict_head(R,X),
     used_by_pref_premises(R).
used_by_pref_premises(R) :- head(R,_),
     derived_by_pref_prems(X) : body(R,X), in(R).
used_by_pref_premises(R) :- strict_head(R,_),
     derived_by_pref_prems(X) : strict_body(R,X),
     in(R).
derived_by_pref_rules(X) :- in(X), premise(X).
derived_by_pref_rules(X) :- axiom(X).
derived_by_pref_rules(X) :- head(R,X),
     used_by_pref_rules(R).
derived_by_pref_rules(X) :- strict_head(R,X),
     used_by_pref_rules(R).
used_by_pref_rules(R) :- pref_rule(R),
     derived_by_pref_rules(X) : body(R,X).
used_by_pref_rules(R) :- strict_rule(R),
     derived_by_pref_rules(X) : strict_body(R,X).
suspect_includes_premises :- suspect(X), premise(X).
rebutted_suspect :- contradicts(X,Y), head(R,X),
     suspect(R), derived_by_pref_prems(Y),
     suspect_includes_premises.
rebutted_suspect :- contradicts(X,Y), head(R,X),
     suspect(R), derived_by_pref_rules(Y).
subset :- in(X) : suspect(X).
:- subset.
:- rebutted_suspect.
```