A Versatile Framework for Formula-Based Enforcement and Synthesis in Abstract Argumentation

Andreas Niskanen[1] , **Jean-Guy Mailly**[2] , **Yannis Dimopoulos**[3] and **Pavlos Moraitis**[4,5]
[1]Department of Computer Science, University of Helsinki, Finland
[2]IRIT, Université Toulouse Capitole, France
[3]Department of Computer Science, University of Cyprus, Cyprus
[4]LIPADE, Université Paris Cité, France
[5]Argument Theory, France

Abstract

Argumentation dynamics provides techniques for revising argumentation theories in real-world domains (e.g., an autonomous medical diagnostic agent). Within this context, enforcement in abstract argumentation has become a prominent topic. Enforcement aims to modify an argumentation framework to satisfy given acceptability conditions while minimizing change from the original framework. Motivated by the need to address both *syntactic* and *semantic* notions of change, we propose *formula-based enforcement*, a generic framework that strictly generalizes existing approaches, by additionally covering cases they cannot handle, including semantic change. We analyze its complexity under central argumentation semantics, obtaining results from NP-completeness to completeness for the third level of the polynomial hierarchy. For second-level complete variants, we present an exact procedure based on MaxSAT solving and counterexample-guided abstraction refinement (CEGAR), and evaluate it empirically.

1 Introduction

Formal argumentation has become a prominent research area in artificial intelligence over the past decades, with applications spanning from reasoning over inconsistent logical knowledge bases [Besnard and Hunter, 2008] to decision-making [Amgoud *et al.*, 2008], negotiation [Dimopoulos *et al.*, 2019], persuasion [Hunter, 2018] and legal reasoning [Atkinson and Bench-Capon, 2023], among others. Among the most central topics, abstract argumentation frameworks (AFs) [Dung, 1995] are directed graphs where nodes are arguments and edges are attacks between those arguments. Arguments acceptability is evaluated through different semantics which specify jointly acceptable sets of arguments called extensions [Baroni *et al.*, 2018]. Argumentation dynamics has emerged as a central research direction in computational argumentation [Doutre and Mailly, 2018] with various forms of enforcement [Baumann and Brewka, 2010; Dyrkolbotn, 2014; Wallner *et al.*, 2017; Haret *et al.*, 2018; Dauphin and Satoh, 2018; Xu *et al.*, 2018; Baumann and Brewka, 2019; Borg and Bex, 2021; Baumann *et al.*, 2021; Heine and Ulbricht, 2024] studied as central computational problems. In enforcement, the goal is to modify an original AF in order to satisfy a given constraint on the acceptance status of specific arguments, e.g., so that a set of arguments is an extension [Baumann and Brewka, 2010; Baumann, 2012; Coste-Marquis *et al.*, 2015; Wallner *et al.*, 2017; Niskanen *et al.*, 2018] or is (credulously or skeptically) accepted [Kontarinis *et al.*, 2013; Doutre *et al.*, 2014; Niskanen *et al.*, 2016; Dauphin and Satoh, 2018]. Among the possible solution AFs, those obtained by minimizing syntactic change to the original AF—i.e., by applying as few modifications as possible to the set of arguments or the attack structure—have been widely investigated from both formal and computational perspectives [Baumann, 2012; Baumann and Brewka, 2013; Doutre *et al.*, 2014; Coste-Marquis *et al.*, 2015; Niskanen *et al.*, 2016; Wallner *et al.*, 2017; Niskanen *et al.*, 2018; Ulbricht and Baumann, 2019; Craandijk and Bex, 2022; Heine and Ulbricht, 2024].

While applying minimal syntactic change is a desirable property from various perspectives such as efficiency and simplicity, in some situations a *semantic notion of change* is more appropriate [Coste-Marquis *et al.*, 2014a; Coste-Marquis *et al.*, 2014b; Diller *et al.*, 2018; Baumann and Brewka, 2019]. For example, consider a medical diagnostic system that has been reliably assisting practitioners for some time but is later found by experts to produce an incorrect inference. The system must be corrected, but ideally in a way that preserves its other, validated conclusions. The repair should therefore be as localized as possible, affecting only the erroneous outcome—i.e., *minimizing semantic change*.

We introduce *formula-based enforcement*, a general framework for enforcement in abstract argumentation allowing various forms of constraints on the resulting AF, including ones for minimizing syntactic or semantic change. Specifically, we use three sets of propositional formulas to express different forms of hard or soft constraints: the first set specifies which semantic constraints must be satisfied by an extension of the AF; the second set forbids extensions of the AF from satisfying specific constraints; and the third set imposes constraints on the attack relation of the result. Each of these constraints can either be hard or soft. This novel framework has both theoretical and practical significance. For example,

consider a diagnostic system that initially deems δ_1 and δ_2 acceptable diagnoses while rejecting δ_3. If medical experts later determine that only δ_1 should be accepted, our approach seamlessly updates the system to preserve δ_1, reject δ_2, and continue rejecting δ_3, while approaches based purely on minimizing syntactic change could lead to an AF with unwanted additional extensions (including some that contain δ_3).

The paper is structured as follows. We begin with a motivating medical diagnosis example, illustrating real-world scenarios *which cannot be fully addressed by existing enforcement methods*. We then provide the necessary formal background on abstract argumentation and argumentation dynamics. Next, we introduce our formula-based enforcement framework, and show that it generalizes various existing approaches to enforcement [Coste-Marquis *et al.*, 2015; Niskanen *et al.*, 2016; Heine and Ulbricht, 2024] and synthesis [Niskanen *et al.*, 2019] in abstract argumentation. We analyze the computational complexity of the corresponding decision problem under five central extension-based semantics, establishing results ranging from NP to Σ_3^P-completeness. For solving Σ_2^{P}-complete variants of formula-based enforcement, building on earlier algorithmic approaches to hard problems in argumentation dynamics [Niskanen *et al.*, 2016; Wallner *et al.*, 2017; Niskanen *et al.*, 2019; Niskanen and Järvisalo, 2020], we develop an exact algorithm based on maximum satisfiability (MaxSAT) solving [Bacchus *et al.*, 2021] and counterexample-guided abstraction refinement (CEGAR) [Clarke *et al.*, 2003; Clarke *et al.*, 2004]. Finally, we empirically evaluate an implementation of the MaxSAT-based CEGAR algorithm on a novel form of enforcement in which both semantic and syntactic change is minimized.[1]

2 Motivation

Consider an argumentation-based diagnostic system built from medical expertise to automate diagnoses such as eye diseases (e.g., [Kakas *et al.*, 2019]). Using expert-defined argumentation theories, the system derives diagnoses from patient symptoms and clinical indicators, supporting tasks such as emergency triage in ophthalmology. However, its diagnoses may sometimes differ from those obtained by physicians after full examination. In such cases, the system must revise its argument-based knowledge by incorporating medical exam results, for instance by adding or removing attacks between arguments so that outcomes align with the ground truth. Any proposed revisions must be validated by experts.

We assume two types of arguments: δ-arguments, supporting diagnoses from observed symptoms, and α-arguments, concerning parameters (e.g., age, medical history) or clinical indicators (e.g., intraocular pressure). This distinction, however, can be omitted. We now examine these cases in detail (considered in Section 4) assuming stable semantics for evaluating argument acceptability. In each scenario, $\mathcal{F}_a$ represents the system-diagnosed diseases and $\mathcal{F}_b$ represents the medical exam results (green arguments in both graphs). We illustrate possible revisions to $\mathcal{F}_a$ to incorporate these results, noting that alternative revisions may also exist.

[1] Sketches of proofs and additional results are included in the supplementary material: https://hal.science/hal-05623895.

In Figure (1a), δ_2 is the disease diagnosed by the system while δ_1 is the disease diagnosed by the medical exam. By updating the attack relation—adding (α_2, δ_2) and removing (α_2, δ_1)—δ_1 becomes the only acceptable δ argument and the system's unique decision, belonging to the extension $\mathcal{E} = \{\delta_1, \alpha_2, \alpha_3\}$. This illustrates a case where one argument is added to the conclusions while another is removed.

In Figure (1c), δ_2 is the disease diagnosed by the system and is confirmed by the medical exam. However, an additional non-contradictory disease, δ_3, previously omitted, is also identified and must belong to the same extension as δ_2. To include δ_3 among the conclusions, the attack (α_2, δ_3) is removed, resulting in the extension $\mathcal{E} = \{\delta_2, \delta_3, \alpha_2, \alpha_3\}$.

In Figure (1e), δ_2 is diagnosed by the system but not confirmed by the medical exam, while a new disease, δ_3—not initially suggested by medical experts—is identified. Consequently, δ_2 must be removed from the conclusions, and δ_3 added. This is accomplished by deleting the attack (α_2, δ_3) and adding (α_2, δ_2), resulting in δ_3 being accepted in the (unique) extension $\mathcal{E} = \{\delta_3, \alpha_2, \alpha_3\}$.

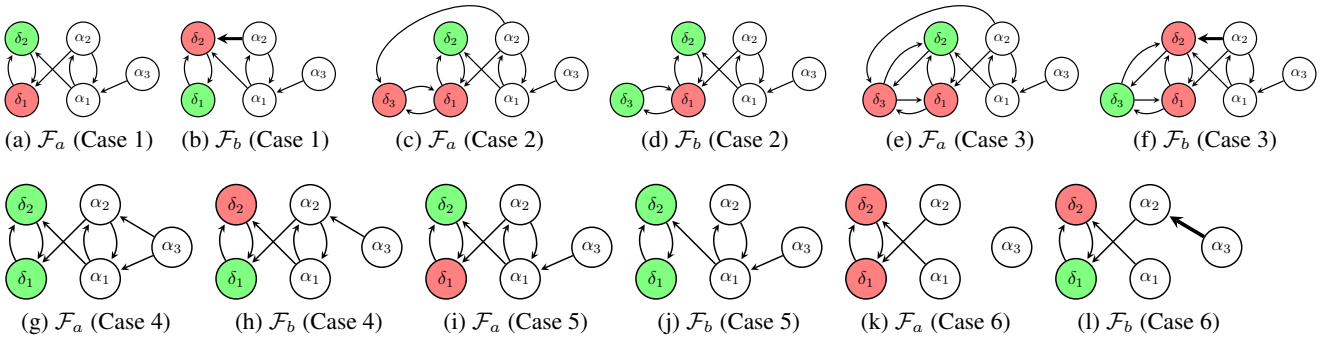
In Figure (1g), the system cannot choose between δ_1 and δ_2, both credulously accepted in $\mathcal{E}_1 = \{\delta_1, \alpha_3\}$ and $\mathcal{E}_2 = \{\delta_2, \alpha_3\}$. The medical exam confirms δ_1 and rules out δ_2, requiring the removal of δ_2. Deleting the attack (α_3, α_1) leads to the unique extension $\mathcal{E} = \{\delta_1, \alpha_1, \alpha_3\}$, reaching the target.

In Figure (1i), the system diagnoses δ_2, while the medical exam identifies both δ_1 and δ_2 as potential diseases. Since δ_1 and δ_2 are in conflict, both should be credulously accepted. Making δ_1 credulously accepted, while keeping it separate from δ_2, is achieved by deleting the attack (α_2, δ_1), resulting in two extensions: $\mathcal{E}_1 = \{\delta_1, \alpha_2, \alpha_3\}$ and $\mathcal{E}_2 = \{\delta_2, \alpha_2, \alpha_3\}$.

Finally, in Figure (1k), the system cannot produce a diagnosis, as both δ_1 and δ_2 are rejected. The medical exam, however, identifies δ_1 as the diagnosed disease, requiring it to be added to the conclusions. Adding the attack (α_3, α_2) makes δ_1 accepted, yielding the unique extension $\mathcal{E} = \{\delta_1, \alpha_1, \alpha_3\}$.

3 Abstract Argumentation and Dynamics

An *abstract argumentation framework* (AF) [Dung, 1995] is a directed graph $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ where $\mathcal{A}$ is a (finite) set of *arguments* and $\mathcal{R} \subseteq \mathcal{A} \times \mathcal{A}$ is an *attack relation*. Arguments acceptability is based on the notion of *extensions*, i.e., sets of collectively acceptable arguments, defined via various *semantics*. Most semantics are based on *conflict-free* sets and the notion of defense. Formally, a set $\mathcal{E} \subseteq \mathcal{A}$ is conflict-free if for all $a, b \in \mathcal{E}$, $(a, b) \notin \mathcal{R}$. A set $\mathcal{E} \subseteq \mathcal{A}$ defends an argument $a \in \mathcal{A}$ if for all $(b, a) \in \mathcal{R}$ there exists $c \in \mathcal{E}$ with $(c, b) \in \mathcal{R}$. A conflict-free set is *admissible* if it defends all its elements. An admissible set is a *complete* extension if it contains all the arguments that it defends. The $\subseteq$-maximal admissible sets are the *preferred* extensions, and the unique $\subseteq$-minimal complete extension is the *grounded* extension. Finally, a conflict-free set is a *stable* extension if it attacks all the arguments outside of it. We denote by $\sigma(\mathcal{F})$ the set of σ-extensions of $\mathcal{F}$, with $\sigma \in \{\mathrm{cf}, \mathrm{ad}, \mathrm{co}, \mathrm{pr}, \mathrm{gr}, \mathrm{stb}\}$ standing (resp.) for conflict-free, admissible, complete, preferred, grounded, and stable semantics. The arguments $\mathrm{cred}_\sigma(\mathcal{F}) = \bigcup \sigma(\mathcal{F})$ (resp. $\mathrm{skep}_\sigma(\mathcal{F}) = \bigcap \sigma(\mathcal{F})$) are *credulously* (resp. *skeptically*) accepted. Fi-



(a) $\mathcal{F}_a$ (Case 1) (b) $\mathcal{F}_b$ (Case 1) (c) $\mathcal{F}_a$ (Case 2) (d) $\mathcal{F}_b$ (Case 2) (e) $\mathcal{F}_a$ (Case 3) (f) $\mathcal{F}_b$ (Case 3)

(g) $\mathcal{F}_a$ (Case 4) (h) $\mathcal{F}_b$ (Case 4) (i) $\mathcal{F}_a$ (Case 5) (j) $\mathcal{F}_b$ (Case 5) (k) $\mathcal{F}_a$ (Case 6) (l) $\mathcal{F}_b$ (Case 6)

Figure 1: All cases of motivation examples

nally, given $\mathcal{E} \in \sigma(\mathcal{F})$, the set of rejected arguments is $\mathcal{E}^O = \{a \in \mathcal{A} \mid \exists b \in \mathcal{E} \text{ s.t. } (b, a) \in \mathcal{R}\}$. See [Dung, 1995; Baroni *et al.*, 2018] for more details on semantics.

Argumentation dynamics has received considerable attention in recent years [Doutre and Mailly, 2018; Baumann *et al.*, 2021]. Notably, in *extension enforcement* [Baumann and Brewka, 2010], given an AF $\mathcal{F}$ and a set of arguments $\mathcal{T}$, the goal is to find an AF $\mathcal{F}'$ such that $\mathcal{T}$ is (included in) an extension of $\mathcal{F}'$. Most of existing work focuses on minimizing syntactic change, i.e., $\mathcal{F}'$ must be as close as possible to $\mathcal{F}$ (w.r.t. the number of attacks added or deleted) [Baumann, 2012; Coste-Marquis *et al.*, 2015; Wallner *et al.*, 2017]. In *status enforcement* [Niskanen *et al.*, 2016], given an AF $\mathcal{F}$ and disjoint sets of arguments $\mathcal{P}$ and $\mathcal{N}$, the goal is to find an AF $\mathcal{F}'$ so that all arguments in $\mathcal{P}$ are credulous (resp. skeptical) conclusions of $\mathcal{F}'$, and no argument in $\mathcal{N}$ is a credulous (resp. skeptical) conclusion of $\mathcal{F}'$. Again, $\mathcal{F}'$ must be as close as possible to $\mathcal{F}$ w.r.t. the attack relation. Other related notions [Niskanen *et al.*, 2019; Heine and Ulbricht, 2024] are formally described in Section 4.1, where we show that they can be captured as specific instances of formula-based enforcement studied in this work.

4 Formula-Based Enforcement

In this section, we formally define formula-based enforcement. Given a finite set of arguments $\mathcal{A}$, we consider three sets of propositional variables, namely $\text{IN}(\mathcal{A}) = \{\text{in}_a \mid a \in \mathcal{A}\}$, $\text{OUT}(\mathcal{A}) = \{\text{out}_a \mid a \in \mathcal{A}\}$ and $\text{ATT}(\mathcal{A}) = \{\text{att}_{a,b} \mid a, b \in \mathcal{A}\}$ which we use to specify semantic and syntactic constraints (either hard or soft) on the resulting solution AF.

Definition 1. A *formula-based enforcement instance* is a tuple $\mathcal{I} = (\mathcal{A}, \Phi^+, \Phi^-, \Phi^{\mathcal{R}}, w, \sigma)$, where Φ^+ and Φ^- are sets of propositional formulas over $\text{IN}(\mathcal{A}) \cup \text{OUT}(\mathcal{A})$, $\Phi^{\mathcal{R}}$ is a set of propositional formulas over $\text{ATT}(\mathcal{A})$, $w\colon \Phi^+ \cup \Phi^- \cup \Phi^{\mathcal{R}} \to \mathbb{N} \cup \{\infty\}$ is a weight function, and σ is a semantics.

Intuitively, formulas in Φ^+ express semantic constraints on what should hold in an extension, while formulas in Φ^- express constraints which should not hold in any extension of the solution AF. In turn, $\Phi^{\mathcal{R}}$ constrain the attack relation of the solution AF. Finally, w determines whether a constraint is hard ($w(\varphi) = \infty$) or soft ($w(\varphi) < \infty$). For instance, to require an argument a to be credulously but not skeptically accepted, while being attacked by b which is not credulously accepted, we need the formulas in_a ($a \in \mathcal{E}$ for some extension $\mathcal{E}$) and $\neg\text{in}_a$ ($a \notin \mathcal{E}'$ for some extension $\mathcal{E}'$) in Φ^+, in_b in Φ^-, and $\text{att}_{b,a}$ in $\Phi^{\mathcal{R}}$. We continue by formally defining how an AF satisfies each of these formulas: we associate the attacks of the AF, as well as each of its extension, with a truth assignment, and then make use of propositional satisfiability.

Definition 2. For an AF $\mathcal{F} = (\mathcal{A}, \mathcal{R})$, we define $\tau^{\mathcal{F}}\colon \text{ATT}(\mathcal{A}) \to \{0, 1\}$ as the truth assignment mapping $\tau^{\mathcal{F}}(\text{att}_{a,b}) = 1$ iff $(a, b) \in \mathcal{R}$. For an extension $\mathcal{E} \in \sigma(\mathcal{F})$, we define $\tau^{\mathcal{E}}\colon \text{IN}(\mathcal{A}) \cup \text{OUT}(\mathcal{A}) \to \{0, 1\}$ mapping $\tau^{\mathcal{E}}(\text{in}_a) = 1$ iff $a \in \mathcal{E}$, and $\tau^{\mathcal{E}}(\text{out}_a) = 1$ iff $a \in \mathcal{E}^O$. Given an instance $\mathcal{I} = (\mathcal{A}, \Phi^+, \Phi^-, \Phi^{\mathcal{R}}, w, \sigma)$, an AF $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ satisfies

- $\varphi \in \Phi^+$ if there exists $\mathcal{E} \in \sigma(\mathcal{F})$ s.t. $\tau^{\mathcal{E}}$ satisfies φ;
- $\varphi \in \Phi^-$ if there is no $\mathcal{E} \in \sigma(\mathcal{F})$ s.t. $\tau^{\mathcal{E}}$ satisfies φ;
- $\varphi \in \Phi^{\mathcal{R}}$ if $\tau^{\mathcal{F}}$ satisfies φ.

We write $\mathcal{F} \models \varphi$ if $\mathcal{F}$ satisfies a formula $\varphi \in \Phi^+ \cup \Phi^- \cup \Phi^{\mathcal{R}}$.

Let $\mathcal{I}$ be a formula-based enforcement instance. An AF $\mathcal{F}$ is a *solution* of $\mathcal{I}$ if it satisfies each formula $\varphi \in \Phi^+ \cup \Phi^- \cup \Phi^{\mathcal{R}}$ with $w(\varphi) = \infty$, i.e., all of the hard constraints. Weights of soft constraints are used to compare the quality of different solutions of $\mathcal{I}$. The *cost* of an AF $\mathcal{F}$ is the sum of the weights of formulas in $\Phi^+ \cup \Phi^- \cup \Phi^{\mathcal{R}}$ not satisfied by $\mathcal{F}$. Formally,

$$\text{COST}(\mathcal{I}, \mathcal{F}) = \sum_{\varphi \in \Phi^+ \cup \Phi^- \cup \Phi^{\mathcal{R}} \mid \mathcal{F} \not\models \varphi} w(\varphi)$$

An AF $\mathcal{F}$ is an *optimal solution* of $\mathcal{I}$ if it achieves *minimum cost* over all AFs with the same set of arguments, i.e., there is no AF $\mathcal{F}' = (\mathcal{A}, \mathcal{R}')$ with $\text{COST}(\mathcal{I}, \mathcal{F}') < \text{COST}(\mathcal{I}, \mathcal{F})$.

The proposed approach offers a generic and versatile framework that captures various enforcement and synthesis problems in the literature. For example, hard constraints in Φ^+ and Φ^- can be used to describe knowledge on the acceptance status of arguments, which must be incorporated into the current AF by revising it in a syntactically minimal sense, which in turn is achieved by including soft constraints in $\Phi^{\mathcal{R}}$ describing the attack structure of the AF. On the other hand, in a pure synthesis setting the set $\Phi^{\mathcal{R}}$ is empty, whereas the

sets Φ^+ and Φ^- can be used to describe positive and negative examples on the extensions of the inferred AF. This formula-based approach is proposed because it allows us to furnish the technical mechanisms to achieve the simultaneous capture and minimization of syntactic and semantic changes within a single unified framework.

4.1 Relation with other forms of enforcement

We continue by formally proving connections to well-known forms of enforcement and synthesis approaches, specifically by showing how they are captured as instances of formula-based enforcement. We begin by introducing several notations for useful formulas. Given $\mathcal{F} = (\mathcal{A}, \mathcal{R})$, we define the set of formulas (unit clauses) $\Phi^{\mathcal{F}} = \{\text{att}_{a,b} \mid (a,b) \in \mathcal{R}\} \cup \{\neg\text{att}_{a,b} \mid (a,b) \notin \mathcal{R}\}$ capturing the syntax of $\mathcal{F}$. For a set of arguments $\mathcal{T} \subseteq \mathcal{A}$, we define several formulas. To represent that $\mathcal{T}$ must be *included in* an extension, we define $\varphi^{\mathcal{T}}_{\subseteq} = \bigwedge_{a \in \mathcal{T}} \text{in}_a$; for representing that $\mathcal{T}$ *is* an extension, we define $\varphi^{\mathcal{T}}_{=} = \varphi^{\mathcal{T}}_{\subseteq} \wedge \bigwedge_{a \in \mathcal{A} \setminus \mathcal{T}} \neg\text{in}_a$. Finally, to constrain $\mathcal{T}$ to be the *rejected* set w.r.t. an extension (i.e. $\mathcal{T} = \mathcal{E}^O$ with $\mathcal{E}$ an extension), we define $\varphi^{\mathcal{T}}_{\text{out}} = \bigwedge_{a \in \mathcal{T}} \text{out}_a \wedge \bigwedge_{a \in \mathcal{A} \setminus \mathcal{T}} \neg\text{out}_a$. Finally, to characterize syntactic change, we use the symmetric difference between A and B, $A \triangle B = (A \setminus B) \cup (B \setminus A)$.

We start with (argument-fixed) extension enforcement [Coste-Marquis *et al.*, 2015; Wallner *et al.*, 2017].

Definition 3. An *extension enforcement instance* is a tuple $(\mathcal{F}, \mathcal{T}, \sigma)$ where $\mathcal{F} = (\mathcal{A}, \mathcal{R})$, $\mathcal{T} \subseteq \mathcal{A}$ is a set of arguments, and σ an argumentation semantics. A solution of the enforcement problem is an AF $\mathcal{F}' = (\mathcal{A}, \mathcal{R}')$ such that there exists $\mathcal{E} \in \sigma(\mathcal{F}')$ with $\mathcal{T} \subseteq \mathcal{E}$ (*non-strict* enforcement). A solution is *optimal* if it minimizes $|\mathcal{R} \triangle \mathcal{R}'|$.

Proposition 1. *Let $(\mathcal{F}, \mathcal{T}, \sigma)$ be an instance of non-strict extension enforcement, and define $\mathcal{I}^{\text{NSTR}} = (\mathcal{A}, \{\varphi^{\mathcal{T}}_{\subseteq}\}, \emptyset, \Phi^{\mathcal{F}}, w, \sigma)$ as the formula enforcement instance where $w(\varphi^{\mathcal{T}}_{\subseteq}) = \infty$, and $w(\varphi) = 1$ for all $\varphi \in \Phi^{\mathcal{F}}$. It holds that $\mathcal{F}^*$ is an optimal solution to $(\mathcal{F}, \mathcal{T}, \sigma)$ iff $\mathcal{F}^*$ is an optimal solution to $\mathcal{I}^{\text{NSTR}}$.*

Other forms of extension enforcement can also be captured, including enforcement under different forms of expansions [Baumann and Brewka, 2010; Baumann, 2012]; roughly speaking, an expansion is a modification of an AF where only addition of arguments and attacks is allowed, but no argument or attack is removed. This can be captured as formula-based enforcement via hard constraints in $\Phi^{\mathcal{R}}$ that guarantee that $\mathcal{F}^*$ is a suitable expansion of $\mathcal{F}$, and using a bounded number of auxiliary arguments (i.e., constructing the result on the set of arguments $\mathcal{A}'$ with $\mathcal{A} \subset \mathcal{A}'$).

We move on to status enforcement [Niskanen *et al.*, 2016].

Definition 4. A *credulous status enforcement instance* is a tuple $(\mathcal{F}, \mathcal{P}, \mathcal{N}, \sigma)$ where $\mathcal{F} = (\mathcal{A}, \mathcal{R})$, $\mathcal{P}, \mathcal{N} \subseteq \mathcal{A}$, and σ is a semantics. A solution of the enforcement problem is $\mathcal{F}' = (\mathcal{A}, \mathcal{R}')$ such that $\mathcal{P} \subseteq \text{cred}_\sigma(\mathcal{F}')$ and $\mathcal{N} \cap \text{cred}_\sigma(\mathcal{F}') = \emptyset$. A solution is *optimal* if it minimizes $|\mathcal{R} \triangle \mathcal{R}'|$.

Proposition 2. *Let $(\mathcal{F}, \mathcal{P}, \mathcal{N}, \sigma)$ be an instance of credulous status enforcement, and define $\mathcal{I}^{\text{CRED}} = (\mathcal{A}, \Phi^{\mathcal{P}}, \Phi^{\mathcal{N}}, \Phi^{\mathcal{F}}, w, \sigma)$ as the formula enforcement instance where $\Phi^{\mathcal{P}} = \{in_a \mid a \in \mathcal{P}\}$, $\Phi^{\mathcal{N}} = \{\neg in_a \mid a \in \mathcal{N}\}$, $w(\varphi) = \infty$ for all $\varphi \in \Phi^{\mathcal{P}} \cup \Phi^{\mathcal{N}}$, and $w(\varphi) = 1$ for all $\varphi \in \Phi^{\mathcal{F}}$. It holds that $\mathcal{F}^*$ is an optimal solution to $(\mathcal{F}, \mathcal{P}, \mathcal{N}, \sigma)$ iff $\mathcal{F}^*$ is an optimal solution to $\mathcal{I}^{\text{CRED}}$.*

We continue with *rejection* enforcement [Heine and Ulbricht, 2024] where the goal is to ensure that a set of arguments is rejected, i.e., is exactly the set of arguments $\mathcal{E}^O$ attacked by an extension $\mathcal{E}$.

Definition 5. A *rejection enforcement instance* is a tuple $(\mathcal{F}, \mathcal{T}, \sigma)$ where $\mathcal{F} = (\mathcal{A}, \mathcal{R})$, $\mathcal{T} \subseteq \mathcal{A}$, and σ is a semantics. A solution of the rejection enforcement instance is an AF $\mathcal{F}' = (\mathcal{A}, \mathcal{R}')$ such that $\mathcal{T} = \mathcal{E}^O$ for some $\mathcal{E} \in \sigma(\mathcal{F}')$. Moreover, $\mathcal{F}'$ is *consensus preserving* if $\sigma(\mathcal{F}) \subseteq \sigma(\mathcal{F}')$. Finally, $\mathcal{F}'$ is *optimal* if it minimizes $|\mathcal{R} \triangle \mathcal{R}'|$.

The original definition of rejection enforcement uses expansions. As discussed previously, expansions are captured by adding suitable hard constraints to $\Phi^{\mathcal{R}}$ and auxiliary arguments. For readability, we disregard these constraints here.

Proposition 3. *Let $(\mathcal{F}, \mathcal{T}, \sigma)$ be a rejection enforcement instance, and define $\mathcal{I}^{\text{REJ}} = (\mathcal{A}, \{\varphi^S_{=} \mid S \in \sigma(\mathcal{F})\}, \{\varphi^{\mathcal{T}}_{out}\}, \Phi^{\mathcal{F}}, w, \sigma)$ where $w(\varphi) = \infty$ for all $\varphi \in \Phi^+ \cup \Phi^-$, and $w(\varphi) = 1$ for all $\varphi \in \Phi^{\mathcal{F}}$. It holds that $\mathcal{F}^*$ is a consensus preserving optimal solution to $(\mathcal{F}, \mathcal{T}, \sigma)$ iff it is an optimal solution to $\mathcal{I}^{\text{REJ}}$.*

Finally, we show a connection to AF *synthesis* [Niskanen *et al.*, 2019]: given positive examples of extensions and negative examples of non-extensions, we need to find $\mathcal{F}$ such that positive examples are extensions of $\mathcal{F}$ and no negative example is an extension. Examples which are not satisfied incur a cost, and the goal is to find a solution of minimum cost.

Definition 6. An *AF synthesis instance* is a tuple $(\mathcal{A}, E^+, E^-, \sigma)$ where $\mathcal{A}$ is a finite set of arguments, E^+ and E^- are sets of weighted extensions (S, w) with $S \subseteq \mathcal{A}$ and $w > 0$ an integer, and σ is a semantics. A solution of the synthesis problem is an AF $\mathcal{F} = (\mathcal{A}, \mathcal{R})$. The *cost* of an $\mathcal{F}$ is $\sum_{(S,w) \in E^+ | S \notin \sigma(\mathcal{F})} w + \sum_{(S,w) \in E^- | S \in \sigma(\mathcal{F})} w$. An AF is *optimal* if it achieves minimum cost.

Proposition 4. *Let $(\mathcal{A}, E^+, E^-, \sigma)$ be an instance of AF synthesis, and define $\mathcal{I}^{\text{SYNTH}} = (\mathcal{A}, \Phi^{E^+}, \Phi^{E^-}, \emptyset, w, \sigma)$ as the formula enforcement instance where $\Phi^{E^+} = \{\varphi^S_{=} \mid (S, w_S) \in E^+\}$, $\Phi^{E^-} = \{\varphi^S_{=} \mid (S, w_S) \in E^-\}$, and for each $(S, w_S) \in E^+ \cup E^-$, $w(\varphi^S_{=}) = w_S$. It holds that $\mathcal{F}^*$ is an optimal solution to $(\mathcal{A}, E^+, E^-, \sigma)$ iff $\mathcal{F}^*$ is an optimal solution to $\mathcal{I}^{\text{SYNTH}}$.*

4.2 Novel forms of enforcement and synthesis

Our framework can capture novel notions of synthesis and enforcement. Starting with synthesis, we can express complex constraints that need to be satisfied by the result, e.g. $((\neg\text{in}_a \wedge \neg\text{in}_b) \rightarrow \text{in}_c) \in \Phi^+$ requires that if there is an extension which does not contain a nor b, then c belongs to the extension. Moreover, formulas can capture credulous (resp. skeptical) constraints, i.e., those which must be satisfied by some (resp. each) extension. For instance, $(\text{in}_a \vee \text{in}_b) \in \Phi^+$ results in an AF that has at least one extension that contains either a

or b, whereas the formula $(\neg \mathrm{in}_a \wedge \neg \mathrm{in}_b) \in \Phi^-$, induces an AF where either a or b belongs to every extension.

In the context of enforcement, an important feature of the formula-based approach is its capability to model both *semantic* (i.e., w.r.t. the acceptance of arguments) and *syntactic* (i.e., w.r.t. the structure of the AF) change. These notions can also be combined in various ways. For example, one can give more priority to syntactic change by specifying larger weights to the corresponding constraints. When it is equally important to keep the structure of the AF as well as the acceptance status of its arguments, we can use the same weight for each formula. The formula-based enforcement framework also allows for minimizing semantic change and syntactic change in a lexicographic way by selecting suitable weights. For instance, assigning to "syntactic" formulas $\varphi \in \Phi^{\mathcal{R}}$ the weight $w(\varphi) = 1$, and to "semantic" formulas $\varphi \in \Phi^+ \cup \Phi^-$ the weight $w(\varphi) = |\Phi^{\mathcal{R}}| + 1$, induces higher priority to semantic change, and lower priority to syntactic change.

We also distinguish between hard and soft constraints, with hard constraints specified using infinite weights. For instance, assume a scenario where both α and β must absolutely be credulously accepted, and *if possible* they should be accepted together, i.e., in the same extension (but if it is not possible, then they might be accepted in different extensions). Finally, syntactic change $|\mathcal{R} \triangle \mathcal{R}'|$ should be minimized. In this case we define the formula enforcement instance with $\Phi^+ = \{\mathrm{in}_\alpha, \mathrm{in}_\beta, \mathrm{in}_\alpha \wedge \mathrm{in}_\beta\}$, $\Phi^- = \emptyset$, $\Phi^{\mathcal{R}} = \{\mathrm{att}_{a,b} \mid (a,b) \in \mathcal{R}\} \cup \{\neg \mathrm{att}_{a,b} \mid (a,b) \notin \mathcal{R}\}$. The unit clauses in Φ^+ are hard constraints with $w(\mathrm{in}_\alpha) = w(\mathrm{in}_\beta) = \infty$, while the third formula has weight $w(\mathrm{in}_\alpha \wedge \mathrm{in}_\beta) = |\mathcal{A}|^2 + 1$ higher than the number of formulas in $\Phi^{\mathcal{R}}$, which in turn have unit weight.

The following concept illustrates how lexicographic optimization can be applied to capture more advanced forms of enforcement. An instance of *credulous status enforcement with soft joint acceptance constraints* $(\mathcal{F}, \mathcal{P}, \mathcal{N}, \mathcal{P}^\wedge, \mathcal{N}^\wedge, \sigma)$, is an extension of credulous status enforcement (Def. 4), where $\mathcal{P}^\wedge, \mathcal{N}^\wedge$ are collections of positive and negative target sets. The goal is to enforce as many positive target sets collectively (in the sense of non-strict extension enforcement), and as few negative target sets as possible (reminiscent of positive and negative examples in AF synthesis). We use the following notation: given a set U, a set $A \subseteq U$, and a set of sets $\mathcal{B} \subseteq 2^U$, $L(A, \mathcal{B})$ is true iff there is a set $S \in \mathcal{B}$ s.t. $A \subseteq S$.

Definition 7. An instance of *credulous status enforcement with soft joint acceptance constraints* is a tuple $(\mathcal{F}, \mathcal{P}, \mathcal{N}, \mathcal{P}^\wedge, \mathcal{N}^\wedge, \sigma)$, where $\mathcal{F} = (\mathcal{A}, \mathcal{R})$, $\mathcal{P}, \mathcal{N} \subseteq \mathcal{A}$, $\mathcal{P}^\wedge, \mathcal{N}^\wedge \subseteq 2^{\mathcal{A}}$, and σ is an argumentation semantics. A solution of the enforcement problem is an AF $\mathcal{F}' = (\mathcal{A}, \mathcal{R}')$ such that each $p \in \mathcal{P}$ is credulously accepted w.r.t. σ, no $n \in \mathcal{N}$ is credulously accepted w.r.t. σ, and $\mathcal{F}'$ minimizes $\sum_{S^+ \in \mathcal{P}^\wedge} [\neg L(S^+, \sigma(\mathcal{F}'))] + \sum_{S^- \in \mathcal{N}^\wedge} [L(S^-, \sigma(\mathcal{F}'))]$. A solution is *syntactically optimal* if it minimizes $|\mathcal{R} \triangle \mathcal{R}'|$.

We extend the notation $\varphi^{\mathcal{T}}_{\subseteq}$ from a set of arguments $\mathcal{T}$ to a collection of sets $\mathcal{S}$ by defining $\varphi^{\mathcal{S}}_{\subseteq} = \{\varphi^{\mathcal{T}}_{\subseteq} \mid \mathcal{T} \in \mathcal{S}\}$.

Proposition 5. *Let $(\mathcal{F}, \mathcal{P}, \mathcal{N}, \mathcal{P}^\wedge, \mathcal{N}^\wedge, \sigma)$ be an instance of credulous status enforcement with soft joint acceptance constraints, and define $\mathcal{I}^{\text{CRED}}_{\text{SEM,SYN}} = (\mathcal{A}, \Phi^{\mathcal{P}} \cup \Phi^{\mathcal{P}^\wedge}, \Phi^{\mathcal{N}} \cup \Phi^{\mathcal{N}^\wedge}, \Phi^{\mathcal{F}}, w, \sigma)$ as the formula-based enforcement instance with $\Phi^{\mathcal{P}} = \{in_a \mid a \in \mathcal{P}\}$, $\Phi^{\mathcal{P}^\wedge} = \varphi^{\mathcal{P}^\wedge}_{\subseteq}$, $\Phi^{\mathcal{N}} = \{in_a \mid a \in \mathcal{N}\}$, $\Phi^{\mathcal{N}^\wedge} = \varphi^{\mathcal{N}^\wedge}_{\subseteq}$, $w(\varphi) = \infty$ for all $\varphi \in \Phi^{\mathcal{P}} \cup \Phi^{\mathcal{N}}$, $w(\varphi) = |\mathcal{A}|^2 + 1$ for all $\Phi^{\mathcal{P}^\wedge} \cup \Phi^{\mathcal{N}^\wedge}$, and $w(\varphi) = 1$ for all $\Phi^{\mathcal{F}}$. It holds that $\mathcal{F}^*$ is a syntactically optimal solution to $(\mathcal{F}, \mathcal{P}, \mathcal{N}, \mathcal{P}^\wedge, \mathcal{N}^\wedge, \sigma)$ iff $\mathcal{F}^*$ is an optimal solution to $\mathcal{I}^{\text{CRED}}_{\text{SEM,SYN}}$.*

In addition to using $\mathcal{P}^\wedge$ to model that as many extensions as possible need to be preserved, note that semantic distance in terms of credulously accepted arguments can be minimized by including $\bigcup \sigma(F)$ into $\mathcal{P}$ and $\mathcal{A} \setminus \bigcup \sigma(F)$ into $\mathcal{N}$.

Finally, we demonstrate our framework's instantiation through the real-world cases of the medical diagnosis problem introduced in Section 2. Through these instantiations, the target argumentation framework $\mathcal{F}_b$ is obtained from $\mathcal{F}_a$. For **Case 1**: $\Phi^+ = \{\mathrm{in}_{\delta_1}, \mathrm{in}_{\alpha_2}, \mathrm{in}_{\alpha_3}\}$, $\Phi^- = \{\mathrm{in}_{\delta_2}\}$, $\Phi^{\mathcal{R}} = \{\mathrm{att}_{\alpha_2,\delta_1}, \mathrm{att}_{\alpha_3,\alpha_1}, \mathrm{att}_{\alpha_1,\alpha_2}, \mathrm{att}_{\alpha_2,\alpha_1}, \mathrm{att}_{\delta_1,\delta_2}, \mathrm{att}_{\delta_2,\delta_1}, \mathrm{att}_{\alpha_1,\delta_2}\} \cup \{\neg \mathrm{att}_{\delta_1,\alpha_2}, \neg \mathrm{att}_{\delta_1,\alpha_3}, \neg \mathrm{att}_{\alpha_3,\alpha_2}, \neg \mathrm{att}_{\alpha_2,\alpha_3}, \neg \mathrm{att}_{\alpha_3,\delta_2}, \neg \mathrm{att}_{\delta_2,\alpha_3}, \neg \mathrm{att}_{\alpha_1,\delta_1}, \neg \mathrm{att}_{\delta_1,\alpha_1}, \neg \mathrm{att}_{\alpha_3,\delta_1}\}$. We do not repeat $\Phi^{\mathcal{R}}$ for the other cases since it is similar. For **Case 2**: $\Phi^+ = \{\mathrm{in}_{\delta_2}, \mathrm{in}_{\delta_3}, \mathrm{in}_{\alpha_2}, \mathrm{in}_{\alpha_3}, \mathrm{in}_{\delta_2} \wedge \mathrm{in}_{\delta_3}\}$ and $\Phi^- = \{\mathrm{in}_{\delta_1}\}$. In this case $\mathrm{in}_{\delta_2} \wedge \mathrm{in}_{\delta_3} \in \Phi^+$ encodes the requirement that δ_2 and δ_3 must belong to the same extension. For **Case 3**: $\Phi^+ = \{\mathrm{in}_{\delta_3}, \mathrm{in}_{\alpha_2}, \mathrm{in}_{\alpha_3}\}$ and $\Phi^- = \{\mathrm{in}_{\delta_1}, \mathrm{in}_{\delta_2}\}$. For **Case 4**: $\Phi^+ = \{\mathrm{in}_{\delta_1}, \mathrm{in}_{\alpha_1}, \mathrm{in}_{\alpha_3}\}$ and $\Phi^- = \{\mathrm{in}_{\delta_2}\}$. For **Case 5**: $\Phi^+ = \{\neg \mathrm{in}_{\delta_1}, \mathrm{in}_{\delta_1}, \neg \mathrm{in}_{\delta_2}, \mathrm{in}_{\delta_2}, \mathrm{in}_{\alpha_2}, \mathrm{in}_{\alpha_3}\}$ and $\Phi^- = \{\mathrm{in}_{\delta_1} \wedge \mathrm{in}_{\delta_2}\}$. In this case, $\neg \mathrm{in}_{\delta_1}$, in_{δ_1}, $\neg \mathrm{in}_{\delta_2}$ and in_{δ_2} encode the requirement that δ_1 and δ_2 must be credulously accepted, while $\mathrm{in}_{\delta_1} \wedge \mathrm{in}_{\delta_2} \in \Phi^-$ expresses that they cannot belong to the same extension. Finally, for **Case 6**: $\Phi^+ = \{\mathrm{in}_{\delta_1}, \mathrm{in}_{\alpha_1}, \mathrm{in}_{\alpha_3}\}$ and $\Phi^- = \{\mathrm{in}_{\delta_2}\}$.

In all these cases, formulas $\varphi \in \Phi^+ \cup \Phi^-$ are hard constraints ($w(\varphi) = \infty$). In contrast, formulas $\varphi' \in \Phi^{\mathcal{R}}$ are soft constraints with $w(\varphi') = 1$, representing the initial AF. Consequently, our method computes a new AF (i.e., $\mathcal{F}_b$) that is as close as possible (in syntactic terms) to the original (i.e., $\mathcal{F}_a$) while ensuring all semantic constraints are satisfied. We emphasize that our approach is capable of capturing all these cases, as well as their combinations, whereas existing approaches in the literature are unable to do so.

5 Complexity Results

We study the complexity of the following problem: given $\mathcal{I} = (\mathcal{A}, \Phi^+, \Phi^-, \Phi^{\mathcal{R}}, w, \sigma)$, and a non-negative integer $k \in \mathbb{N}$, is there an AF $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ with $\text{COST}(\mathcal{I}, \mathcal{F}) \leq k$?

As usual, grounded semantics exhibits lower complexity than admissible, complete and stable semantics, which are in turn more tractable than preferred semantics.

Proposition 6. *Formula-based enforcement is* NP*-complete for $\sigma = \mathrm{gr}$; Σ^{P}_2-complete for $\sigma \in \{\mathrm{ad}, \mathrm{co}, \mathrm{stb}\}$; and Σ^{P}_3-complete for $\sigma = \mathrm{pr}$.*

Restricted variants of formula-based enforcement may exhibit milder complexity than the general case.

Proposition 7. *Formula-based enforcement under $\sigma \in \{\mathrm{ad}, \mathrm{co}, \mathrm{stb}\}$ is* NP*-complete when $\Phi^- = \emptyset$.*

Algorithm 1 MaxSAT-based CEGAR for formula enforcement. **Input**: Arguments $\mathcal{A}$, sets of formulas $\Phi^+, \Phi^-, \Phi^{\mathcal{R}}$, weight function w, semantics $\sigma \in \{\mathrm{ad}, \mathrm{stb}\}$.

1: $(\varphi_{\text{hard}}, \varphi_{\text{soft}}) \leftarrow \textsc{Abstraction}_\sigma(\mathcal{A}, \Phi^+, \Phi^-, \Phi^{\mathcal{R}}, w)$
2: **while true do**
3: $(result, \tau) \leftarrow \textsc{MaxSAT}(\varphi_{\text{hard}}, \varphi_{\text{soft}})$
4: **if** $result = unsat$ **then return false**
5: $\mathcal{R} \leftarrow \textsc{AttackStructure}(\tau)$; $\mathcal{F} \leftarrow (\mathcal{A}, \mathcal{R})$
6: $\varphi_{\text{cex}} \leftarrow \textsc{CounterExample}_\sigma(\Phi^-, w, \mathcal{F}, \tau)$
7: $(result, \tau_{\text{cex}}) \leftarrow \textsc{SAT}(\varphi_{\text{cex}})$
8: **if** $result = unsat$ **then return** $\mathcal{F}$
9: $\mathcal{E} \leftarrow \textsc{Extension}(\tau_{\text{cex}})$
10: $\varphi_{\text{hard}} \leftarrow \varphi_{\text{hard}} \wedge \textsc{Refine}(\Phi^-, w, \mathcal{F}, \mathcal{E}, \tau_{\text{cex}})$

6 Algorithms

Now we present an algorithm for solving Σ_2^{P}-complete formula enforcement under admissible and stable semantics.[2] Our approach is based on using a MaxSAT solver [Bacchus *et al.*, 2021] in an iterative counterexample-guided abstraction refinement (CEGAR) [Clarke *et al.*, 2003; Clarke *et al.*, 2004] procedure. The core idea is to iteratively solve an *abstraction* in NP, i.e., an over-approximation of the set of (optimal) solutions, obtaining solution candidates. At each iteration we check whether the current candidate is a valid solution by searching for a *counterexample*. If a counterexample is found, the abstraction is *refined* by adding a constraint which prevents the counterexample from occurring in future iterations. This continues until no counterexamples are found—i.e., the candidate solution is valid—or there are no more solutions to the abstraction, i.e., the problem has no solution either.

For background on SAT and MaxSAT, recall that for a Boolean variable x, there are two literals, x and $\neg x$. A clause C is a disjunction ($\vee$) of literals. A conjunctive normal form (CNF) formula φ is a conjunction ($\wedge$) of clauses. A truth assignment τ maps each variable in φ to 0 (false) or 1 (true), and extends to literals via $\tau(\neg x) = 1 - \tau(x)$, to clauses via $\tau(C) = \max\{\tau(l) \mid l \in C\}$, and to CNF formulas via $\tau(\varphi) = \min\{\tau(C) \mid C \in \varphi\}$. Given a CNF formula φ, the *Boolean satisfiability* problem (SAT) asks whether there is an assignment τ with $\tau(\varphi) = 1$. If there is, φ is satisfiable, and otherwise φ is unsatisfiable. In the MaxSAT problem [Bacchus *et al.*, 2021], we are given a set of hard clauses φ_{hard} and a set of weighted soft clauses φ_{soft} consisting of pairs (C, w_C). The task is to find a truth assignment τ with $\tau(\varphi_{\text{hard}}) = 1$ which minimizes the cost $\sum_{(C,w_C)\in\varphi_{\text{soft}}} w_C(1 - \tau(C))$.

We begin by detailing our MaxSAT encoding of the NP abstraction, which, in line with Proposition 7, relaxes the constraints expressed by formulas in Φ^- . The $\text{att}_{a,b}$ variables for each $a, b \in \mathcal{A}$ are directly used in the encoding to represent the solution AF. For each $\varphi \in \Phi^+$, we declare fresh variables in_a^φ and out_a^φ for each $a \in \mathcal{A}$, representing an extension witnessing the satisfaction of φ. Then, for each $\varphi \in \Phi^+$ and $\psi \in \Phi^+ \cup \Phi^-$, we let $\psi^\varphi = \psi[\text{in}_a \mapsto \text{in}_a^\varphi \mid a \in \mathcal{A}][\text{out}_a \mapsto \text{out}_a^\varphi \mid a \in \mathcal{A}]$, that is, ψ^φ is the formula ψ with variables in_a and out_a replaced with variables corresponding to formula φ. Finally, for each $\varphi \in \Phi^-$ with $w(\varphi) < \infty$, we declare a fresh variable sat_φ to indicate the satisfaction of φ.

Our algorithm is presented as Algorithm 1. We start by initializing an abstraction (line 1) as a MaxSAT instance $(\varphi_{\text{hard}}, \varphi_{\text{soft}})$. The hard clauses are defined as

$$\varphi_{\text{hard}} = \bigwedge_{\varphi\in\Phi^+} \textsc{Ext}_\sigma^\varphi(\mathcal{A}, \Phi^-, w) \wedge \bigwedge_{\substack{\varphi\in\Phi^+ \\ w(\varphi)=\infty}} \varphi^\varphi \wedge \bigwedge_{\substack{\varphi\in\Phi^{\mathcal{R}} \\ w(\varphi)=\infty}} \varphi$$

where the first conjunction encodes for each formula $\varphi \in \Phi^+$ a σ-extension witnessing its satisfaction, and the latter two enforce that each formula $\varphi \in \Phi^+ \cup \Phi^{\mathcal{R}}$ with $w(\varphi) = \infty$ must be satisfied. To encode a σ-extension, we define $\textsc{Ext}_\sigma^\varphi(\mathcal{A}, \Phi^-, w)$ as $\bigwedge_{a\in\mathcal{A}}(\text{out}_a^\varphi \leftrightarrow \bigvee_{b\in\mathcal{A}}(\text{att}_{b,a} \wedge \text{in}_b^\varphi)) \wedge \textsc{Sem}_\sigma^\varphi(\mathcal{A}) \wedge \textsc{Aux}^\varphi(\Phi^-, w)$, where the first conjunction encodes the interpretation of out-variables, $\textsc{Sem}_\sigma^\varphi(\mathcal{A})$ encodes the semantics—in line with MaxSAT encodings for extension and status enforcement [Wallner *et al.*, 2017; Niskanen *et al.*, 2016]—via $\textsc{Sem}_{\text{cf}}^\varphi(\mathcal{A}) = \bigwedge_{a,b\in\mathcal{A}}(\text{att}_{a,b} \rightarrow (\neg\text{in}_a^\varphi \vee \neg\text{in}_b^\varphi))$, $\textsc{Sem}_{\text{ad}}^\varphi(\mathcal{A}) = \textsc{Sem}_{\text{cf}}^\varphi(\mathcal{A}) \wedge \bigwedge_{a,b\in\mathcal{A}}((\text{att}_{b,a} \wedge \text{in}_a^\varphi) \rightarrow \text{out}_b^\varphi)$, and $\textsc{Sem}_{\text{stb}}^\varphi(\mathcal{A}) = \textsc{Sem}_{\text{cf}}^\varphi(\mathcal{A}) \wedge \bigwedge_{a\in\mathcal{A}}(\text{in}_a^\varphi \vee \text{out}_a^\varphi)$. Finally, $\textsc{Aux}^\varphi(\Phi^-, w) = \bigwedge_{\psi\in\Phi^-, w(\psi)=\infty} \neg\psi^\varphi \wedge \bigwedge_{\psi\in\Phi^-, w(\psi)<\infty}(\neg\text{sat}_\psi \vee \neg\psi^\varphi)$, accounting for negative formulas by stating that for $\psi \in \Phi^-$, if $w(\psi) = \infty$ or sat_ψ is set to true, ψ is not satisfied by the witnessing extension for φ.[3] The weighted soft clauses φ_{soft} are $\{(\varphi, w(\varphi)) \mid \varphi \in \Phi^{\mathcal{R}}, w(\varphi) < \infty\} \cup \{(\varphi^\varphi, w(\varphi)) \mid \varphi \in \Phi^+, w(\varphi) < \infty\} \cup \{(\text{sat}_\varphi, w(\varphi)) \mid \varphi \in \Phi^-, w(\varphi) < \infty\}$, directly modeling the cost function of formula-based enforcement.

We iteratively solve the abstraction, querying a MaxSAT solver for a candidate solution (line 3). If no candidate solution exists, there is no solution to the formula-based enforcement problem either, and we return **false** (line 4). Otherwise, we extract the candidate attack structure $\mathcal{R}$ from the optimal MaxSAT solution τ via $\textsc{AttackStructure}(\tau) = \{(a, b) \in \mathcal{A} \times \mathcal{A} \mid \tau(\text{att}_{a,b}) = 1\}$. It remains to check whether $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ is an optimal solution to the formula-based enforcement problem by searching for a counterexample, namely an extension which satisfies a formula $\varphi \in \Phi^-$ for which either $w(\varphi) = \infty$ or $\tau(\text{sat}_\varphi) = 1$. We construct a propositional formula (line 6), defined as

$$\varphi_{\text{cex}} = \bigwedge_{a\in\mathcal{A}} (\text{out}_a \leftrightarrow \bigvee_{(b,a)\in\mathcal{R}} \text{in}_b) \wedge \textsc{Sem}_\sigma(\mathcal{F}) \wedge \Big(\bigvee_{\varphi\in\Phi^-, w(\varphi)=\infty} \varphi \vee \bigvee_{\varphi\in\Phi^-, w(\varphi)<\infty, \tau(\text{sat}_\varphi)=0} \varphi \Big).$$

Here, $\textsc{Sem}_\sigma(\mathcal{F})$ is the SAT encoding of semantics σ [Besnard and Doutre, 2004], i.e., $\textsc{Sem}_{\text{cf}}(\mathcal{F}) = \bigwedge_{(a,b)\in\mathcal{R}}(\text{in}_a \vee \text{in}_b)$, $\textsc{Sem}_{\text{ad}}(\mathcal{F}) = \textsc{Sem}_{\text{cf}}(\mathcal{F}) \wedge \bigwedge_{a\in\mathcal{A}}(\text{in}_a \rightarrow \bigvee_{(b,a)\in\mathcal{A}} \text{out}_b)$, and $\textsc{Sem}_{\text{stb}}(\mathcal{F}) = \textsc{Sem}_{\text{cf}}(\mathcal{F}) \wedge \bigwedge_{a\in\mathcal{A}}(\text{in}_a \vee \text{out}_a)$.

[2] The algorithm can be adapted to complete semantics as well; however, our focus is on the new form of credulous status enforcement, for which admissible and complete semantics coincide.

[3] When $\Phi^+ = \emptyset$, we use a trivial formula to encode an arbitrary σ-extension and add the same constraints on negative formulas.

We call a SAT solver on the formula φ_{cex} (line 7). If it is unsatisfiable, we return the current solution $\mathcal{F}$ (line 8) as an optimal solution to the formula enforcement problem. Otherwise, we extract the counterexample extension $\text{EXTENSION}(\tau_{\text{cex}}) = \{a \in \mathcal{A} \mid \tau_{\text{cex}}(\text{in}_a) = 1\}$ (line 9). We refine the abstraction by adding a constraint to φ_{hard} preventing $\mathcal{E}$ from reoccurring as an extension in subsequent candidate solutions. In particular, we adapt the strong refinement developed for extension and status enforcement [Niskanen and Järvisalo, 2020] to our setting. If there is a formula $\varphi \in \Phi^-$ with $w(\varphi) = \infty$ satisfied by $\mathcal{E}$, we add the clause

$$\text{REFINE}(\mathcal{F}, \mathcal{E}) = \bigvee_{\substack{(a,b)\in\mathcal{R} \\ a\in\mathcal{E}, b\in\mathcal{E}^O}} \neg\text{att}_{a,b} \vee \bigvee_{\substack{(a,b)\in(\mathcal{A}\times\mathcal{A})\setminus\mathcal{R} \\ a\notin\mathcal{E}^O, b\in\mathcal{E}}} \text{att}_{a,b}$$

requiring that we change the attack structure so that the counterexample extension $\mathcal{E}$ no longer exists. If, in turn, there is a formula $\varphi \in \Phi^-$ with $w(\varphi) < \infty$ and $\tau(\text{sat}_\varphi) = 0$ satisfied by $\mathcal{E}$, we instead add a set of clauses $(\text{sat}_\varphi \vee \text{REFINE}(\mathcal{F}, \mathcal{E}))$ for each such formula φ. These clauses require that in order to set sat_φ to false, we need to change the attack structure so that $\mathcal{E}$ is no longer an extension.

7 Empirical Evaluation

Our implementation of the MaxSAT-based CEGAR algorithm from Section 6 is available online in open source at https://bitbucket.org/andreasniskanen/phiaf. We use PySAT (version 1.8.dev23) [Ignatiev *et al.*, 2018] to interface with the MaxSAT solver RC2 [Ignatiev *et al.*, 2019] and the SAT solver Glucose (version 3.0) [Audemard and Simon, 2009]. The experiments were run on 2.50-GHz Intel(R) Xeon(R) Gold 6248 CPUs. We imposed a 600-second per-instance time and a 32-GB memory limit and measured CPU time via runsolver [Roussel, 2011]. We generated benchmark instances for credulous status enforcement with soft joint acceptance constraints as follows. Our initial AFs are based on the $D(n, p, q)$ random model [Gao, 2017] of AFs with n arguments, in which an attack between a pair of arguments exists with probability p, and is symmetric with conditional probability q. Given probability p, we set $q = q(p)$ to the threshold of the phase transition exhibited by the model (specifically so that $4q/(1+q)^2 = p$) to obtain nontrivial instances. For each $n = 20, 30, \ldots, 200$ and $p = 0.025, 0.05, 0.1, 0.2, 0.4$, we sample 10 AFs $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ from $D(n, p, q(p))$. For each AF, we generate the input $\mathcal{P}, \mathcal{N}, \mathcal{P}^\wedge, \mathcal{N}^\wedge$ as follows. For $\mathcal{P}$ and $\mathcal{N}$, we sample two sets S, T of $n/10$ arguments from $\mathcal{A}$ without replacement, and set $\mathcal{P} = (S \cup \bigcup \sigma(F)) \setminus T$ and $\mathcal{N} = \mathcal{A} \setminus \mathcal{P}$. Finally, for both $\mathcal{P}^\wedge$ and $\mathcal{N}^\wedge$, we sample 5 sets of size $n/10$ from $\mathcal{A} \setminus \mathcal{N}$. In total, this results in 950 input instances.

Our results are summarized in Table 1 for stable semantics and in Table 2 for admissible semantics. Under stable, the implementation scales up to 70 arguments, with $>$ 80 % of instances solved up to $n = 70$ with $<$ 65-second mean CPU times. Sparse input AFs seem to give rise to easier instances, as for $p = 0.025$ all instances are solvable up to $n = 90$, while for $p \in \{0.2, 0.4\}$ this is the case up to $n = 40$. In contrast, enforcement under admissible semantics is in general empirically harder, with $> 80\%$ of instances solved only up to $n = 30$. For an explanation, note that since we encode an extension for each argument in $(\bigcup \sigma(\mathcal{F}) \cup \mathcal{P}) \setminus \mathcal{N}$, $|\bigcup \sigma(\mathcal{F})|$ directly affects the size of the initial MaxSAT encoding. In our setup 147/950 initial AFs satisfy $|\bigcup \text{ad}(\mathcal{F})| - |\bigcup \text{stb}(\mathcal{F})| > 0$, and this difference is $\approx 0.17n$ on average when nonzero. In addition, the number of iterations (i.e., solver calls) is considerably higher, averaging ≈ 2100 under admissible and ≈ 46 under stable. Intuitively, since there are considerably more admissible than stable extensions, counterexamples are more likely. Both effects are more pronounced on sparse AFs, which due to minimizing syntactic distance result in sparse candidate AFs.

n	$p = 0.025$	$p = 0.05$	$p = 0.1$	$p = 0.2$	$p = 0.4$	all
20	**10** (0.36)	**10** (0.27)	**10** (0.31)	**10** (0.19)	**10** (0.39)	**50** (0.31)
30	**10** (1.03)	**10** (0.58)	**10** (0.39)	**10** (1.76)	**10** (0.37)	**50** (0.83)
40	**10** (2.53)	**10** (1.1)	**10** (7.01)	**10** (0.86)	**10** (32.22)	**50** (8.74)
50	**10** (5.06)	8 (2.36)	**10** (2.32)	**10** (4.06)	**10** (16.67)	48 (6.25)
60	**10** (6.43)	9 (5.91)	**10** (33.83)	9 (5.46)	8 (9.4)	46 (12.61)
70	**10** (17.03)	9 (44.86)	9 (2.74)	8 (10.08)	7 (64.66)	43 (26.32)
80	**10** (65.17)	**10** (21.46)	9 (20.07)	6 (6.08)	3 (362.82)	38 (57.15)
90	**10** (57.92)	8 (70.89)	6 (9.26)	4 (80.37)	1 (1.49)	29 (52.58)
100	8 (122.76)	6 (40.97)	6 (27.44)	6 (15.55)	1 (74.47)	27 (57.79)
110	**10** (146.73)	9 (72.45)	6 (68.38)	2 (3.64)	1 (558.19)	28 (110.54)
120	9 (113.46)	8 (122.41)	6 (144.58)	4 (147.75)	0 (–)	27 (128.1)
130	7 (182.38)	8 (84.24)	4 (17.15)	4 (14.53)	1 (463.44)	24 (105.87)
140	7 (185.13)	8 (101.92)	4 (43.14)	2 (8.34)	0 (–)	21 (109.55)
150	4 (176.43)	3 (211.55)	4 (15.3)	2 (13.58)	0 (–)	13 (109.9)
160	6 (366.49)	5 (74.93)	3 (19.15)	2 (17.8)	0 (–)	16 (166.66)
170	7 (274.04)	5 (164.21)	4 (27.14)	0 (–)	0 (–)	16 (177.99)
180	7 (160.92)	5 (181.54)	2 (27.35)	3 (29.75)	0 (–)	17 (128.12)
190	2 (293.61)	6 (90.13)	0 (–)	1 (46.45)	0 (–)	9 (130.49)
200	3 (56.28)	4 (211.68)	3 (41.2)	1 (93.7)	0 (–)	11 (112.08)

Table 1: Number of solved instances (out of 10) and mean CPU time over solved instances for each n and p under stable semantics.

n	$p = 0.025$	$p = 0.05$	$p = 0.1$	$p = 0.2$	$p = 0.4$	all
20	4 (257.73)	7 (103.61)	7 (18.2)	**10** (9.49)	**10** (1.58)	38 (52.48)
30	5 (187.55)	8 (86.15)	8 (2.66)	**10** (15.34)	**10** (3.3)	41 (44.75)
40	7 (76.01)	7 (88.56)	5 (42.69)	6 (54.39)	8 (1.47)	33 (51.62)
50	5 (167.03)	9 (47.41)	4 (22.72)	5 (119.32)	9 (48.46)	32 (74.54)
60	5 (11.3)	1 (79.04)	3 (94.14)	2 (180.47)	6 (120.5)	17 (88.35)
70	5 (151.73)	3 (48.88)	1 (458.73)	2 (228.68)	2 (361.65)	13 (195.74)
80	3 (26.48)	0 (–)	0 (–)	1 (8.61)	1 (70.58)	5 (31.73)
90	3 (157.39)	0 (–)	0 (–)	0 (–)	1 (4.88)	4 (119.26)

Table 2: Number of solved instances (out of 10) and mean CPU time over solved instances for n and p under admissible semantics. Instances beyond $n > 90$ reached the timeout limit.

8 Conclusion

We introduced a novel approach for argumentation dynamics based on the enforcement of formulas in the extensions of an abstract argumentation framework. This approach captures all existing notions of enforcement and synthesis proposed in the literature, as well as new variants that minimize both syntactic and semantic distance, such as status enforcement with soft joint acceptance constraints, which are not addressed by existing approaches and are introduced in this work. We pinpointed the exact complexity of the corresponding decision problem under five key semantics, and provided a computational approach for solving second-level-complete variants in the form of a MaxSAT-based CEGAR algorithm. Our evaluation demonstrates the scalability of the approach to 30-70 arguments.

Ethical Statement

There are no ethical issues.

Acknowledgments

Work financially supported by the Research Council of Finland (grant 347588), the EuroHPC Joint Undertaking and its members including top-up funding by Ministry of Education and Culture of Finland, and the French National Research Agency (ANR grants AIDAL ANR-22-CPJ1-0061-01, AGGREEY ANR-22-CE23-0005 and GRAIL ANR-25-CE23-5597). The authors thank the Finnish Computing Competence Infrastructure (FCCI) for computational and data storage resources.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EuroHPC. Neither the European Union nor EuroHPC can be held responsible for them.

References

[Amgoud *et al.*, 2008] Leila Amgoud, Yannis Dimopoulos, and Pavlos Moraitis. Making decisions through preference-based argumentation. In *KR 2008*, pages 113–123. AAAI Press, 2008.

[Atkinson and Bench-Capon, 2023] Katie Atkinson and Trevor J. M. Bench-Capon. ANGELIC II: An improved methodology for representing legal domain knowledge. In *ICAIL 2023*, pages 12–21. ACM, 2023.

[Audemard and Simon, 2009] Gilles Audemard and Laurent Simon. Predicting learnt clauses quality in modern SAT solvers. In *IJCAI 2009*, pages 399–404, 2009.

[Bacchus *et al.*, 2021] Fahiem Bacchus, Matti Järvisalo, and Ruben Martins. Maximum satisfiability. In *Handbook of Satisfiability - Second Edition*, chapter 24, pages 929–991. IOS Press, 2021.

[Baroni *et al.*, 2018] Pietro Baroni, Martin Caminada, and Massimiliano Giacomin. Abstract argumentation frameworks and their semantics. In *Handbook of Formal Argumentation*, chapter 4, pages 159–236. College Publications, 2018.

[Baumann and Brewka, 2010] Ringo Baumann and Gerhard Brewka. Expanding argumentation frameworks: Enforcing and monotonicity results. In *COMMA 2010*, pages 75–86. IOS Press, 2010.

[Baumann and Brewka, 2013] Ringo Baumann and Gerhard Brewka. Spectra in abstract argumentation: An analysis of minimal change. In *LPNMR 2013*, pages 174–186. Springer, 2013.

[Baumann and Brewka, 2019] Ringo Baumann and Gerhard Brewka. Extension removal in abstract argumentation - An axiomatic approach. In *AAAI 2019*, pages 2670–2677. AAAI Press, 2019.

[Baumann *et al.*, 2021] Ringo Baumann, Sylvie Doutre, Jean-Guy Mailly, and Johannes P. Wallner. Enforcement in formal argumentation. In *Handbook of Formal Argumentation*, volume 2, chapter 8, pages 445–510. College Publications, 2021.

[Baumann, 2012] Ringo Baumann. What does it take to enforce an argument? Minimal change in abstract argumentation. In *ECAI 2012*, pages 127–132. IOS Press, 2012.

[Besnard and Doutre, 2004] Philippe Besnard and Sylvie Doutre. Checking the acceptability of a set of arguments. In *NMR 2004*, pages 59–64, 2004.

[Besnard and Hunter, 2008] Philippe Besnard and Anthony Hunter. *Elements of Argumentation*. MIT Press, 2008.

[Borg and Bex, 2021] AnneMarie Borg and Floris Bex. Enforcing sets of formulas in structured argumentation. In *KR 2021*, pages 130–140, 2021.

[Clarke *et al.*, 2003] Edmund M. Clarke, Orna Grumberg, Somesh Jha, Yuan Lu, and Helmut Veith. Counterexample-guided abstraction refinement for symbolic model checking. *J. ACM*, 50(5):752–794, 2003.

[Clarke *et al.*, 2004] Edmund M. Clarke, Anubhav Gupta, and Ofer Strichman. SAT-based counterexample-guided abstraction refinement. *IEEE Trans. Comput. Aided Des. Integr. Circuits Syst.*, 23(7):1113–1123, 2004.

[Coste-Marquis *et al.*, 2014a] Sylvie Coste-Marquis, Sébastien Konieczny, Jean-Guy Mailly, and Pierre Marquis. On the revision of argumentation systems: Minimal change of arguments statuses. In *KR 2014*, pages 52–61. AAAI Press, 2014.

[Coste-Marquis *et al.*, 2014b] Sylvie Coste-Marquis, Sébastien Konieczny, Jean-Guy Mailly, and Pierre Marquis. A translation-based approach for revision of argumentation frameworks. In *JELIA 2014*, pages 397–411. Springer, 2014.

[Coste-Marquis *et al.*, 2015] Sylvie Coste-Marquis, Sébastien Konieczny, Jean-Guy Mailly, and Pierre Marquis. Extension enforcement in abstract argumentation as an optimization problem. In *IJCAI 2015*, pages 2876–2882. AAAI Press, 2015.

[Craandijk and Bex, 2022] Dennis Craandijk and Floris Bex. Enforcement heuristics for argumentation with deep reinforcement learning. In *AAAI 2022*, pages 5573–5581. AAAI Press, 2022.

[Dauphin and Satoh, 2018] Jérémie Dauphin and Ken Satoh. Dialogue games for enforcement of argument acceptance and rejection via attack removal. In *PRIMA 2018*, pages 449–457. Springer, 2018.

[Diller *et al.*, 2018] Martin Diller, Adrian Haret, Thomas Linsbichler, Stefan Rümmele, and Stefan Woltran. An extension-based approach to belief revision in abstract argumentation. *Int. J. Approx. Reason.*, 93:395–423, 2018.

[Dimopoulos *et al.*, 2019] Yannis Dimopoulos, Jean-Guy Mailly, and Pavlos Moraitis. Argumentation-based negotiation with incomplete opponent profiles. In *AAMAS 2019*, pages 1252–1260. IFAAMAS, 2019.

[Doutre and Mailly, 2018] Sylvie Doutre and Jean-Guy Mailly. Constraints and changes: A survey of abstract argumentation dynamics. *Argument Comput.*, 9(3):223–248, 2018.

[Doutre *et al.*, 2014] Sylvie Doutre, Andreas Herzig, and Laurent Perrussel. A dynamic logic framework for abstract argumentation. In *KR 2014*. AAAI Press, 2014.

[Dung, 1995] Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artif. Intell.*, 77(2):321–358, 1995.

[Dyrkolbotn, 2014] Sjur Kristoffer Dyrkolbotn. How to argue for anything: Enforcing arbitrary sets of labellings using AFs. In *KR 2014*, pages 626–629. AAAI Press, 2014.

[Gao, 2017] Yong Gao. A random model for argumentation framework: Phase transitions, empirical hardness, and heuristics. In *IJCAI 2017*, pages 503–509, 2017.

[Haret *et al.*, 2018] Adrian Haret, Johannes Peter Wallner, and Stefan Woltran. Two sides of the same coin: Belief revision and enforcing arguments. In *IJCAI 2018*, pages 1854–1860, 2018.

[Heine and Ulbricht, 2024] Anne-Marie Heine and Markus Ulbricht. From acceptance to rejection in abstract argumentation. In *KR 2024*, pages 431–441, 2024.

[Hunter, 2018] Anthony Hunter. Towards a framework for computational persuasion with applications in behaviour change. *Argument Comput.*, 9(1):15–40, 2018.

[Ignatiev *et al.*, 2018] Alexey Ignatiev, António Morgado, and João Marques-Silva. PySAT: A Python toolkit for prototyping with SAT oracles. In *SAT 2018*, LNCS, pages 428–437. Springer, 2018.

[Ignatiev *et al.*, 2019] Alexey Ignatiev, António Morgado, and João Marques-Silva. RC2: an efficient MaxSAT solver. *J. Satisf. Boolean Model. Comput.*, 11(1):53–64, 2019.

[Kakas *et al.*, 2019] Antonis C. Kakas, Pavlos Moraitis, and Nikolaos I. Spanoudakis. *GORGIAS*: Applying argumentation. *Argument Comput.*, 10(1):55–81, 2019.

[Kontarinis *et al.*, 2013] Dionysios Kontarinis, Elise Bonzon, Nicolas Maudet, Alan Perotti, Leon van der Torre, and Serena Villata. Rewriting rules for the computation of goal-oriented changes in an argumentation system. In *CLIMA 2013*, pages 51–68. Springer, 2013.

[Niskanen and Järvisalo, 2020] Andreas Niskanen and Matti Järvisalo. Strong refinements for hard problems in argumentation dynamics. In *ECAI 2020*, pages 841–848, 2020.

[Niskanen *et al.*, 2016] Andreas Niskanen, Johannes Peter Wallner, and Matti Järvisalo. Optimal status enforcement in abstract argumentation. In *IJCAI 2016*, pages 1216–1222. IJCAI/AAAI Press, 2016.

[Niskanen *et al.*, 2018] Andreas Niskanen, Johannes Peter Wallner, and Matti Järvisalo. Extension enforcement under grounded semantics in abstract argumentation. In *KR 2018*, pages 178–183. AAAI Press, 2018.

[Niskanen *et al.*, 2019] Andreas Niskanen, Johannes Peter Wallner, and Matti Järvisalo. Synthesizing argumentation frameworks from examples. *J. Artif. Intell. Res.*, 66:503–554, 2019.

[Roussel, 2011] Olivier Roussel. Controlling a solver execution with the runsolver tool. *J. Satisf. Boolean Model. Comput.*, 7(4):139–144, 2011.

[Ulbricht and Baumann, 2019] Markus Ulbricht and Ringo Baumann. If nothing is accepted - Repairing argumentation frameworks. *J. Artif. Intell. Res.*, 66:1099–1145, 2019.

[Wallner *et al.*, 2017] Johannes Peter Wallner, Andreas Niskanen, and Matti Järvisalo. Complexity results and algorithms for extension enforcement in abstract argumentation. *J. Artif. Intell. Res.*, 60:1–40, 2017.

[Xu *et al.*, 2018] Kang Xu, Beishui Liao, and Huaxin Huang. Updating argumentation frameworks for enforcing extensions. In *CLAR 2018*, pages 63–79. Springer, 2018.