# A Logical Theory about Dynamics in Abstract Argumentation

Richard Booth[1], Souhila Kaci[2], Tjitze Rienstra[1,2], and Leendert van der Torre[1]

[1] Université du Luxembourg
6 rue Richard Coudenhove-Kalergi, Luxembourg
{richard.booth,tjitze.rienstra,leon.vandertorre}@uni.lu
[2] LIRMM (CNRS/Université Montpellier 2)
161 rue Ada, Montpellier, France
souhila.kaci@lirmm.fr

**Abstract.** We address dynamics in abstract argumentation using a logical theory where an agent's belief state consists of an argumentation framework (AF, for short) and a constraint that encodes the outcome the agent believes the AF *should* have. Dynamics enters in two ways: (1) the constraint is strengthened upon learning that the AF should have a certain outcome and (2) the AF is expanded upon learning about new arguments/attacks. A problem faced in this setting is that a constraint may be inconsistent with the AF's outcome. We discuss two ways to address this problem: First, it is still possible to form consistent *fallback beliefs*, i.e., beliefs that are most plausible given the agent's AF and constraint. Second, we show that it is always possible to find AF expansions to restore consistency. Our work combines various individual approaches in the literature on argumentation dynamics in a general setting.

**Keywords:** Argumentation, Dynamics, Knowledge Representation.

## 1 Introduction

In Dung-style argumentation [1] the argumentation framework (AF for short) is usually assumed to be static. There are, however, many scenarios where argumentation is a dynamic process: Agents may learn that an AF must have a certain outcome and may learn about new arguments/attacks. These are two basic issues that a theory about argumentation dynamics should address.

Some of these aspects have received attention in recent years. For example, the so called *enforcing problem* [2] is concerned with the question of whether and how an AF can be modified to make a certain set of arguments accepted. Other work studies the impact on the outcome of an AF when a new argument comes into play [3] or studies the issue of reasoning with incomplete AFs [4].

We address the problem by answering the following research questions: *How can we model an agent's belief about the outcome of an AF?* and *How can we characterize the effects of an agent learning that the AF should have a certain outcome, or learning about new arguments/attacks?*

W. Liu, V.S. Subrahmanian, and J. Wijsen (Eds.): SUM 2013, LNAI 8078, pp. 148–161, 2013.

The basis of our approach is a logical *labeling language*, interpreted by labelings that assign to each argument a label indicating that it is *accepted, rejected* or *undecided* [5]. Formulas in this language are statements about the acceptance of the arguments of an AF. This allows us to reason about the outcome of an AF in terms of beliefs, rather than extensions or labelings.

We take an agent's belief state to consist of an AF and a formula encoding a constraint on the outcome of the AF. The constraint is strengthened upon learning that the AF should have a certain outcome. Furthermore, the agent's AF is expanded upon learning about new arguments and attacks. These two operations are modeled by a *constraint expansion* and *AF expansion* operator.

A problem faced in this setting is that the constraint on the AF's outcome may be inconsistent with its actual outcome, preventing the agent from forming consistent beliefs. We call such a state *incoherent*. We appeal to the intuition that an AF provides the agent with the ability to argue for the plausibility of the beliefs that it induces. Incoherence thus means that the agent is unable to argue for the plausibility of her beliefs using the AF.

We show that there are two ways to deal with this. First, we show that, given an incoherent belief state, it is always possible to come up with an expansion of the AF that restores coherence. Such AF expansions can be thought of as providing the missing arguments necessary to argue for her beliefs. Second, we show that it is always possible to form consistent *fallback beliefs*, which represent the "most rational" outcome of the agent's AF, given the constraint. Finally, we present an answer-set program for computing fallback belief, i.e., for determining whether or not some formula is a fallback belief in a particular belief state.

Our theory about argumentation dynamics combines several individual approaches in the literature in a general setting. For example, the issue of restoring coherence is related to the enforcing problem [2]; other ways to characterize the effect of an AF expansion have been studied in [3] and our notion of fallback belief is related to principles developed in [4].

A brief outline of this paper: In section 2 we introduce our labeling logic, together with the necessary basics of argumentation theory. Next, we present our belief state model and associated expansion operators in section 3. We then discuss in sections 4 and 5 how to deal with incoherent belief states, i.e., by restoring coherence via AF expansion and by using fallback belief. In section 6 we present an ASP encoding for computing fallback belief. Having focused in these sections on the complete semantics, we turn in section 7 to a discussion of a number of additional semantics. In section 8 we discuss related work and we conclude and discuss future work in section 9.

## 2 Preliminaries

We start out with some preliminaries concerning Dung-style abstract argumentation theory [1]. According to this theory, argumentation can be modeled using an *argumentation framework*, which captures two basic notions, namely arguments and attacks among arguments. We limit ourselves to the abstract setting,

meaning that we do not specify the content of arguments in a formal way. Nevertheless, arguments should be understood to consist of a *claim* and a *reason*, i.e., some consideration that counts in favor of believing the claim to be true, while attacks among arguments stem from conflicts between different claims and reasons. We assume in this paper that argumentation frameworks are finite.

**Definition 1.** *An* argumentation framework *(AF for short) is a pair* $(A, R)$ *where* $A$ *is a finite set of* arguments *and* $R \subseteq A \times A$ *is an* attack *relation.*

Given an AF $(A, R)$ we say that $x$ is an *attacker* of $y$, whenever $(x, y) \in R$. The outcome of an AF consists of possible points of view on the acceptability of its arguments. In the literature, these points of view are represented either by sets of acceptable arguments, called *extensions* or by *argument labelings*, which are functions assigning to each argument a label *in*, *out* or *undecided*, indicating that the argument is respectively accepted, rejected or neither [5]. The two representations are essentially reformulations of the same idea as they can be mapped 1-to-1 such that extensions correspond to sets of in-labeled arguments [5]. For the current purpose we choose to adopt the labeling-based approach.

**Definition 2.** *A* labeling *of an AF* $F = (A, R)$ *is a function* $L : A \to \{I, O, U\}$. *We denote by* $I(L), O(L)$ *and* $U(L)$ *the set of all arguments* $x \in A$ *such that* $L(x) = I$, $L(x) = O$ *or* $L(x) = U$, *respectively, and by* $\mathcal{M}_F$ *the set of all labelings of* $F$.

Various conditions are used to single out labelings that represent rational points of view. The following gives rise to what is called the *complete* semantics:

**Definition 3.** *Let* $F = (A, R)$ *be an AF and* $L \in \mathcal{M}_F$ *a labeling. We say that* $L$ *is* complete *iff for each* $x \in A$ *it holds that:*
- $L(x) = I$ *iff* $\forall y \in A$ *s.t.* $(y, x) \in R$, $L(y) = O$,
- $L(x) = O$ *iff* $\exists y \in A$ *s.t.* $(y, x) \in R$ *and* $L(y) = I$,

Thus, under the complete semantics, the outcome of an AF consists of labelings in which an argument is in iff its attackers are out and is out iff it has an attacker that is in. Many of the other semantics proposed in the literature, such as the *grounded*, *preferred* and *stable* semantics [1] are based on selecting particular subsets of the set of complete labelings:

**Definition 4.** *Let* $L$ *be a complete labeling of the AF* $F$. $L$ *is called:*
- grounded *iff there is no complete labeling* $L'$ *of* $F$ *s.t.* $I(L') \subset I(L)$,
- preferred *iff there is no complete labeling* $L'$ *of* $F$ *s.t.* $I(L) \subset I(L')$,
- stable *iff* $U(L) = \emptyset$.

We focus on the complete semantics but briefly discuss the others in section 7.

*Example 1.* Consider the AF shown in figure 1, which has three complete labelings, namely `IOOI`, `OIOI` and `UUUU`. (We denote labelings by strings of the form `ABC`... where `A`, `B`, `C`, ... are the labels of the arguments $a, b, c, \ldots$)

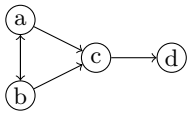


**Fig. 1.** An argumentation framework

A flexible way to reason about the outcome of an AF is by using a logical *labeling language*. Formulas in this language assign a label to an argument or are boolean combinations of such assignments. The language, given an AF $F = (A, R)$, is denoted by $\mathcal{L}_F$ and is generated by the following BNF, where $x \in A$:

$$\phi := \mathtt{in}_x \mid \mathtt{out}_x \mid \mathtt{u}_x \mid \neg\phi \mid \phi \vee \phi \mid \top \mid \bot.$$

We also use the connectives $\wedge, \rightarrow, \leftrightarrow$, defined as usual in terms of $\neg$ and $\vee$. Next, we define a *satisfaction* relation between labelings and formulas:

**Definition 5.** *Let $F$ be an AF. The satisfaction relation $\models_F \subseteq \mathcal{M}_F \times \mathcal{L}_F$ is defined by:*
- $L \models_F \mathtt{in}_x$ *iff* $L(x) = I$,
- $L \models_F \mathtt{out}_x$ *iff* $L(x) = O$,
- $L \models_F \mathtt{u}_x$ *iff* $L(x) = U$,
- $L \models_F \phi \vee \psi$ *iff* $L \models_F \phi$ *or* $L \models_F \psi$,
- $L \models_F \neg\phi$ *iff* $L \not\models_F \phi$,
- $L \models_F \top$ *and* $L \not\models_F \bot$.

*A* model *of a formula $\phi$ is a labeling $L \in \mathcal{M}_F$ such that $L \models_F \phi$. We denote by $[\phi]_F$ the set of labelings satisfying $\phi$, defined by $[\phi]_F = \{L \in \mathcal{M}_F \mid L \models_F \phi\}$. We write $\phi \models_F \psi$ iff $[\phi]_F \subseteq [\psi]_F$ and $\phi \equiv_F \psi$ iff $[\phi]_F = [\psi]_F$.*

Whenever the AF we talk about is clear from the context, we drop the subscript $F$ from $\models_F$, $[\dots]_F$ and $\equiv_F$.

Using this labeling language, we can reason about the outcome of an AF by talking about *beliefs* induced by the AF. These beliefs can be represented by a formula $\phi$ such that $[\phi]$ is exactly the set of complete labelings of $F$. It is worth noting that $\phi$ can be formulated in a straightforward way:

**Proposition 1.** *Let $F = (A, R)$ be an AF. It holds that a labeling $L$ is a complete labeling of $F$ iff $L$ is a model of the formula*

$$\wedge_{x \in A}((\mathtt{in}_x \leftrightarrow (\wedge_{(y,x) \in R} \mathtt{out}_y)) \wedge (\mathtt{out}_x \leftrightarrow (\vee_{(y,x) \in R} \mathtt{in}_y))).$$

*Example 2.* Among the beliefs induced by AF in figure 1 are $\neg\mathtt{out}_d$ and $(\mathtt{in}_a \vee \mathtt{in}_b) \leftrightarrow \mathtt{in}_d$ and $\neg(\mathtt{in}_a \wedge \mathtt{in}_b)$.

Finally, *conflict-freeness* is considered to be a necessary (but not sufficient) condition for any labeling to be considered rational. We will make use of the following definition:

**Definition 6.** *Let $F = (A, R)$ be an AF. A labeling $L$ of $F$ is said to be* conflict-free *iff $L$ is a model of the formula*

$$\wedge_{x \in A}(\mathtt{in}_x \rightarrow ((\wedge_{(y,x) \in R} \mathtt{out}_y) \wedge (\wedge_{(x,y) \in R} \mathtt{out}_y))).$$

*We denote this formula by $Cf_F$. We say that $\phi$ is is conflict-free iff $Cf_F \not\models \neg\phi$.*

Thus, in a conflict-free labeling any neighbor of an in-labeled argument is out. Note that we deviate from the usual definition (see e.g. [6]), which allows neighbors of an in-labeled argument to be undecided. The reason is that, given our definition, conflict-freeness can be seen to generalize completeness in a dynamic setting, in the sense that a conflict-free labeling of an AF is always (part of) a complete labeling of some expansion of the AF. The benefit of this will become clear in the following sections.

*Example 3.* Some examples of conflict-free labelings of the AF in figure 1 are `IOOO`, `UUOI` and `OOOO`. Examples of labelings that are not are `IIOO` and `UUIO`.

## 3 Belief States

On the one hand, AFs interpreted under the complete semantics induce beliefs about the status of arguments (and consequently about argument's claims and reasons) that are rational in the sense that the arguments and attacks in the AF can be used to argue for the plausibility of these beliefs. For example, given the AF $(\{b, a\}, \{(b, a)\})$, the belief $\mathtt{out}_a$ can, informally speaking, be argued for by pointing out that $a$ is attacked by $b$ which, in turn, is not attacked and should thus be accepted. Furthermore, these beliefs are defeasible, because learning about new arguments and attacks may cause old beliefs to be retracted.

On the other hand, an agent may learn or come to desire some claim to be true or false, without being aware of arguments to argue for the plausibility of it. This bears on the outcome that the AF *should* have, according to the agent. To model scenarios like these, we define an agent's belief state to consist not only of an AF, but also a constraint that the agent puts on its outcome.

**Definition 7.** *A belief state is a pair $S = (F, K)$, where $F = (A, R)$ is an AF and $K \in \mathcal{L}_F$ the agent's* constraint. *We define $K(S)$ by $K(S) = K$ and $Bel(S)$ by $[Bel(S)] = \{L \in [K] \mid L \text{ is a complete labeling of } F\}$. We say that the agent* believes *$\psi$ iff $Bel(S) \models \psi$ and that $S$ is* coherent *iff $Bel(S) \not\models \bot$.*

Thus, the belief $Bel(S)$ of an agent is formed by the outcome of the AF in conjunction with the constraint. Intuitively, the plausibility of the agent's belief can be argued for only if it is consistent, i.e., only if the belief state is coherent. An incoherent state is thus a state in which the agent is prevented from forming beliefs that can be shown to be plausible via the AF.

We turn again to incoherence in the following section. We first define two expansion operators: one that strengthens the agent's constraint and one that expands the AF. The *constraint expansion operator* takes as input a belief state and a formula $\phi$ representing a constraint that is to be incorporated into the new belief state. It is defined as follows.

**Definition 8.** *Let $F$ be an AF, $S = (F, K)$ a belief state and $\phi \in \mathcal{L}_F$. The* constraint expansion *of $S$ by $\phi$, denoted $S \oplus \phi$ is defined by $S \oplus \phi = (F, K \wedge \phi)$.*

*Example 4.* Let $S_1 = (F, \top)$ where $F$ is the AF shown in figure 1. We do not have $Bel(S_1) \models \mathtt{in}_d$. That is, the agent does not believe that $d$ is in. Consider the constraint expansion $S_2 = S_1 \oplus (\mathtt{in}_a \vee \mathtt{in}_b)$. Now we have $Bel(S_2) \models \mathtt{in}_d$. That is, after learning that either $a$ or $b$ is in, the agent believes that $d$ is in.

As to expanding the AF, we make two assumptions: First, we assume that arguments and attacks are not "forgotten". This means that elements can be added to an AF but not removed. Second, we assume that attacks between arguments are determined once the arguments are known. This means that no new attacks can be added between arguments already present in the agent's AF. Such expansions are called *normal expansions* by Baumann and Brewka [2]. We call a set of new arguments and attacks an *AF update*:

**Definition 9.** *Let $F = (A, R)$ be an AF. An* AF update *for $F$ is a pair $F^* = (A^*, R^*)$ where $A^*$ is a set of* added arguments*, such that $A \cap A^* = \emptyset$ and $R^* \subseteq ((A \cup A^*) \times (A \cup A^*)) \setminus (A \times A)$ a set of* added attacks*.*

The *AF expansion operator* is defined as follows:

**Definition 10.** *Let $F = (A, R)$ be an AF, $S = (F, K)$ a belief state and $F^* = (A^*, R^*)$ an AF update for $F$. The* AF expansion *of $S$ by $F^*$, denoted by $S \otimes F^*$ is defined by $S \otimes F^* = ((A \cup A^*, R \cup R^*), K)$.*

*Example 5.* Consider the belief state $S_1 = (F, \mathtt{out}_a \vee \mathtt{out}_b)$ where $F$ is the AF shown in figure 1. Note that we do not have, e.g., $Bel(S_1) \models \mathtt{in}_b$. Now consider the AF expansion $S_2 = (S_1 \otimes (\{e\}, \{(e, a)\})$. Now we do have $Bel(S_2) \models \mathtt{in}_b$.

The two operators just defined allow us to study our belief state model in a dynamic setting, where an agent's belief state changes after new constraints on the AF's outcome are acquired or after adding new arguments and attacks.

## 4 Restoring Coherence through AF Expansion

In the previous section we presented a belief state model which includes, besides the agent's AF, a constraint on its outcome. We also explained that incoherence (i.e., the belief induced by the AF being inconsistent with the constraint) prevents the agent from forming beliefs that can be shown to be plausible via the agent's AF. The question is then: can the AF be expanded in such a way that the beliefs induced by it *are* consistent with the agent's constraints? In other words: can we restore coherence by expanding the AF in some way? Consider the following example.

*Example 6.* Let $S_1 = (F, \top)$ where $F$ is the AF shown in figure 1. Suppose the agent learns that both $a$ and $b$ are out. The resulting state $S_2 = S_1 \oplus (\mathtt{out}_a \wedge \mathtt{out}_b)$ is incoherent, i.e., we have $Bel(S_2) \models \bot$. Now suppose the agent learns about

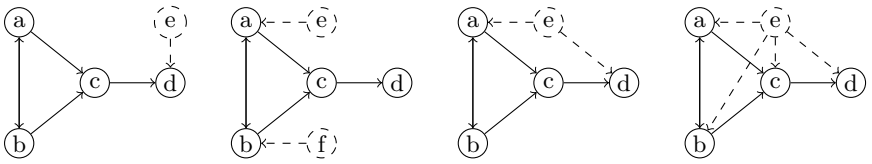


**Fig. 2.** Four argumentation framework updates

arguments $e$ and $f$, attacking $a$ and $b$. The corresponding AF update is shown as $F_2^*$ in figure 2. The resulting state is $S_3 = S_2 \otimes (\{e, f\}, \{(e, a), (f, b)\})$. Coherence is now restored: $Bel(S_3) \not\models \bot$. In $S_3$ the agent believes, e.g., that $c$ is in and $d$ is out: $Bel(S_3) \models \mathtt{in}_c \wedge \mathtt{out}_d$. Notice that $F_4^*$, too, restores coherence in state $S_2$, whereas $F_1^*$ and $F_3^*$ do not.

This example shows that it is indeed possible to expand an AF such that coherence is restored. Note, also, that the AF updates $F_2^*$ and $F_4^*$ can be understood to provide the "missing explanation" for the agent's constraint $\mathtt{out}_a \wedge \mathtt{out}_b$. That is, $a$ and $b$ are out *because* there are arguments attacking (among possibly other arguments) $a$ and $b$. We can show that, as long as the agent's constraint is conflict-free, there always exists some AF expansion that restores coherence. That the agent's constraint is required to be conflict-free follows from the fact that attacks between existing arguments cannot be removed. Proofs are omitted due to space constraints.

**Theorem 1.** *Let $(F, K)$ be an incoherent belief state where $K$ is conflict-free. There exists an AF update $F^*$ for $F$ such that $(F, K) \otimes F^*$ is coherent.*

This result essentially says that incoherence of a belief state can be understood to mean that the agent's AF is incomplete and needs to be expanded with additional arguments and attacks. A related result, called the *conservative strong enforcing* result, was presented by Baumann and Brewka [2]. However, this result deals only with the possibility of making some set of arguments accepted. By contrast, we deal with arbitrary formulas expressible in the logical labeling language.

## 5 Fallback Belief

In example 6, the agent learns that $a$ and $b$ are out, resulting in the belief state becoming incoherent and beliefs becoming inconsistent. Nevertheless, it is still possible to form reasonable, consistent beliefs given this constraint, even without performing a coherence restoring AF expansion. To see what we mean, it is enough to just look at the AF in figure 1 and see that, once $a$ and $b$ are out, $c$ should be in and $d$ should be out. However, there are no complete labelings

satisfying these assignments of labels. Thus to form such beliefs, which we call *fallback* beliefs, we must adopt a different method.

The starting point is to define a *rationality order* over conflict-free labelings, used to determine their relative rationality. Consider an assignment, to each AF $F$, of a total pre-order (i.e., a complete, transitive and reflexive order) $\preceq_F$ over conflict-free labelings of $F$. Given a set $M \subseteq [Cf_F]$ we define $min_{\preceq_F}(M)$ by $min_{\preceq_F}(M) = \{L \in M \mid \forall L' \in M, L \preceq_F L'\}$. Following terminology used in belief revision, we call such an assignment *faithful* if the minimal labelings according to $\preceq_F$ are exactly the complete labelings of $F$.

**Definition 11.** *A* faithful assignment *assigns to each AF $F$ a total pre-order $\preceq_F \subseteq [Cf_F] \times [Cf_F]$ s.t. $L \in min_{\preceq_F}([Cf_F])$ iff $L$ is a complete labeling of $F$. If $L \preceq_F L'$, we say that $L$ is* at least as rational *as $L'$.*

In an incoherent state, i.e., when all fully rational labelings of the AF $F$ are ruled out, the agent can fall back on the remaining labelings that are most rational according to the ordering $\preceq_F$. These labelings can be used to form fallback beliefs, the idea being that they represent the best outcome of the AF given the agent's constraint. Given a belief state $S$, we denote the fallback belief in $S$ by $Bel^*(S)$. The type of belief we end up with can be characterized by an appropriate adaptation of the well known KM postulates [7]:

**Theorem 2.** *The following are equivalent:*

1. *There exists a faithful assignment mapping each $F$ to a total pre-order $\preceq_F$ such that for each $K$, $[Bel^*((F,K))] = min_{\preceq_F}([K] \cap [Cf_F])$.*
2. *For each $S = (F,K)$, $Bel^*$ satisfies:*
   *P1: $Bel^*(S) \models K(S) \wedge Cf_F$.*
   *P2: If $S$ is coherent then $Bel^*(S) \equiv Bel(S)$.*
   *P3: If $K(S)$ is conflict-free then $Bel^*(S)$ is conflict-free.*
   *P4: If $F_1 = F_2$ and $K_1 \equiv K_2$ then $Bel^*((F_1,K_1)) \equiv Bel^*((F_2,K_2))$.*
   *P5: $Bel^*(S) \wedge \psi \models Bel^*(S \oplus \psi)$.*
   *P6: If $Bel^*(S) \wedge \psi$ is conflict-free then $Bel^*(S \oplus \psi) \models Bel^*(S) \wedge \psi$.*

Thus, if we define $Bel^*$ by $[Bel^*((F,K))] = min_{\preceq_F}([K] \cap [Cf_F])$ then fallback belief behaves like the one-shot revision, by the constraint $K$, of the outcome of $F$ under the complete semantics. The postulates in proposition 2 now embody conditions of minimal change w.r.t. the fully rational outcome of the AF, rather than an arbitrary KB. The original postulates were discussed by Katsuno and Mendelzon [7], who built on the AGM approach to belief revision [8]. Here we content ourselves with pointing out how our postulates differ from the original ones. First of all, P1, P2, P3 and P6 are changed to account for the fact that only conflict-free labelings are considered possible. Second, in P5 and P6 conjunction is substituted with $\oplus$, Finally, P4 requires the AFs (and thus orderings) in the two belief states to be equivalent, as well as the constraint.

The question we need to answer now is: when is one conflict-free labeling of an AF $F$ to be more rational than another? That is, how should $\preceq_F$ order arbitrary conflict-free labelings of $F$? A natural way to do this is by looking at the arguments that are *illegally* labeled [6]. This is defined as follows:

**Definition 12.** *Let $F = (A, R)$ be an AF and $L \in \mathcal{M}_F$ a labeling of $F$. An argument $x \in A$ is said to be:*
- Illegally in *iff* $L(x) = I$ *and* $\exists y \in A, (y, x) \in R$ *and* $L(y) \neq O$,
- Illegally out *iff* $L(x) = O$ *and* $\nexists y \in A, (y, x) \in R$ *such that* $L(y) = I$,
- Illegally undecided *iff* $L(x) = U$ *and* $\exists y \in A, (y, x) \in R$ *and* $L(y) = I$ *or* $\nexists y \in A, (y, x) \in R$ *such that* $L(y) = U$.

*We denote by $Z_F^I(L), Z_F^O(L)$ and $Z_F^U(L)$ the sets of arguments that are, respectively, illegally in, out and undecided in $L$.*

Intuitively, an illegally labeled argument indicates a local violation of the condition imposed on the argument's label according to the complete semantics. It can be checked, for example, that a labeling $L$ is a complete labeling iff it has no arguments illegally labeled. It can also be checked that, in a conflict-free labeling, arguments are never illegally in. Thus in judging the relative rationality of a conflict-free labeling $L$, we only have to look at the sets $Z_F^O(L)$ and $Z_F^U(L)$.

What, exactly, do the sets $Z_F^O(L)$ and $Z_F^U(L)$ tell us about how rational $L$ is? To answer this we have to look at what it takes to turn $L$ into a complete labeling. We say that an AF update that turns $L$ into (part of) a complete labeling of the (expanded) AF is an AF update that *completes* $L$. Formally:

**Definition 13.** *Let $F^* = (A^*, R^*)$ be an AF update for $F = (A, R)$ and $L$ a conflict-free labeling of $F$. We say that $F^*$* completes *$L$ iff there is a complete labeling $L'$ of the AF $(A \cup A^*, R \cup R^*)$ such that $(L' \downarrow A) = L$, where $(L \downarrow A)$ is a function defined by $(L \downarrow A)(x) = L(x)$, for all $x \in A$.*

As a measure for the "impact" of an AF update, Baumann looked at the number of added attacks [9]. In our setting it is more appropriate to look at the number of arguments in the existing AF that are attacked by the AF update. We call this this the *attack degree* of the AF update.

**Definition 14.** *Let $F^* = (A^*, R^*)$ be an AF update for $F = (A, R)$. We denote by $\delta_F(F^*)$ the* attack degree *of $F^*$, defined by $\delta_F(F^*) = |\{x \in A \mid \exists y \in A^*, (y, x) \in R^*\}|$.*

The key is that the sets $Z_F^O(L)$ and $Z_F^U(L)$ inform us about the minimal impact it would take to complete $L$, or to turn $L$ into a fully rational point of view. That is, it informs us about the minimal attack degree of an AF update that completes $L$:

**Proposition 2.** *Let $L$ be a conflict-free labeling of an AF $F$. If $F^*$ completes $L$ then $\delta_F(F^*) \geq |Z_F^O(L) \cup Z_F^U(L)|$.*

We use the cardinality of the sets $Z_F^O(L)$ and $Z_F^U(L)$ as the criterion to define the rationality order $\preceq_F$, making the assumption that the agent believes that conflict-free labelings that require less impact to be turned into a complete labeling are more rational. We now define a faithful assignment as follows: Let $F$ be an AF and $L, L' \in [Cf_F]$,

$$L \preceq_F L' \text{ iff } |Z_F^O(L) \cup Z_F^U(L)| \leq |Z_F^O(L') \cup Z_F^U(L')|$$

Now, the outcome of the AF according to the agent's fallback belief is the outcome that would hold if some minimal impact, coherence restoring AF update would be performed.

*Example 7.* The table below represents $\preceq_F$ for the AF $F$ shown in figure 1.

| 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| OIOI | OIO$\underline{\text{O}}$ UUU$\underline{\text{O}}$ | $\underline{\text{OO}}$IO UU$\underline{\text{OO}}$ | $\underline{\text{OOO}}$I $\underline{\text{OUU}}$O | $\underline{\text{OOOO}}$ $\underline{\text{OUOU}}$ $\underline{\text{OUOO}}$ |
| UUUU | OIO$\underline{\text{U}}$ IOO$\underline{\text{O}}$ | $\underline{\text{OU}}$UU UU$\underline{\text{OU}}$ | $\underline{\text{OOU}}$U $\underline{\text{UOO}}$I | $\underline{\text{OOOU}}$ $\underline{\text{UOOO}}$ |
| IOOI | UU$\underline{\text{O}}$I IOO$\underline{\text{U}}$ | $\underline{\text{UO}}$UU | $\underline{\text{OUO}}$I $\underline{\text{UOU}}$O | $\underline{\text{OOUO}}$ $\underline{\text{UOOU}}$ |

The table groups labelings by to the number of arguments illegally labeled. These arguments are underlined and the numbers are shown in the column headers. This determines the ordering $\preceq_F$ as follows: $L \prec_F L'$ iff $L$ is in another column to the left of $L'$. We have the following fallback beliefs:

- $Bel^*(F, \mathtt{out}_a \wedge \mathtt{out}_b) \models \mathtt{in}_c$ (if $a$ and $b$ are out then $c$ is in).
- $Bel^*(F, \mathtt{in}_c) \models \mathtt{out}_a \wedge \mathtt{out}_b$ (if $c$ is in then $a$ and $b$ must be out).
- $Bel^*(F, \mathtt{out}_d) \models \neg(\mathtt{in}_a \wedge \mathtt{in}_b)$ (even if $d$ is out, $a$ and $b$ cannot both be in).
- $Bel^*(F, \mathtt{out}_d) \models \mathtt{u}_a \rightarrow \mathtt{u}_c$ (even if $d$ is out, if $a$ is undecided then so is $c$).

Note that none of these inferences can be made by looking only at the complete labelings of $F$.

As the following theorem states more formally, and as we pointed out above, fallback belief is formed by assuming the most rational outcome of an AF in an incoherent state to be the outcome that would hold after a coherence restoring AF update with minimal impact. That is, if coherence is restored using an AF update with a minimal attack degree, then the agent's regular belief in the updated state includes the agent's fallback belief in the old state.

**Theorem 3.** *Let $S$ be an incoherent belief state and $F_1^*$ a minimal coherence restoring update (i.e., $S \otimes F_1^*$ is coherent and there is no $F_2^*$ such that $S \otimes F_2^*$ is coherent and $\delta_F(F_2^*) < \delta_F(F_1^*)$). It holds that $Bel(S \otimes F_1^*) \models Bel^*(S)$.*

*Example 8.* Let $S = (F, \mathtt{out}_c)$ be a belief state with $F = (\{a, b, c, d, e\}, \{(a, b), (b, c), (d, e), (e, c)\})$. We have $[Bel^*(S)] = \{\mathtt{IO}\underline{\mathtt{O}}\mathtt{IO}, \underline{\mathtt{O}}\mathtt{IOIO}, \mathtt{IOO}\underline{\mathtt{O}}\mathtt{I}\}$. Three minimal coherence restoring AF updates are: $F_1^* = (\{f\}, \{(f, c)\})$, $F_2^* = (\{f\}, \{(f, a)\})$ and $F_3^* = (\{f\}, \{(f, d)\})$. We have that $Bel(S \otimes F_n^*) = \psi$, where $[\psi] = \{\mathtt{IOOIOI}\}$ if $n = 1$; $[\psi] = \{\mathtt{OIOIO}\}$, if $n = 2$ and $[\psi] = \{\mathtt{IOOOII}\}$, if $n = 3$. It can be checked that, for all $n \in \{1, 2, 3\}$, $Bel(S \otimes F_n^*) \models \phi$ and thus $Bel(S \otimes F_n^*) \models Bel^*(S)$.

## 6 Computing Fallback Beliefs with ASP

Answer-set programming has proven to be a useful mechanism to compute extensions of AFs under various semantics [10,11,12]. The idea is to encode both the AF and a so called *encoding* of the semantics in a single program of which the stable models correspond to the extensions of the AF.

In this section we show that the problem of deciding whether a formula $\phi$ is a fallback belief in a state $(F, K)$ can be solved, too, using an answer-set program. The encoding, shown in listing 6, turns out to be surprisingly simple, and

works as follows. The AF is assumed to be encoded (line 1) using the predicates `arg/1` and `att/2`. For example, the AF of figure 1 is encoded by the facts `arg(a)`, `arg(b)`, `arg(c)`, `arg(d)`, `att(a,b)`, `att(a,c)`, `att(b,a)`, `att(b,c)` and `att(c,d)`. The choice rule on line 2 ensures that each argument $x \in A$ gets one of three labels, expressed by the predicates `in/1`, `out/1` and `undec/1`. On lines 3 and 4 conflict-freeness is ensured. Given just these constraints, stable models correspond to conflict-free labelings of $F$. Lines 5-10 are used to establish whether an argument $x \in A$ is illegally labeled, expressed by the predicate `illegal(x)`. The cardinality of this predicate is minimized on line 12. Finally, the agent's constraint is assumed to be encoded (line 11) using statements restricting the possible labels assigned to arguments. For example, the constraint $\texttt{out}_a \vee \texttt{out}_b$ is encoded by the choice rule `1 {out(a), out(b)} 2`, and the constraint $\texttt{out}_a \wedge \texttt{out}_b$ by the two facts `out(a)` and `out(b)`. The (optimal) stable models now correspond to maximally rational conflict-free labelings that satisfy the constraint.

```
 1  % <-- Framework encoding here -->
 2  1 { in(X), out(X), undec(X) } 1 :- arg(X).
 3  out(Y) :- att(X, Y), in(X).
 4  out(X) :- att(X, Y), in(Y).
 5  legally_out(X) :- out(X), att(Y, X), in(Y).
 6  legally_undec(X) :- undec(X), att(Y, X), undec(Y).
 7  illegally_out(X) :- out(X), not legally_out(X).
 8  illegally_undec(X) :- undec(X), not legally_undec(X).
 9  illegal(X) :- illegally_out(X).
10  illegal(X) :- illegally_undec(X).
11  % <-- Constraint encoding here -->
12  #minimize { illegal(X) }.
```

Program 1. An answer set program to compute fallback belief

The program is compatible with the *Gringo* grounder (version 3.0.5) and *Clasp* answer set solver (version 2.1.2) [13]. The optimal stable models can be obtained by running the solver with the option `--opt-all`. The final step of the complete procedure amounts to checking whether the formula $\phi$ is true in every optimal stable model. Alternatively, the set of stable models of the program can be converted into a formula in disjunctive normal form that represents the agent's whole fallback belief.

## 7 Additional Semantics

We have focused in this paper on the complete semantics. Some of the notions we introduced can be adapted to other semantics in a straightforward way. For example, we can define a family of types of *s-belief* for a semantics $s \in \{Co, St, Pr, Gr\}$ (for Complete, Stable, Preferred, Grounded) as follows:

**Definition 15.** *Let $F = (A, R)$ be an AF, $S = (F, K)$ be the agent's belief state and $s \in \{Co, St, Pr, Gr\}$. We define $Bel_s(S)$ by $[Bel_s(S)] = \{L \in [K] \mid L$ is an $s$-labeling of $F\}$. We say that the agent* $s$-believes *$\phi$ iff $Bel_s(S) \models \phi$.*

It can be checked that we have $Bel_{Gr}(S) \models Bel_{Co}(S)$ and $Bel_{St}(S) \models Bel_{Pr}(S) \models Bel_{Co}(S)$. This follows directly from the fact that grounded labelings are also complete, stable also preferred, and so on. Now consider e.g. the following notion of *s-coherence*:

**Definition 16.** *Let $S$ be a belief state and $s \in \{Co, St, Pr, Gr\}$. We say that $S$ is $s$-coherent iff $Bel_s(S) \not\models \bot$.*

Given the notions of $s$-belief and $s$-coherence we can generalize theorem 1:

**Theorem 4.** *Let $s \in \{Co, St, Pr, Gr\}$ and let $(F, K)$ be an $s$-incoherent belief state where $K$ is conflict-free. There exists an AF update $F^*$ for $F$ such that $(F, K) \otimes F^*$ is $s$-coherent.*

Fallback belief, however, is less straightforward to adapt, as the corresponding rationality orderings would have to combine different criteria, i.e. minimizing/-maximizing in-labeled arguments w.r.t. set-inclusion and minimizing illegally labeled arguments, meaning we have to deal with partial pre-orders.

## 8 Related Work

In this section we give a short overview of related work. We already mentioned the relation of our work with the *enforcing problem* [2]. The authors present a result stating that every conflict-free extension can be enforced (i.e., made accepted under a semantics) with an appropriate AF expansion. In our setting we consider more general types of enforcing, not limited only to acceptance of sets of arguments. Our theorem 4 thus strengthens their possibility result.

Next, different ways to characterize the impact of AF expansions have been studied. This includes minimality w.r.t. the number of added attacks, studied in the context of the enforcing problem [9]. Further criteria were defined in the study of the impact on the outcome of an AF of adding an argument [3]. A limitation in that work is that it considers only additions of a single argument. A slightly different perspective is taken in the work of Liao, Lin and Koons [14], where the impact of adding arguments and attacks plays a role in the efficient recomputation of the extensions of an AF.

The ordering presented in section 5 is related to a preferential model semantics for argumentation [15] and a study of nonmonotonic inference relations to reason with AFs [4]. Also related are *open labelings* [16], which have a purpose similar to ours, i.e., to identify arguments to attack in order to make a labeling consistent with an AF, and an approach where arguments are labeled with formulas expressing instructions on what to attack in order to change the argument's status under the grounded semantics [17]. We should also mention other work in which (parts of) argumentation theory are formalized using logics. They

include models using modal logics [18,19]; translations of the problem of computing extensions to problems in classical logic or ASP [20,12]; and a study of a logical language consisting of attack and defense connectives [21].

Finally, our model is related to the concept of a *constrained AF*, where an AF is combined with a constraint on the status of the arguments [22]. However, these constraints must be consistent with the AF's outcome under the admissible semantics, limiting the types of constraints that can be dealt with. Furthermore, this work does not explore the relation between constraints and AF expansions.

## 9 Conclusion and Future Work

We believe that theories about dynamics in abstract argumentation should address two issues: First, agents may learn or come to desire that an AF must have a certain outcome and second, agents may expand their AF. Our solution centers on the issue of dealing with incoherence after constraining the AF's outcome. Two ways to deal with this are AF expansion and by using fallback belief.

We plan to extend our work in a number of directions. First, our model allows iterated updates only under the assumption that new observations never contradict old ones. In order to allow this we have to look at revising the agent's constraint in the light of conflicting observations. Second, a number of generalizations are possible. For example, we may drop the requirement that observations are conflict-free and we can allow removal of arguments and attacks.

Finally, we plan to investigate connections between the areas of abstract argumentation and belief revision beyond those presented in this paper. We believe that the approach of using a logical labeling language to reason about the outcome of an AF is an essential step towards establishing such connections.

**Acknowledgements.** Richard Booth is supported by the National Research Fund, Luxembourg (DYNGBaT project).

## References

1. Dung, P.M.: 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)
2. Baumann, R., Brewka, G.: Expanding argumentation frameworks: Enforcing and monotonicity results. In: Baroni, P., Cerutti, F., Giacomin, M., Simari, G.R. (eds.) COMMA. Frontiers in Artificial Intelligence and Applications, vol. 216, pp. 75–86. IOS Press (2010)
3. Cayrol, C., de Saint-Cyr, F., Lagasquie-Schiex, M.: Change in abstract argumentation frameworks: Adding an argument. Journal of Artificial Intelligence Research 38(1), 49–84 (2010)
4. Booth, R., Kaci, S., Rienstra, T., van der Torre, L.: Monotonic and non-monotonic inference for abstract argumentation. In: FLAIRS (2013)

5. Caminada, M.: On the issue of reinstatement in argumentation. In: Fisher, M., van der Hoek, W., Konev, B., Lisitsa, A. (eds.) JELIA 2006. LNCS (LNAI), vol. 4160, pp. 111–123. Springer, Heidelberg (2006)
6. Baroni, P., Caminada, M., Giacomin, M.: An introduction to argumentation semantics. Knowledge Eng. Review 26(4), 365–410 (2011)
7. Katsuno, H., Mendelzon, A.O.: Propositional knowledge base revision and minimal change. Artificial Intelligence 52(3), 263–294 (1991)
8. Alchourrón, C.E., Gärdenfors, P., Makinson, D.: On the logic of theory change: Partial meet contraction and revision functions. Journal of symbolic logic, 510–530 (1985)
9. Baumann, R.: What does it take to enforce an argument? Minimal change in abstract argumentation. In: Raedt, L.D., Bessière, C., Dubois, D., Doherty, P., Frasconi, P., Heintz, F., Lucas, P.J.F. (eds.) ECAI. Frontiers in Artificial Intelligence and Applications, vol. 242, pp. 127–132. IOS Press (2012)
10. Toni, F., Sergot, M.: Argumentation and answer set programming. In: Balduccini, M., Son, T.C. (eds.) Gelfond Festschrift. LNCS (LNAI), vol. 6565, pp. 164–180. Springer, Heidelberg (2011)
11. de la Banda, M.G., Pontelli, E. (eds.): ICLP 2008. LNCS, vol. 5366. Springer, Heidelberg (2008)
12. Egly, U., Gaggl, S.A., Woltran, S.: Answer-set programming encodings for argumentation frameworks. Argument and Computation 1(2), 147–177 (2010)
13. Gebser, M., Kaufmann, B., Kaminski, R., Ostrowski, M., Schaub, T., Schneider, M.: Potassco: The potsdam answer set solving collection. AI Communications 24(2), 107–124 (2011)
14. Liao, B.S., Jin, L., Koons, R.C.: Dynamics of argumentation systems: A division-based method. Artif. Intell. 175(11), 1790–1814 (2011)
15. Roos, N.: Preferential model and argumentation semantics. In: Proceedings of the 13th International Workshop on Non-Monotonic Reasoning, NMR 2010 (2010)
16. Gratie, C., Florea, A.M.: Argumentation semantics for agents. In: Cossentino, M., Kaisers, M., Tuyls, K., Weiss, G. (eds.) EUMAS 2011. LNCS, vol. 7541, pp. 129–144. Springer, Heidelberg (2012)
17. Boella, G., Gabbay, D.M., Perotti, A., van der Torre, L., Villata, S.: Conditional labelling for abstract argumentation. In: Modgil, S., Oren, N., Toni, F. (eds.) TAFA 2011. LNCS, vol. 7132, pp. 232–248. Springer, Heidelberg (2012)
18. Grossi, D.: On the logic of argumentation theory. In: van der Hoek, W., Kaminka, G.A., Lespérance, Y., Luck, M., Sen, S. (eds.) AAMAS, pp. 409–416. IFAAMAS (2010)
19. Schwarzentruber, F., Vesic, S., Rienstra, T.: Building an epistemic logic for argumentation. In: del Cerro, L.F., Herzig, A., Mengin, J. (eds.) JELIA 2012. LNCS, vol. 7519, pp. 359–371. Springer, Heidelberg (2012)
20. Besnard, P., Doutre, S.: Checking the acceptability of a set of arguments. In: Delgrande, J.P., Schaub, T. (eds.) NMR, pp. 59–64 (2004)
21. Boella, G., Hulstijn, J., van der Torre, L.W.N.: A logic of abstract argumentation. In: Parsons, S., Maudet, N., Moraitis, P., Rahwan, I. (eds.) ArgMAS 2005. LNCS (LNAI), vol. 4049, pp. 29–41. Springer, Heidelberg (2006)
22. Coste-Marquis, S., Devred, C., Marquis, P.: Constrained argumentation frameworks. In: Doherty, P., Mylopoulos, J., Welty, C.A. (eds.) KR, pp. 112–122. AAAI Press (2006)