Computational Models of Argument
H. Prakken et al. (Eds.)

doi:10.3233/FAIA200507

An Epistemic Interpretation of Abstract Dialectical Argumentation

Jesse HEYNINCK [a,1], and Gabriele KERN-ISBERNER [a]
[a] *Department of Computer Science, TU Dortmund, Germany*

Abstract. Formal argumentation is a well-established and influential knowledge representation formalism that is at the center of recent developments in explainable artificial intelligence. Many extensions to formal argumentation have been proposed, and to cope with the multiplicity of such generalizations, *abstract dialectical frameworks* (in short, ADFs) have been proposed by Brewka and Woltran. This generality comes at a cost, since the semantics underlying ADFs are arguably not as transparent as those of abstract argumentation frameworks. This opacity is witnessed among others by revisions of several of the central semantics for abstract dialectical frameworks. In this paper, we intend to give a clear conceptual foundation of abstract dialectical frameworks by intepreting abstract dialectical frameworks in *epistemic logic*. In particular, we show how interpretations and their refinements can be straightforwardly embedded in epistemic logic as S5-structures that model the interpretation as knowledge. Given such an interpretation, it turns out that all major semantics for ADFs coincide with the possible world structures that are *autoepistemically sound* according to the seminal paper by Moore with respect to the theory expressed by the ADF.

Keywords. Autoepistemic Logic, Computational Argumentation, Abstract Dialectical Frameworks

1. Introduction

Formal argumentation is one of the major approaches to knowledge representation and has been heralded for its potential in explainable artificial intelligence (see e.g. [26]). In the seminal paper [8], *abstract argumentation frameworks* where conceived of as directed graphs where nodes represent arguments and edges between these nodes represent attacks. So-called *argumentation semantics* determine which sets of arguments can be reasonably upheld together given such an argumentation graph. Various authors have remarked that other relations between arguments are worth consideration. For example, in [6], *bipolar argumentation frameworks* are developed, where arguments can support as well as attack each other. The last decades saw a proliferation of such extensions of the original formalism of [8], and it has often proven hard to compare the resulting different dialects of the formal argumentation formalism. To cope with the result-

[1] Corresponding Author: Jesse Heyninck. Email: jesse.heyninck@tu-dortmund.de

ing multiplicity, [5,4] introduced *abstract dialectical argumentation* that aims to unify these different dialects. Just like in [8], *abstract dialectical frameworks* (in short, ADFs) are directed graphs. In contradistinction to abstract argumentation frameworks, however, in ADFs, edges between nodes do not necessarily represent attacks but can encode *any* relationship between arguments. Such a generality is achieved by associating an *acceptance condition* with each argument, which is a boolean formula in terms of the parents of the argument that expresses the conditions under which an argument can be accepted. As such, ADFs are able to capture all of the major extensions of abstract argumentation and offer a general framework for argumentation based inference. This generality arguably results in a loss of transparency of the semantics of ADFs. Such an opacity is witnessed by revisions of several of the central semantics for ADFs. For example, the stable semantics from [5] was revised in [4] because it did not adequately capture the stable model semantics from logic programming. Likewise, the admissible semantics received reformulations in [1] and [22] in view of both reasons of intuitiveness and representational adequacy. Such a lack of transparency is especially worrying given the ambitions of formal argumentation in contributing to explainable AI. Therefore, we make first steps towards a clear conceptual foundations of ADFs by interpreting ADFs in *epistemic logic*. In particular, we show how interpretations can be interpreted as S5-structures for the beliefs in the arguments accepted by the interpretations in question. Under such an interpretation, it turns out that all major semantics for ADFs coincide with the S5-structures that are *autoepistemically sound* according to [21] with respect to the knowledge expressed by the ADF.

Outline of the Paper In Section 2, we give preliminaries on propositional logic (Section 2.1), ADFs (Section 2.2) and epistemic and autoepistemic logic (Section 2.3). In Section 3, we reinterpret interpretations as S5-structures known from epistemic logic, and show that such an interpretation fulfills some basic sanity criteria. In Section 4 we show that such an interpretation can be used to translate ADFs into autoepistemic logic. In Section 5 we make some remarks about translating autoepistemic logic into ADFs. We end the paper by discussing related work (Section 6) and making some concluding remarks (Section 7).

2. Preliminaries

In the following, we briefly recall some general preliminaries on propositional logic as well as technical details on ADFs [4].

2.1. Propositional Logic

For a set At of atoms let $\mathcal{L}(\mathsf{At})$ be the corresponding propositional language constructed using the usual connectives $\wedge$ (*and*), $\vee$ (*or*), $\neg$ (*negation*) and $\rightarrow$ (*material implication*). A (classical) *interpretation* (also called *possible world*) ω for a propositional language $\mathcal{L}(\mathsf{At})$ is a function $\omega : \mathsf{At} \rightarrow \{\top, \bot\}$. Let $\Omega(\mathsf{At})$ denote the set of all interpretations for At. We simply write Ω if the set of atoms is implicitly given. An interpretation ω *satisfies* (or is a *model* of) an atom $a \in \mathsf{At}$,

denoted by $\omega \models a$, if and only if $\omega(a) = \top$. The satisfaction relation $\models$ is extended to formulas as usual. As an abbreviation we sometimes identify an interpretation ω with its *complete conjunction*, i. e., if $a_1, \ldots, a_n \in \mathsf{At}$ are those atoms that are assigned $\top$ by ω and $a_{n+1}, \ldots, a_m \in \mathsf{At}$ are those atoms that are assigned $\bot$ by ω we identify ω by $a_1 \ldots a_n \overline{a_{n+1}} \ldots \overline{a_m}$ (or any permutation of this). For example, the interpretation ω_1 on $\{a, b, c\}$ with $\omega(a) = \omega(c) = \top$ and $\omega(b) = \bot$ is abbreviated by $a\bar{b}c$. For $\Phi \subseteq \mathcal{L}(\mathsf{At})$ we also define $\omega \models \Phi$ if and only if $\omega \models \phi$ for every $\phi \in \Phi$. Define the set of models $\mathsf{Mod}(X) = \{\omega \in \Omega(\mathsf{At}) \mid \omega \models X\}$ for every formula or set of formulas X. A formula or set of formulas X_1 *entails* another formula or set of formulas X_2, denoted by $X_1 \vdash X_2$, if $\mathsf{Mod}(X_1) \subseteq \mathsf{Mod}(X_2)$.

2.2. Abstract Dialectical Frameworks

We briefly recall some technical details on ADFs following loosely the notation from [4]. An ADF D is a tuple $D = (S, L, C)$ where S is a set of *statements*, $L \subseteq S \times S$ is a set of *links*, and $C = \{C_s\}_{s \in S}$ is a set of *acceptance functions*, which are total functions $C_s : 2^{par_D(s)} \to \{\top, \bot\}$ for each $s \in S$ with $par_D(s) = \{s' \in S \mid (s', s) \in L\}$. An acceptance function C_s defines the cases when the statement s can be accepted (truth value $\top$), depending on the acceptance status of its parents in D. By abuse of notation, we will often identify an acceptance function C_s with its equivalent *acceptance condition* which models the acceptable cases as a propositional formula $\phi \in \mathcal{L}(par_D(s))$.

Example 1. *We consider the following ADF* $D_1 = (\{a, b, c\}, L, C)$ *with:*
$L = \{(a, b), (b, a), (a, c), (b, c)\}$ *and:* $C_a = \neg b$, $C_b = \neg a$, $C_c = a \vee b$.
Informally, the acceptance conditions can be read as "a is accepted if b is not accepted", "b is accepted if a is not accepted" and "c is accepted if either a is accepted or b is accepted".

An ADF $D = (S, L, C)$ is interpreted through 3-valued interpretations $v : S \to \{\top, \bot, u\}$, which assign to each statement in S either the value $\top$ (true, accepted), $\bot$ (false, rejected), or u (unknown). A 3-valued interpretation v can be extended to arbitrary propositional formulas over S via Kleene semantics: $v(\neg\phi) = \bot[\top]$ iff $v(\phi) = \top[\bot]$, and $v(\neg\phi) = u$ iff $v(\phi) = u$. $v(\phi \wedge \psi) = \top$ iff $v(\phi) = v(\psi) = \top$, $v(\phi \wedge \psi) = \bot$ iff $v(\phi) = \bot$ or $v(\psi) = \bot$, and $v(\phi \wedge \psi) = u$ otherwise, and similarly for disjunction. $\mathcal{V}$ is the set of all three-valued interpretations.

Then $v \in \mathcal{V}$ is a *model* of D if for all $s \in S$, if $v(s) \neq u$ then $v(s) = v(C_s)$.

We define an order $\leq_i$ over $\{\top, \bot, u\}$ by making u the minimal element: $u <_i \top$ and $u <_i \bot$, and this order is lifted pointwise as follows (given two interpretations v, w over S): $v \leq_i w$ iff $v(s) \leq_i w(s)$ for every $s \in S$.[2] The set of two-valued interpretations extending an interpretation v is defined as $[v]^2 = \{\omega \in \Omega(S) \mid v \leq_i \omega\}$. Given a set of interpretations V, $\sqcap_i V(s) = v(s)$ if for every $v' \in V$, $v'(s) = v(s)$ and $\sqcap_i V(s) = u$ otherwise. $\Gamma_D(v) : S \to \{\top, \bot, u\}$ where $s \mapsto \sqcap_i \{\omega(C_s) \mid \omega \in [v]^2\}$.

Definition 1. *Let* $D = (S, L, C)$ *be an ADF with* $v : S \to \{\top, \bot, u\}$ *an interpretation:*

[2]Notice that, in general, a three-valued interpretation will be denoted with v whereas a two-valued interpretation is denoted with ω.

- v *is* complete for D *iff* $v = \Gamma_D(v)$.
- v *is* preferred for D *iff* v *is a* $\leq_i$*-maximally complete interpretation for* D.
- v *is* grounded for D *iff* v *is a* $\leq_i$*-minimally complete interpretation for* D.

We denote by $\mathsf{Cmp}(D)$, $\mathsf{Prf}(D)$ *respectively* $\mathsf{Grn}(D)$ *the sets of complete, preferred respectively grounded interpretations of* D.

Notice that any complete (and therefore preferred and grounded) interpretation of D is also a model of D. We finally define inference relations for ADFs:

Definition 2. *Given an ADF* $D = (S, L, C)$ *and* $s \in S$ *and* $\mathsf{sem} \in \{\mathsf{Prf}, \mathsf{Cmp}, \mathsf{Cmp}\}$, *we define:* $D \overset{\cap}{\underset{\mathsf{sem}}{\sim}} s[\neg s]$ *iff* $v(s) = \top[\bot]$ *for all* $v \in \mathsf{sem}(D)$.[3]

Example 2 (Example 1 continued)**.** *The ADF of Example 1 has three complete models* v_1, v_2, v_3 *with:* $v_1(a) = \top$, $v_1(b) = \bot$,$v_1(c) = \top$, $v_2(a) = \bot$, $v_2(b) = \top$,$v_2(c) = \top$, $v_3(a) = u$,$v_3(b) = u$,$v_3(c) = u$.

v_3 *is the grounded interpretation whereas* v_1 *and* v_2 *are both preferred.*

2.3. Epistemic and Autoepistemic Logic

We recall the syntax and semantics of S5 [17]. We use $\mathbf{L}$ to denote the epistemic belief operator. By an epistemic language we mean any language $\mathcal{L}^{\mathbf{L}}$ such that $\mathbf{L}\phi \in \mathcal{L}^{\mathbf{L}}$ if $\phi \in \mathcal{L}^{\mathbf{L}}$. We denote $\mathcal{L}$ as the fragment of $\mathcal{L}^{\mathbf{L}}$ that contains all the formulas containing no occurence of the belief operator $\mathbf{L}$ and we shall from now on assume that $\mathcal{L}$ coincides with the language of propositional logic.

Definition 3. *Given* Ω, *a possible world structure over* Ω *is a set* $Q \subseteq \Omega$.

The set of all possible world structures is thus[4] $\wp(\Omega)$ and is a complete lattice under $\subseteq$. Such possible world structures can be used to model beliefs by interpretting a set of worlds Q as the states an agent considers as possible. This is the standard idea underlying the semantics of the modal logic S5 where entailment is defined as follows:

Definition 4. *Let* $Q \cup \{\omega\} \subseteq \Omega$ *and* $\phi \in \mathcal{L}^{\mathbf{L}}$*:*

- *for* $\phi \in \mathsf{At}$, $Q,\omega \models \phi$ *if* $\omega \models \phi$
- $Q,\omega \models \mathbf{L}\phi$ *if* $Q,\omega' \models \phi$ *for every* $\omega' \in Q$
- $Q,\omega \models \phi \wedge \psi$ *if* $Q,\omega \models \phi$ *and* $Q,\omega \models \psi$
- $Q,\omega \models \neg\phi$ *if* $Q,\omega \not\models \phi$

Finally, $Q,\omega \models \phi \rightarrow \psi$ *iff* $Q,\omega \models \neg\phi \vee \psi$ *and* $Q,\omega \models \phi \vee \psi$ *iff* $Q,\omega \models \neg(\neg\phi \wedge \neg\psi)$.

Example 3. *Consider the formula* $\neg\mathbf{L}b \rightarrow a$ *and the possible world structure* $\{a\bar{b}, \bar{a}\bar{b}\}$. *Observe for example that* $\{a\bar{b}, \bar{a}\bar{b}\}, a\bar{b} \models \neg\mathbf{L}b \rightarrow a$ *whereas* $\{a\bar{b}, \bar{a}\bar{b}\}, \bar{a}\bar{b} \not\models \neg\mathbf{L}b \rightarrow a$.

[3] Since the grounded extension is unique for any ADF [4], $\cap$ is ommited from $\overset{}{\underset{\mathsf{Grn}}{\sim}}$.

[4] Notice that we use $\wp$ as the power-set and not as the Weierstrass function.

[21] noticed that it is interesting to look at those possible world structures that represent "knowledge of a perfect, rational, introspective agent" [3]. In more detail, given a set of formulas $\Delta \subseteq \mathcal{L}^{\mathbf{L}}$, Moore suggests to look at those sets of possible worlds that model Δ and are closed under introspection. In terms of possible world structures, this translates to possible world structures that are fixpoints of the following operator (see [3]) (given $Q \subseteq \Omega$ and $\Delta \subseteq \mathcal{L}_{\mathbf{L}}$):

$$\Psi_\Delta(Q) = \{\omega \in \Omega \mid Q, \omega \models \bigwedge \Delta\}$$

Definition 5. *A set of worlds $Q \subseteq \Omega$ is an autoepistemic extension (in short, AEE) for $\Delta \subseteq \mathcal{L}^{\mathbf{L}}$ iff $\Psi_\Delta(Q) = Q$. An AEE Q is* consistent *iff $Q \neq \emptyset$.*

Example 4. *Let $\Delta = \{\neg \mathbf{L}b \rightarrow a; \neg \mathbf{L}a \rightarrow b\}$. We have the following autoepistemic extensions for Δ: $\{ab, \overline{a}b\}$ and $\{ab, a\overline{b}\}$. Notice that e.g. $\{\overline{a}\overline{b}\}$ is not an autoepistemic extension since $\{\overline{a}\overline{b}\}, \overline{a}\overline{b} \models \neg \mathbf{L}b \wedge \neg a$, i.e. $\{\overline{a}\overline{b}\}, \overline{a}\overline{b} \not\models \neg \mathbf{L}b \rightarrow a$. Therefore, $\overline{a}\overline{b} \notin \Psi_\Delta(\{\overline{a}\overline{b}\})$ and thus $\{\overline{a}\overline{b}\}$ does not constitute a fixed point under Ψ_Δ.*

In [21], a syntactic characterization of autoepistemic extensions was given as follows, which we recall for completeness:

Definition 6. *A* (syntactic) autoepistemic extension *of a set of autoepistemic formulas $\Delta \subseteq \mathcal{L}^{\mathbf{L}}$ is any theory $\mathcal{E} \subseteq \mathcal{L}^{\mathbf{L}}$ that satisfies (where $\phi \in \mathcal{L}^{\mathbf{L}}$):*

$$\mathcal{E} = Cn(\Delta \cup \{\mathbf{L}\phi \mid \mathcal{E} \vdash \phi\} \cup \{\neg \mathbf{L}\phi \mid \mathcal{E} \nvdash \phi\})$$

The syntactic characterization of autoepistemic extensions and the one in terms of possible worlds are equivalent (see e.g. [20]):

Theorem 1. *Given $\Delta \subseteq \mathcal{L}^{\mathbf{L}}$, $Q \subseteq \Omega$ is an autoepistemic extension of Δ iff $\{\phi \in \mathcal{L}^{\mathbf{L}} \mid \forall \omega \in Q : Q, \omega \models \phi\}$ is a syntactic autoepistemic extension of Δ.*

Furthermore, it will prove useful below to consider *maximally informative* and *minimally informative* autoepistemic extensions:[5]

Definition 7. *Given $\Delta \subseteq \mathcal{L}^{\mathbf{L}}$:*

- *$Q \subseteq \Omega$ is a* maximally informative AEE *iff it is an autoepistemic extension and there is no autoepistemic extension $Q' \subseteq \Omega$ s.t. $Q' \subset Q$.*
- *$Q \subseteq \Omega$ is a* minimally informative AEE *iff it is an autoepistemic extension and there is no autoepistemic extension $Q' \subseteq \Omega$ s.t. $Q' \supset Q$.*

We can define an inference relation based on autoepistemic logics as follows:

Definition 8. *Given an autoepistemic knowledge base Δ:*

- $\Delta \mathrel{|\!\sim}^{\cap}_{AEL} \phi$ *iff $\phi \in \mathcal{E}$ for every autoepistemic extension $\mathcal{E}$ of Δ.*
- $\Delta \mathrel{|\!\sim}^{\cap,\mathsf{max}}_{AEL} \phi$ *iff $\phi \in \mathcal{E}$ for every maximally informative AEE $\mathcal{E}$ of Δ.*
- $\Delta \mathrel{|\!\sim}^{\cap,\mathsf{min}}_{AEL} \phi$ *iff $\phi \in \mathcal{E}$ for every minimally informative AEE $\mathcal{E}$ of Δ.*

[5] Notice that a maximally informative AEE is $\subseteq$-minimal: this is so because we consider sets of worlds, and thus minimizing these sets means maximizing the informational content of these sets of worlds. Likewise, minimally informative AAEs are $\subseteq$-maximal.

3. An Epistemic Embedding of ADF-Interpretations

In ADFs, instead of restricting relations between arguments to attack or support, arguments can have *any* relation between each other. This abstraction is achieved by assigning acceptance conditions to arguments in terms of their parents. Given an ADF, semantics encode what are reasonable stances for an agent given the information encoded by an ADF in the following sense: a node can only be accepted if we have good reasons for accepting it, and having good reasons to accept a node means that we should accept the node in question. E.g. in Example 1, a can only be accepted if b is rejected, and likewise if b is rejected, a should be accepted. Formally speaking, the semantics of ADFs are based on 3-valued interpretations v over S. $v(s) = \top$ means that s is believed. Likewise, $v(s) = \bot$ encodes belief in s being false, whereas $v(s) = u$ encodes suspension of belief about s, i.e. neither believing s being true nor believing s being false. Epistemic logic allows us to give a straightforward epistemic embedding of a 3-valued interpretation. In more detail, given an ADF $D = (S, L, C)$ and 3-valued interpretation v over S, we can associate a possible world structure with v as follows:

Definition 9. *Let $D = (S, L, C)$ and $v \in \mathcal{V}$. We define $Q_v = \{\omega \in \Omega(S) \mid v \leq_i \omega\}$*

Under this interpretation, Q_v can be seen to be the set of all worlds that are possibilities (given v) for being the *actual world*. For example, if $v(s) = \top$, it will be the case that for every $\omega \in Q_v$, $\omega \models s$, i.e. in every candidate for the actual world, s is the case and consequently Q_v models belief in s. Likewise, if $v(s) = u$, there are candidates for the actual world where s is true and candidates for the actual world where s is false, and thus Q_v models neither belief in s nor belief in $\neg s$. One can observe that $Q_v = [v]^2$, i.e. the semantics of ADFs already implicitly assume possible world structures. The following result shows that $v(s) = \top[\bot]$ indeed corresponds to belief in s by Q_v:

Proposition 1. *For any interpretation $v \in \mathcal{V}$:*

- $v(s) = \top$ *iff* $Q_v, \omega \models \mathbf{L}s$ *(for any $\omega \in \Omega(S)$),*
- $v(s) = \bot$ *iff* $Q_v, \omega \models \mathbf{L}\neg s$ *(for any $\omega \in \Omega(S)$),*
- $v(s) = u$ *iff* $Q_v, \omega \models \neg\mathbf{L}s \wedge \neg\mathbf{L}\neg s$ *(for any $\omega \in \Omega(S)$),*

Proof. Suppose first that $v(s) = \top$. Then for every $\omega \in Q_v$, $\omega \models s$ and thus $Q_v, \omega \models \mathbf{L}s$. Suppose now that $Q_v, \omega \models \mathbf{L}s$, i.e. for every $\omega \in Q_v$, $\omega \models s$ and suppose towards a contradiction that $v(s) \neq \top$. But then there is an $\omega' \in \Omega(S)$ s.t. $v \leq_i \omega'$ and $\omega'(s) = \bot$. Since $\omega' \in Q_v$, this contradicts $Q_v, \omega \models \mathbf{L}s$. The other cases are analogous. □

The epistemic embedding of interpretations also allows for an intuitive analogue of the information ordering $\leq_i$ over $\mathcal{V}$. Recall that this ordering represents the amount of information represented by an interpretation v. Within our epistemic interpretation of $\mathcal{V}$, $v \leq_i v'$ means that the interpretation v' is commited to the same or more beliefs than v, i.e. whenever $Q_v, \omega \models \mathbf{L}\phi$ then $Q_{v'}, \omega \models \mathbf{L}\phi$ (for any $\omega \in \Omega$). This is the case when $Q_v \supseteq Q_{v'}$, i.e. the information $Q_{v'}$ gives us about the actual world is at least as specific as the information about the ac-

tual world given by Q_v. This intuition is vindicated by the following proposition (whose proof is straightforward and left out in view of spatial considerations):

Proposition 2. *$v \leq_i v'$ iff $Q_v \supseteq Q_{v'}$.*

4. Interpreting ADF-semantics in Autoepistemic Logic

In this section, we use the epistemic embedding of three-valued interpretations v over a set of nodes S to translate all of the major semantics for ADFs in autoepistemic logic. We first formulate a translation that is adequate for complete semantics. This translation allows us to show that preferred respectively grounded interpretations correspond to autoepistemic extensions that are maximally respectively minimally informative. In Section 4.2, we finally show that the translation fulfills some desirable properties.

4.1. Translating ADFs into Autoepistemic Logic

The basic idea behind our translation is the following: believing a condition C_s of a node s means that the node must be true, which formally translates as the premise $\mathbf{L}C_s \rightarrow s$. Likewise, believing the condition C_s is false means that the node must be false (i.e. $\mathbf{L}\neg C_s \rightarrow \neg s$). In other words, positive (respectively negative) beliefs in nodes imply truth (respectively falsity) of the corresponding nodes.

Definition 10. *Given an an ADF $D = (S, L, C)$, $\Delta(D) := \{\mathbf{L}C_s \rightarrow s; \mathbf{L}\neg C_s \rightarrow \neg s \mid s \in S\}$*

It will prove useful to have a method to define an interpretation v_Q on the basis of a possible world structure $Q \subseteq \Omega(S)$ as follows: $v_Q := \sqcap_i Q$.

The critical reader might perhaps wonder if the translation does not require the "reversed" conditionals $\mathbf{L}s \rightarrow C_s$ and $\mathbf{L}\neg s \rightarrow \neg C_s$, which encode a form of *explanatory closure* of ADFs which states that for every node that is believed (respectively disbelieved), an agent should be able to give a reason for this belief (respectively disbelief). This is done by adding the premises $\mathbf{L}s \rightarrow C_s$ and $\mathbf{L}\neg s \rightarrow \neg C_S$. In fact, for any $s \in S$ and any AEE of $\Delta(D)$, Q will also imply both of the above implications:[6]

Fact 1. *Given an ADF $D = (S, L, C)$ and an autoepistemic extension Q of $\Delta(D)$, $Q, \omega \models (\mathbf{L}s \rightarrow C_s) \wedge (\mathbf{L}\neg s \rightarrow \neg C_S)$ for any $\omega \in Q$ and any $s \in S$.*

Proof. [7] Consider the ADF $D = (S, L, C)$ and suppose $Q \subseteq \Omega(S)$ is an AEE of $\Delta(D)$. Suppose now that $\omega \in Q$, $s \in S$ and $Q, \omega \models \mathbf{L}s$. Then $v_Q(s) = \top$ and since v_Q is complete (with Theorem 2) and thus also a model, $v_Q(C_s) = \top$ and thus $Q, \omega' \models \mathbf{L}C_s$ for any $\omega' \in \Omega(S)$. This implies that $\omega(C_s) = \top$ for any $\omega \in Q$.

[6] We thank an anonymous reviewer of a previous version of this paper for noticing this.

[7] Notice that the proof of this fact makes use of Theorem 2, which is shown later in this paper. However, since the proof of Theorem 2 does not in any way depend on this fact, this does not cause any logic circularity.

Altogether this shows that for any $s \in S$, $Q, \omega \models \mathbf{L}s \rightarrow C_s$ for any $s \in S$. The proof for $\mathbf{L}\neg s \rightarrow \neg C_s$ is analogous.

□

It is perhaps interesting to note, however, that an alternative translation $\Delta^*(D) = \{\mathbf{L}s \rightarrow C_s, \mathbf{L}\neg s \rightarrow \neg C_S \mid s \in S\}$ is *not* adequate, i.e. there might be AEEs that are not complete:

Example 5. *Let $D = (\{a\}, L, C)$ with $C_a = \top$. The interpretation $v(a) = \top$ is grounded and preferred. Since $\Delta^*(D) = \{\mathbf{L}a \rightarrow \top, \mathbf{L}\neg a \rightarrow \bot\}$, there are two AEEs of $\Delta^*(D)$: $\{a, \overline{a}\}$ and $\{a\}$. To see that $\{a, \overline{a}\}$ is an AEE, notice that (for any $\omega \in \Omega(\{a\})$) $\{a, \overline{a}\}, \omega \models \neg\mathbf{L}a \wedge \neg\mathbf{L}\neg a$ and thus $\Delta^*(D)$ is satisfied trivially.*

We are now ready to prove the main adequacy results. We first need an intermediate result whose proof is left out in view of spatial considerations:

Lemma 1. *If Q is an AEE of $\Delta(D)$ then $[v_Q]^2 = Q$.*

Theorem 2. *Given an ADF $D = (S, L, C)$, the following statements hold:*

1. *If $Q \subseteq \Omega(S)$ is a consistent autoepistemic extension of $\Delta(D)$ then v_Q is a complete interpretation of D;*
2. *If v is a complete interpretation of D then Q_v is an autoepistemic extension of $\Delta(D)$.*

Proof. Ad 1: Suppose that Q is a consistent autoepistemic extension of $\Delta(D)$. We show that for any $s \in S$, $\Gamma_D(v_Q)(s) = v_Q(s)$. We show the case for $v_Q(s) = u$, the other cases are similar and left out in view of space restrictions.

Suppose indeed that $v_Q(s) = u$, i.e. there are some $\omega, \omega' \in Q$ s.t. $\omega(s) = \top$ and $\omega'(s) = \bot$. Since $\omega \in Q$ and Q is an AEE of $\Delta(D)$ and $\mathbf{L}\neg C_s \rightarrow \neg s \in \Delta(D)$, $Q, \omega \models s \rightarrow \neg\mathbf{L}\neg C_s$. Likewise (since $\mathbf{L}C_s \rightarrow s \in \Delta(D)$), $Q, \omega' \models \neg s \rightarrow \neg\mathbf{L}C_s$. This implies that $Q, \omega \models \neg\mathbf{L}C_s$ and $Q, \omega' \models \neg\mathbf{L}\neg C_s$. This implies that there are some $\omega'', \omega''' \in Q$ s.t. $Q, \omega'' \models C_s$ and $Q, \omega''' \models \neg C_s$. Since $Q = [v_Q]^2$ by Lemma 1, this means $\Gamma_D(v_q)(s) = u$.

Thus we have established that $v_Q(s) = x$ implies $\Gamma_D(v_Q)(s) = x$ for every $x \in \{\top, \bot, u\}$. The cases for $\Gamma_D(v_Q)(s) = x$ follow with contraposition from this and since $\{\top, \bot, u\}$ exhausts all possible values of $\Gamma_D(v_Q)$.

The proof of 2. is left out in view of spatial considerations.

□

From this the following corollary follows for the complete semantics:

Corollary 1. *Given ADF $D = (S, L, C)$ and $s \in S$: $D \mathrel{|\!\sim}^{\cap}_{\mathsf{Cmp}} s[\neg s]$ iff $\Delta(D) \mathrel{|\!\sim}^{\cap}_{AEL} s[\neg s]$.*

We now turn to grounded and preferred semantics. We first need the following Lemma:

Lemma 2.
1. *Given some $v \in \mathcal{V}$, $v = v_{Q_v}$.*
2. *Given an ADF $D = (S, L, C)$, if $Q \subseteq \Omega(S)$ is an AEE of $\Delta(D)$, then also $Q = Q_{v_Q}$.*

Proof. We sketch the proof of 2., 1 is analogous but simpler. Suppose for this $D = (S, L, C)$ and $Q \subseteq \Omega(S)$ is an AEE of $\Delta(D)$. Clearly $Q_{v_Q} \supseteq Q$. Suppose now towards a contradiction there is an $\omega \in Q_{v_Q} \setminus Q$. For any $\omega \in Q_{v_Q}$, $\omega \models s[\neg s]$ iff $v_Q(s) = \top[\bot]$, i.e. $\omega \models s[\neg s]$ iff $Q, \omega' \models \mathbf{L}s[\mathbf{L}\neg s]$ for any $\omega' \in \Omega(S)$. Thus, for every $s \in S$ s.t. $\omega(s) \neq v_Q(s)$, $v_Q(s) = u$, i.e. $Q, \omega' \models \neg\mathbf{L}s \wedge \neg\mathbf{L}\neg s$ and thus there are some $\omega', \omega'' \in Q$ s.t. $\omega'(s) = \omega(s)$ and $\omega''(s) \neq \omega(s)$. Furthermore, since Q is an AEE of $\Delta(D)$ and $\mathbf{L}C_s \rightarrow s \in \Delta(D)$ and $\mathbf{L}\neg C_s \rightarrow \neg s \in \Delta(D)$, $Q, \omega' \models \neg\mathbf{L}C_s \wedge \neg\mathbf{L}\neg C_s$ for any $\omega' \in Q$.

But then $Q, \omega'' \not\models \mathbf{L}C_s \rightarrow s$, contradiction to Q being an AEE of $\Delta(D)$ and $\omega'' \in Q$). But then $Q, \omega \models (\mathbf{L}C_s \rightarrow s) \wedge (\mathbf{L}\neg C_s \rightarrow \neg s)$. Altogether, we have established that: if $v_Q(s) = u$ then $Q, \omega \models (\mathbf{L}C_s \rightarrow s) \wedge (\mathbf{L}\neg C_s \rightarrow \neg s)$. We can easily show the same for any $s \in S$ s.t. $v_Q(s) \in \{\top, \bot\}$, which implies $Q, \omega \models \Delta(D)$ and thus $\omega \in Q$, contradiction to the supposition.

□

We notice that Lemma 2 does not in general hold for sets of possible worlds. To see this, consider the set $Q = \{a\bar{b}, \bar{a}, b\}$. Then $v_Q(a) = v_Q(b) = u$ and $Q_{v_Q} = \{ab, a\bar{b}, \bar{a}b, \bar{a}\bar{b}\}$. The proofs of the following Theorems are straightforward in view of Theorem 2, Proposition 2 and Lemma 2 and left out in view of spatial restrictions.

Theorem 3. *Given an ADF $D = (S, L, C)$, the following statements hold:*

1. *If $Q \subseteq \Omega(S)$ is a minimally informative AEE of $\Delta(D)$ then v_Q is the grounded interpretation of D;*
2. *If v is the grounded interpretation of D then Q_v is a minimaly informative AEE of $\Delta(D)$.*

Theorem 4. *Given an ADF $D = (S, L, C)$, the following statements hold:*

1. *If $Q \subseteq \Omega(S)$ is a maximally informative AEE of $\Delta(D)$ then v_Q is a preferred interpretation of D;*
2. *If v is a preferred interpretation of D then Q_v is a maximally informative AEE of $\Delta(D)$.*

From these theorems the following corollary follows for the grounded and preferred semantics:

Corollary 2. *For any ADF $D = (S, L, C)$ and $s \in S$, the following statements hold:*

- $D \mathrel{|\!\sim}^{\cap}_{\mathsf{Prf}} s[\neg s]$ *iff* $\Delta(D) \mathrel{|\!\sim}^{\cap,\mathsf{max}}_{AEL} s[\neg s]$.
- $D \mathrel{|\!\sim}_{\mathsf{Grn}} s[\neg s]$ *iff* $\Delta(D) \mathrel{|\!\sim}^{\cap,\mathsf{min}}_{AEL} s[\neg s]$.

4.2. Properties of the Translation

In [9], several desirable properties for translations between non-monotonic formalisms where suggested: *faithfulness*, *polynomiality* and *modularity*. A faithful translation is a translation that preserves adequacy between the autoepistemic extensions and the semantics of ADFs. The faithfulness of our translation is shown in Theorem 2 for complete semantics, Theorem 3 for grounded semantics and Theorem 4 for preferred semantics.

Polynomiality is motivated by the requirement that the translation should be calculable within reasonable bounds. Clearly, the translation is polynomial: in fact it is linear in the number of nodes.

Modularity was originally defined for translations between circumscription and default logic [12]. Even though the original formulation was slightly different, we follow [24] in his formulation of modularity of a translation from ADFs to a target formalism. Basically, a translation is modular if "local" changes in the translated ADF will only lead to "local" changes in the translation. More formally, for two ADFs $D_1 = (S_1, L_1, C_1)$ and $D_2 = (S_2, L_2, C_2)$, such that $S_1 \cap S_2 = \emptyset$, a translation Δ is modular iff $\Delta(D_1 \cup D_2) = \Delta(D_1) \cup \Delta(D_2)$. It is easy to observe that the translation presented in this paper is modular.[8]

5. From autoepistemic logic to ADFs

The reader might wonder if it is possible to translate autoepistemic logic into ADFs. Such a translation is indeed possible, for the following reason: in [13] a translation from autoepistemic logic to *strong autoepistemic logic* was shown. In the same paper, it was also shown that strong autoepistemic logic can be translated into Reiter's default logic [23]. In [8] Reiter's default logic was translated into abstract argumentation, which can be captured in ADFs. It thus follows that ADFs admit autoepistemic logic under a composition of translations. A direct translation, however, remains to be investigated. We leave this as an avenue for further research.

6. Related Work

The main contribution of this paper is an embedding of ADFs in epistemic logics and a translation from ADFs into autoepistemic logic based on such an embedding. To the best of our knowledge, this is the first time that such an interpretation or translation is spelled out in the literature. There are, however, some related approaches that we wish to mention.

In [10] modal logic is applied to formalize fragments of formal argumentation theory. In particular, [10] establishes a correspondence between a given argumentation framework and a modal logic frame. The idea is that the argumentation framework and the modal frame will have the same number of nodes: for every argument there will be exactly one corresponding world. The meaning of the accessibility relation is, in a sense, inversed: if a attacks b then the world corresponding to b will be an accessible from the world corresponding to a. Consequently, even though both [10] and we interpret argumentation formalisms in some modal logic, the differences should be clear: we consider a translation into epistemic logic instead of a modal logic based on a frame structurally similar to

[8] [24] remarks that it would make sense from a conceptual point of view to generalize modularity to ADFs that have nodes that are not necessarily disjoint, but remarks that technically it is difficult to formulate such a generalized criterion of modularity. We follow [24] in leaving the formulation of such a criterion for future work.

the argumentation framework and we consider the more general ADFs instead of abstract argumentation frameworks.

The connections between ADFs and other formalisms for non-monotonic reasoning have been investigated before. [24] shows that there is a translation from ADFs into normal logic programs. In that paper, it is remarked that in view of the translation from ADFs into normal logic programs, and existing translations from normal logic programs into default logic and from default logic into autoepistemic logic (both by [7]), there exists a translation from ADFs into autoepistemic logic. We now give such a translation and argue for its conceptual adequacy.

Finally, we mention [15,11] where the correspondence between logics for non-monotonic conditionals are investigated. The results of that paper are that a subset of the complete models, namely the 2-valued models (interpretations $v \in \Omega(S)$ s.t. $v(s) = v(C_s)$ for every node s) can be straightforwardly modelled in conditional logics but for complete semantics, such a translation is less straightforward. The translations in this paper together with results on the relation between conditional logics and epistemic logic (e.g. [16]) can be used to shed further light on the correspondence between conditional logics and ADFs.

7. Conclusion and Outlook

In this paper, we have given an epistemic interpretation of ADFs and have formulated an intuitive, faithful, polynomial and modular translation from ADFs into autoepistemic logic. Not only is this interesting from a conceptual point of view, but this translation also is a starting point for further investigations into the connection between ADFs and other formalisms, since there are studies on the relationship between autoepistemic logic and other formalisms, such as default logic [9,7], logic programming [19,18] and circumscription [14]. Furthermore, the epistemic interpretation undertaken in this paper allows us to apply techniques developed in epistemic logic to ADFs. For example, *dynamic epistemic logic* [25] is a well-established field that uses epistemic logic to model changes in knowledge. The epistemic interpretation of ADFs in this paper can take advantage of developments in dynamic epistemic logic (such as [2,25]) to shed further light, among others, on argumentation dynamics (a topic that has been studied mainly for abstract argumentation frameworks until now) and argumentation in multi-agent interactions. In future work, we want to translate other semantics into autoepistemic logic, such as the different formulations of the stable semantics [5,4] and look at extensions of ADFs such as prioritized ADFs [4].

Acknowledgements The research reported in this paper was supported by the German National Science Foundation, DFG-project KE-1413/11-1.

References

[1] João Alcântara and Samy Sá. On three-valued acceptance conditions of abstract dialectical frameworks. *Electronic Notes in Theoretical Computer Science*, 344:3–23, 2019.

[2] Alexandru Baltag, Lawrence S Moss, and Sławomir Solecki. The logic of public announcements, common knowledge, and private suspicions. In *Readings in Formal Epistemology*, pages 773–812. Springer, 2016.

[3] Bart Bogaerts. *Groundedness in logics with a fixpoint semantics*. PhD thesis, 2015.
[4] Gerhard Brewka, Hannes Strass, Stefan Ellmauthaler, Johannes Peter Wallner, and Stefan Woltran. Abstract dialectical frameworks revisited. In *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.
[5] Gerhard Brewka and Stefan Woltran. Abstract dialectical frameworks. In *KR 12*, 2010.
[6] Claudette Cayrol and Marie-Christine Lagasquie-Schiex. On the acceptability of arguments in bipolar argumentation frameworks. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, pages 378–389. Springer, 2005.
[7] Marc Denecker, Victor Marek, and Mirosław Truszczyński. Approximations, stable operators, well-founded fixpoints and applications in nonmonotonic reasoning. In *Logic-based artificial intelligence*, pages 127–144. Springer, 2000.
[8] Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77:321–358, 1995.
[9] Georg Gottlob. The power of beliefs or translating default logic into standard autoepistemic logic. In *Foundations of Knowledge Representation and Reasoning*, pages 133–144. Springer, 1994.
[10] Davide Grossi. On the logic of argumentation theory. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: Volume 1*, pages 409–416. International Foundation for Autonomous Agents and Multiagent Systems, 2010.
[11] Jesse Heyninck, Gabriele Kern-Isberner, and Matthias Thimm. On the correspondence between abstract dialectical frameworks and non-monotonic conditional logics. In *33rd International FLAIRS Conference*, 2020.
[12] Tomasz Imielinski. Results on translating defaults to circumscription. *Artificial Intelligence*, 32(1):131–146, 1987.
[13] Tomi Janhunen. Representing autoepistemic introspection in terms of default rules. In *Proceedings of the 12th European Conference on Artificial Intelligence, ECAI'96*, pages 70–74. John Wiley and Sons, 1996.
[14] Tomi Janhunen. On the intertranslatability of autoepistemic, default and priority logics, and parallel circumscription. In *European Workshop on Logics in Artificial Intelligence*, pages 216–232. Springer, 1998.
[15] Gabriele Kern-Isberner and Matthias Thimm. Towards conditional logic semantics for abstract dialectical frameworks. In Carlos I. Chesnevar et al., editor, *Argumentation-based Proofs of Endearment*, volume 37 of *Tributes*. College Publications, November 2018.
[16] Costas D Koutras, Christos Moyzes, and Christos Rantsoudis. A reconstruction of default conditionals within epistemic logic. *Fundamenta Informaticae*, 166(2):167–197, 2019.
[17] Clarence Irving Lewis, Cooper Harold Langford, and P Lamprecht. *Symbolic logic*. Dover Publications New York, 1959.
[18] Vladimir Lifschitz and Grigori Schwarz. Extended logic programs as autoepistemic theories. In *LPNMR*, pages 101–114, 1993.
[19] V Wiktor Marek and Miroslaw Truszczynski. Reflexive autoepistemic logic and logic programming. *2nd Int. Ws. on LP & NMR*, pages 115–131, 1993.
[20] R Moore. Possible-world semantics for autoepistemic logic. In *Readings in nonmonotonic reasoning*, pages 137–142. Morgan Kaufmann Publishers Inc., 1987.
[21] Robert C Moore. Semantical considerations on nonmonotonic logic. *Artificial intelligence*, 25(1):75–94, 1985.
[22] Sylwia Polberg, Johannes Peter Wallner, and Stefan Woltran. Admissibility in the abstract dialectical framework. In *International Workshop on Computational Logic in Multi-Agent Systems*, pages 102–118. Springer, 2013.
[23] Raymond Reiter. A logic for default reasoning. *Artificial intelligence*, 13:81–132, 1980.
[24] Hannes Strass. Approximating operators and semantics for abstract dialectical frameworks. *Artificial Intelligence*, 205:39–70, 2013.
[25] Hans Van Ditmarsch, Wiebe van Der Hoek, and Barteld Kooi. *Dynamic epistemic logic*, volume 337. Springer Science & Business Media, 2007.
[26] Zhiwei Zeng, Chunyan Miao, Cyril Leung, and Jing Jih Chin. Building more explainable artificial intelligence with argumentation. In *AAAI*, volume 33, pages 8044–8045, 2018.