ASPIC-END: Structured Argumentation with Explanations and Natural Deduction

Jérémie Dauphin and Marcos Cramer

University of Luxembourg**

Abstract. We propose ASPIC-END, an adaptation of the structured argumentation framework ASPIC+ which can incorporate explanations and natural deduction style arguments. We discuss an instantiation of ASPIC-END that models argumentation about explanations of semantic paradoxes (e.g. the Liar paradox), and we show that ASPIC-END satisfies rationality postulates akin to those satisfied by ASPIC+.

1 Introduction

In order to develop tools that intelligently support scientists in their interpretation of data and evaluation of theories, it is important to develop formal models of the argumentation and reasoning about conflicting information found in many academic disciplines. One promising methodology for approaching this problem is *structured argumentation theory* [4], which allows for a fine-grained model of argumentation and argumentative reasoning based on a logical language and evaluated according to the principles developed in *abstract argumentation theory*.

One of the dominant formal frameworks for structured argumentation is the *ASPIC+ framework* [12]. In ASPIC+, arguments are built from axioms and premises as well as from strict and defeasible rules, in a similar manner as proofs are built from axioms and rules in a Hilbert-style proof system. Three kinds of attacks between arguments, *undermines*, *undercuts* and *rebuttals*, are defined between arguments, and finally an *argumentation semantics* from Dung-style abstract argumentation theory [8,1] is applied to determine which sets of arguments can be rationally accepted.

Scientific discourse is characterized not only by the exchange of arguments in favour and against various scientific hypotheses, but also by the attempt to scientifically *explain* observed phenomena. In the context of abstract argumentation, Šešelja and Straßer [16] have therefore proposed to incorporate the notion of *explanation* into argumentation theory, in order to model scientific debate more faithfully. So far, this incorporation of explanation into argumentation theory has not been extended to the case of structured argumentation. One goal of the current paper is to work towards filling this gap by presenting on the one hand a

** Jérémie Dauphin has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 690974 for the project "MIREL: MIning and REasoning with Legal texts".

general framework for incorporating explanation into structured argumentation, and on the other hand a particular proposal for how to define explanations in instantiations of that framework within a specific domain.

Scientific arguments often involve hypothetical reasoning, which involves reasoning based on an assumption or hypothesis that is locally assumed to be true for the sake of the argument, but to which there is no commitment on the global level. Such hypothetical reasoning is captured well by natural deduction proof systems, whereas the Hilbert-style definition of arguments in ASPIC+ cannot account for such hypothetical reasoning.

We propose an adaptation of the ASPIC+ framework called *ASPIC-END* that allows for incorporating explanations and hypothetical reasoning. In order to illustrate the usage of ASPIC-END, we consider its application to argumentation about explanations of semantic paradoxes, a research topic within the field of philosophical logic, and present a specific instantiation of the framework that models a simple example from this domain.

In order to ensure that the ASPIC-END framework behaves as one would rationally expect, we have proved multiple rationality postulates about ASPIC-END, as was previously done for ASPIC+ [11].

The paper is structured as follows: In Section 2, we discuss related work and motivate ASPIC-END. In Section 3, we formally define the ASPIC-END framework, and in Section 4, we instantiate it for argumentation about explanation of semantic paradoxes. In Section 5, we present, motivate and prove six rationality postulates for ASPIC-END, and in Section 6 we conclude.

2 Related work & motivation for ASPIC-END

The work of Dung [8] introduced the theory of *abstract argumentation*, in which one models arguments by abstracting away from their internal structure to focus on the relations of conflict between them. In *structured argumentation*, one models also the internal structure of arguments through a formal language in which arguments and counterarguments can be constructed [4]. One important family of frameworks for structured argumentation is the family of ASPIC-like frameworks, consisting among others of the original ASPIC framework [13], the ASPIC+ framework [12], and the ASPIC- framework [7]. We briefly sketch ASPIC+, as it is the basis for our framework ASPIC-END.

In ASPIC+, one starts with a knowledge base and a set of rules which allow one to make inferences from given knowledge. There are two kinds of rules: *Strict rules* logically entail their conclusion, whereas *defeasible rules* only create a presumption in favour of their conclusion. Arguments are built either by introducing an element of the knowledge base into the framework, or by making an inference based on a rule and the conclusions of previous arguments. Attacks between arguments are constructed either by attacking a fallible premise of an argument (*undermining*), by attacking the conclusion of a defeasible inference made within an argument (*rebuttal*), or by questioning the applicability of such a rule (*undercutting*). Preferences between arguments can be derived from preferences between rules. An abstract argumentation framework has then been built and acceptable arguments can be selected using any abstract argumentation semantics.

Caminada and Amgoud [6] have introduced the notion of *rationality postulates* for structured argumentation frameworks. These are conditions that structured argumentation frameworks would rationally be expected to satisfy, such as closure under strict rules of the output and consistency of the conclusions given consistency of the strict rules. Caminada and Amgoud [6] showed that the original ASPIC system did not satisfy these postulates, but proposed minor changes that made it satisfy them. These changes have been incorporated into ASPIC+ [11].

ASPIC-END features three main differences from ASPIC+. The first is that it allows for arguments to introduce an assumption on which to reason hypothetically, just like in natural deduction. In natural deduction, hypothetical derivations are employed in the inference schemes called $\neg$-Introduction (or *proof by contradiction*), $\rightarrow$-Introduction, and $\vee$-Elimination (or *reasoning by cases*). Allowing for the usage of defeasible rules within hypothetical reasoning leads to specific problems that have been studied for the inference scheme of reasoning by cases in a recent paper by Beirlaen, Heyninck and Straßer [3]. In the current paper we avoid these problems by not allowing defeasible rules within hypothetical reasoning. However, a conclusion made on the basis of an inference scheme involving hypothetical reasoning may still be incorporated into an argument that uses defeasible rules, so that there is some integration of defeasible and hypothetical reasoning. In order to keep the presentation simple, our formal definition of ASPIC-END will only cover the case of the inference scheme of proof by contradiction, but reasoning by cases and $\rightarrow$-Introduction can be treated analogously. Our proof-by-contradiction arguments bear a vague similarity to Caminada's *S-arguments* [5], which can attack an argument by showing that its conclusion leads to an absurdity. But unlike S-arguments, proof-by-contradiction arguments can be embedded into more complex arguments which make use of the negated conclusion of the proof-by-contradiction argument to conclude something else.

The second difference is that ASPIC-END has a notion of *explanations* additionally to the notion of arguments. This feature is based on the work of Šešelja and Straßer [16], who have extended Dung-style abstract argumentation with *explananda* (phenomena that need to be explained) and an *explanatory relation*, which allows arguments to either explain these explananda or deepen another argument's explanation. In Section 3, we will need some definitions from [16]:

Definition 1. An explanatory argumentation framework (EAF) is a tuple $\langle \mathcal{A}, \mathcal{X}, \rightarrow, \dashrightarrow \rangle$, where $\mathcal{A}$ is a set of arguments, $\mathcal{X}$ is a set of explananda, $\rightarrow$ is an attack relation between arguments and $\dashrightarrow$ is an explanatory relation from arguments to either explananda or arguments.

Sets of admissible arguments are then selected:

Definition 2. Let $\Delta = \langle \mathcal{A}, \mathcal{X}, \rightarrow, \dashrightarrow \rangle$ be an EAF, $A \in \mathcal{A}$ and $S \subseteq \mathcal{A}$. We say that S is *conflict-free* iff there are no arguments $B, C \in S$ such that $B \rightarrow C$. We say that S *defends* A iff for every $B \in \mathcal{A}$ such that $B \rightarrow A$, there exists $C \in S$ such that $C \rightarrow B$. We say that S is *admissible* iff S is conflict-free and for all $B \in S$, S defends B.

The most suitable admissible sets are then selected by also taking into account their explanatory power and depth. These are measured by first identifying the explanations present in each set of arguments.

Definition 3. Let $\Delta = \langle \mathcal{A}, \mathcal{X}, \rightarrow, \dashrightarrow \rangle$ be an EAF, $S \subseteq \mathcal{A}$ and $E \in \mathcal{X}$. An *explanation* $X[E]$ for E offered by S is a set $S' \subseteq S$ such that there exists a unique argument $A \in S'$ such that $A \dashrightarrow E$ and for all $A' \in S' \setminus \{A\}$, there exists a path in $\dashrightarrow$ from A' to A.

In order to be able to compare sets of arguments on how many explananda they can explain and in how much detail, the two following measures are required:

Definition 4. Let $\Delta = \langle \mathcal{A}, \mathcal{X}, \rightarrow, \dashrightarrow \rangle$ be an EAF and $S, S' \subseteq \mathcal{A}$. Let $\mathcal{E}$ be the set of explananda S offers an explanation for and $\mathcal{E}'$ the set of explananda S' offers an explanation for. We say that S is *explanatory more powerful than* S' ($S >_p S'$) if and only if $E \supsetneq E'$.

Definition 5. Let $\Delta = \langle \mathcal{A}, \mathcal{X}, \rightarrow, \dashrightarrow \rangle$ be an EAF and $S, S' \subseteq \mathcal{A}$. We say that S is *explanatory deeper than* S' ($S >_d S'$) if and only if for each explanation X' offered by S', there is an explanation X offered by S such that $X' \subseteq X$ and for at least one such X and X' pair, $X' \subsetneq X$.

Šešelja and Straßer [16] define two procedures for selecting the most suitable sets of arguments. The first procedure (for the *argumentative core*) consists in selecting the most explanatory powerful conflict-free sets, from which the maximal most defended sets are then retained. The second procedure (for the *explanatory core*) selects the most explanatory powerful conflict-free sets, from which the most defended sets are taken, and then from those selects the minimal explanatory deepest sets. In our formalism, we will slightly alter and reformulate these procedures.

The third difference is that ASPIC-END allows for arguments about the correct rules of logical reasoning. In ASPIC+, such arguments cannot be modeled, as the rules of logical reasoning represented by strict rules, and arguments involving only strict rules can never be attacked. Argumentation about the correct rules of logical reasoning is quite common within the field of philosophical logic, and additionally occurs not only in other areas of philosophy, e.g. in philosophy of science, but also in the study of logic within fields other than philosophy, e.g. in relation to the applications of logic to linguistics, law and Artificial Intelligence. For example, our *prima facie* intuitions suggest that it is a law of logic that a sentence that is not true must be false. However, the Kripke-Feferman solution to the Liar paradox [15,9] suggests that some sentences, such as the Liar sentence, are neither true nor false, since giving them either one of the two truth values leads to a contradiction. This solution is not putting forward an argument against the falsehood of the sentence by rebutting it, nor is it undermining any of the argument's premises. It is undercutting the argument by attacking the inference made from the negation of truth to falsehood.

It is true that outside the academic disciplines of philosophy and logic, argumentation about the correct rules of logical reasoning is very rare. But the goal of structured argumentation frameworks like ASPIC+ and ASPIC-END is to be largely domain-independent, and to therefore incorporate domain-specific assumptions into instantiations of the framework rather than into the framework itself. Given that there are some domains in which arguments about the correct rules of logical reasoning are sometimes put forward, the restriction that disallows such arguments to be modeled in ASPIC+ should be moved from the definition of the framework to the definition of those instantiations of the framework in which such arguments should indeed be disallowed.

To allow such arguments about the correct laws of logic to be modeled in ASPIC-END, we replace strict rules by *intuitively strict rules* whose applicability can be questioned, as in the case of defeasible rules in ASPIC+, but which behave like strict rules when their applicability is accepted. This means that conclusions of intuitively strict rules cannot be rebutted, just as for strict rules in ASPIC+. Intuitively strict rules represent *prima facie laws of logic*, i.e. purportedly logical inference rules which make sense at first but are open to debate.

3 ASPIC-END

In this section, we define ASPIC-END and motivate the details of its definition.

Definition 6. An *argumentation theory* is a tuple $(\mathcal{L}, \mathcal{R}, n, \leq)$, where:

- $\mathcal{L}$ is a logical language closed under the two unary connectives negation ($\neg$) and assumability ($Assumable$) such that $\bot \in \mathcal{L}$.
- $\mathcal{R} = \mathcal{R}_{is} \cup \mathcal{R}_d$ is a set of intuitively strict ($\mathcal{R}_{is}$) and defeasible ($\mathcal{R}_d$) rules of the form $\varphi_1, \ldots, \varphi_n \rightsquigarrow \varphi$ and $\varphi_1, \ldots, \varphi_n \Rightarrow \varphi$ respectively, where $n \geq 0$ and φ_i, $\varphi \in \mathcal{L}$.
- $n : \mathcal{R} \rightarrow \mathcal{L}$ is a partial function.
- $\mathcal{R}_{ce} := \{(\bot \rightsquigarrow \alpha) \mid \alpha \in \mathcal{L}\} \subseteq \mathcal{R}_{is}$, $\forall r' \in \mathcal{R}_{is} \setminus \mathcal{R}_{ce}$, $n(r') \in \mathcal{L}$, and $\forall r \in \mathcal{R}_{ce}$, $n(r)$ is undefined.
- $\leq$ is a reflexive and transitive relation over R_d which represents preference, with $a < b$ iff $a \leq b$ and $b \nleq a$.

Note that we interpret $\bot$ not just as any contradiction but as the conjunction of all formulas in the language.

We now inductively define how to construct arguments. At the same time, we define five functions on arguments that specify certain features of any given argument: $\mathsf{Conc}(A)$ denotes the conclusion of argument A. $\mathsf{As}(A)$ denotes the set of assumptions under which argument A is operating (so whenever $\mathsf{As}(A) \neq \emptyset$, A is a hypothetical argument). $\mathsf{Sub}(A)$ denotes the set of sub-arguments of A. $\mathsf{DefRules}(A)$ denotes the set of all defeasible rules used in A. $\mathsf{TopRule}(A)$ denotes the last inference rule which has been used in the argument if such a rule exists, and is undefined otherwise.

Definition 7. An *argument* A on the basis of an argumentation theory $\Sigma = (\mathcal{L}, \mathcal{R}, n, \leq)$ has one of the following forms:

1. $A_1, \ldots, A_n \rightsquigarrow \psi$, where $A_1, \ldots, A_n$ are arguments such that there exists an intuitively strict rule $\mathsf{Conc}(A_1), \ldots, \mathsf{Conc}(A_n) \rightsquigarrow \psi$ in $\mathcal{R}_{is}$.
   $\mathsf{Conc}(A) := \psi$, $\mathsf{As}(A) := \mathsf{As}(A_1) \cup \cdots \cup \mathsf{As}(A_n)$,
   $\mathsf{Sub}(A) := \mathsf{Sub}(A_1) \cup \cdots \cup \mathsf{Sub}(A_n) \cup \{A\}$,
   $\mathsf{DefRules}(A) := \mathsf{DefRules}(A_1) \cup \cdots \cup \mathsf{DefRules}(A_n)$,
   $\mathsf{TopRule}(A) := \mathsf{Conc}(A_1), \ldots, \mathsf{Conc}(A_n) \rightsquigarrow \psi$.
2. $A_1, \ldots, A_n \Rightarrow \psi$, where $A_1, \ldots, A_n$ are arguments s.t. $\mathsf{As}(A_1) \cup \cdots \cup \mathsf{As}(A_n) = \emptyset$ and there exists a defeasible rule $\mathsf{Conc}(A_1), \ldots, \mathsf{Conc}(A_n) \Rightarrow \psi$ in $\mathcal{R}_d$.
   $\mathsf{Conc}(A) := \psi$, $\mathsf{As}(A) := \emptyset$,
   $\mathsf{Sub}(A) := \mathsf{Sub}(A_1) \cup \cdots \cup \mathsf{Sub}(A_n) \cup \{A\}$,
   $\mathsf{DefRules}(A) := \mathsf{DefRules}(A_1) \cup \cdots \cup \mathsf{DefRules}(A_n) \cup$
   $\{\mathsf{Conc}(A_1), \ldots, \mathsf{Conc}(A_n) \Rightarrow \psi\}$,
   $\mathsf{TopRule}(A) := \mathsf{Conc}(A_1), \ldots, \mathsf{Conc}(A_n) \Rightarrow \psi$.
3. $\mathsf{Assume}(\varphi)$, where $\varphi \in \mathcal{L}$. $\mathsf{Conc}(A) := \varphi$, $\mathsf{As}(A) := \{\varphi\}$, $\mathsf{Sub}(A) := \{\mathsf{Assume}(\varphi)\}$,
   $\mathsf{DefRules}(A) := \emptyset$, $\mathsf{TopRule}(A)$ is undefined.
4. $\mathsf{ProofByContrad}(\neg\varphi, A')$, where A' is an argument such that $\varphi \in \mathsf{As}(A')$ and $\mathsf{Conc}(A') = \bot$, with:
   $\mathsf{Conc}(A) = \neg\varphi$, $\mathsf{As}(A) = \mathsf{As}(A') \setminus \{\varphi\}$,
   $\mathsf{Sub}(A) = \mathsf{Sub}(A') \cup \{\mathsf{ProofByContrad}(\neg\varphi, A')\}$,
   $\mathsf{DefRules}(A) = \mathsf{DefRules}(A')$,
   $\mathsf{TopRule}(A)$ is undefined.

Notice that we do not allow for the use of defeasible rules within hypothetical arguments. We do however allow for the conclusions of defeasible arguments to be imported inside of a proof by contradiction. This is motivated by the fact that allowing for proofs by contradiction amounts to allowing for transpositions of any rule that can be used within a proof by contradiction, and transpositions are usually assumed only for strict rules in structured argumentation [6,11].

We now need to define the attack relation in our framework. Notice that in ASPIC-END, we also allow for an argument A to attack an argument B which makes an assumption φ if A concludes that φ is not assumable. For example, if one were to assume that the number 5 is yellow, since numbers do not have colors, it should be possible to attack the argument that introduces this assumption and any argument making an inference from this assumption.

Definition 8. Let $\Sigma = (\mathcal{L}, \mathcal{R}, n, \leq)$ be an argumentation theory and A, B two arguments on the basis of Σ. We say that A *attacks* B iff A *rebuts*, *undercuts* or *assumption-attacks* B, where:

- A *rebuts* argument B (on B') iff $\mathsf{Conc}(A) = \neg\varphi$ or $\neg\mathsf{Conc}(A) = \varphi$ for some $B' \in \mathsf{Sub}(B)$ of the form $B''_1, \ldots, B''_n \Rightarrow \varphi$ and $\mathsf{As}(A) = \emptyset$.
- A *undercuts* argument B (on B') iff $\mathsf{Conc}(A) = \neg n(r)$ or $\neg\mathsf{Conc}(A) = n(r)$ for some $B' \in \mathsf{Sub}(B)$ such that $\mathsf{TopRule}(B') = r$, there is no $\varphi \in \mathsf{As}(B')$ such that $\neg\varphi = \mathsf{Conc}(A')$ or $\varphi = \neg\mathsf{Conc}(A')$ for some $A' \in \mathsf{Sub}(A)$, and there

are arguments $B_1, ..., B_n$ such that $B_1 = B'$, $B_n = B$, $B_i \in \mathsf{Sub}(B_{i+1})$ for $1 \leq i < n$ and $\mathsf{As}(A) \subseteq \mathsf{As}(B_1) \cup \cdots \cup \mathsf{As}(B_n)$.

- A *assumption-attacks* B (on B') iff for some $B' \in \mathsf{Sub}(B)$ such that $B' = \mathsf{Assume}(\varphi)$, $\mathsf{Conc}(A) = \neg \mathit{Assumable}(\varphi)$ and $\mathsf{As}(A) = \emptyset$.

We require that any attacking argument A is making fewer assumptions than the B' it attacks, as to prevent arguments from attacking outside of their assumption scope. Note that in the case of rebuttal, since the attacked argument cannot have assumptions, we require that the attacking argument have none either.

In the case of undercutting, we also have the requirement that A does not use the contrary of any assumptions made by B' in any of its inferences, since the attack would not stand in the scope of B'. Additionally, we allow A to make use of any assumptions appearing in the chain of arguments leading B' to B, as these assumptions, even if they have been retracted, still constitute valid grounds on which to form an attack.

Similarly as in ASPIC+, one can also define a notion of successful attack by lifting the preference relation from rules to arguments as follows:

Definition 9. Let $\Sigma = (\mathcal{L}, \mathcal{R}, n, \leq)$ be an argumentation theory and A, B be two arguments on the basis of Σ. We define the *lifting of* $\leq$ *to arguments* $\preceq$ to be such that $A \preceq B$ iff there exists $r_a \in \mathsf{DefRules}(A)$, such that for all $r_b \in \mathsf{DefRules}(B)$, we have $r_a \leq r_b$. We also define $A \prec B$ by replacing $\leq$ with $<$ in the definition of $\preceq$.

Notice that this lifting corresponds to elitist weakest-link as described in [12]. We believe that this ordering is best suited for modeling philosophical and scientific arguments.

We now define what it means for an attack to be successful:

Definition 10. Let $\Sigma = (\mathcal{L}, \mathcal{R}, n, \leq)$ be an argumentation theory, A, B be two arguments on the basis of Σ. We say that A *successfully rebuts* B iff A rebuts B on B' for some argument B' and $A \not\prec B'$, and that A *defeats* B iff A assumption-attacks, undercuts or successfully rebuts B.

The aim of our system is to generate an EAF as defined in Section 2. For this three things need to be specified: A set $\mathcal{X}$ of explananda, a condition under which an argument explains an explanandum, and a condition under which an argument explains another argument. The first two of these three details are domain-specific, and are thus to be specified in an instantiation of the ASPIC-END framework. The third one, on the other hand, should be the same in all domains. The reason for this can be found in the informal clarification that Šešelja and Straßer [16] provided for what it means to say that an argument b explains an argument a: “argument b can be used to explain one of the premises of argument a [. . .] or the link between the premises and the conclusion.”

In the context of structured argumentation, this informal clarification can be turned into a formal definition:

Definition 11. Let A, B be arguments. We say that B *explains* A (on A') iff $A' \in \mathsf{Sub}(A)$, $\mathsf{As}(B) \subseteq \mathsf{As}(A')$ and at least one of the following two cases holds:

- $A' \notin \mathsf{Sub}(B)$ and either $A' = (\rightsquigarrow \mathsf{Conc}(B))$ or $A' = (\Rightarrow \mathsf{Conc}(B))$.
- $\mathsf{Conc}(B) = n(\mathsf{TopRule}(A'))$ and $\nexists B' \in \mathsf{Sub}(B)$ such that $\mathsf{TopRule}(B') = \mathsf{TopRule}(A')$.

Intuitively, the idea behind this definition is that an argument B explains another argument A if B non-trivially concludes one of A's premises or one of the inference rules used by A.

We now have all the elements needed to build an EAF.

Definition 12. Let $\Sigma = (\mathcal{L}, \mathcal{R}, n, \leq)$ be an argumentation theory. Let $\mathcal{X}$ be a set of explananda, and let $\mathcal{C}$ be a criterion for determining whether an argument constructed from Σ explains a given explanandum $E \in \mathcal{X}$. The *explanatory argumentation framework* (EAF) *defined by* $(\Sigma, \mathcal{X}, C)$ is a tuple $\langle \mathcal{A}, \mathcal{X}, \rightarrow, \dashrightarrow \rangle$, where:

- $\mathcal{A}$ is the set of all arguments that can be constructed from Σ satisfying Definition 7;
- $(A, B) \in \rightarrow$ iff A defeats B, where $A, B \in \mathcal{A}$;
- $(A, E) \in \dashrightarrow$ iff criterion $\mathcal{C}$ is satisfied with respect to A and E, where $A \in \mathcal{A}$ and $E \in \mathcal{X}$;
- $(A, B) \in \dashrightarrow$ iff A explains B according to Definition 11, where $A, B \in \mathcal{A}$.

Once such a framework has been generated, we want to be able to extract the most interesting sets of arguments. Such a set should be able to explain as many explananda in as much detail as possible, while being self-consistent and plausible.

We define two kinds of extensions corresponding to the two selection procedures defined by Šešelja and Straßer [16]. As suggested in the informal discussion in their paper, we chose to give higher importance to the criterion of defense compared to the criterion of explanatory power. This prevents some absurd theories which manage to explain all explananda but cannot defend themselves against all attacks from beating plausible theories which fail to explain some of the explananda but are sound and fully defended.

Definition 13. Let $\Sigma = (\mathcal{L}, \mathcal{R}, n, \leq)$ be an argumentation theory, $\Delta = \langle \mathcal{A}, \mathcal{X}, \rightarrow, \dashrightarrow \rangle$ the EAF defined by Σ and $S \subseteq \mathcal{A}$ a set of arguments.

1. We say that S is *satisfactory* iff S is admissible and there is no $S' \subseteq \mathcal{A}$ such that $S' >_p S$ and S' is admissible.
2. We say that S is *insightful* iff S is satisfactory and there is no $S' \subseteq \mathcal{A}$ such that $S' >_d S$ and S' is satisfactory.
3. We say that S is an *argumentative core extension* (*AC-extension*) of Δ iff S is satisfactory and there is no $S' \supset S$ such that S' is satisfactory.
4. We say that S is an *explanatory core extension* (*EC-extension*) of Δ iff S is insightful and there is no $S' \subset S$ such that S' is insightful.

The AC-extensions are sets of arguments which represent the theories explaining the most explananda, together with all other compatible beliefs present in the framework. EC-extensions represent the core of those theories and only include the arguments which defend or provide details for them.

We define the conclusions of the arguments in a given extension as follows:

Definition 14. Let $\Sigma = (\mathcal{L}, \mathcal{R}, n, \leq)$ be an argumentation theory, $\Delta = \langle \mathcal{A}, \mathcal{X}, \rightarrow, \dashrightarrow \rangle$ be the EAF defined by Σ and S be an extension of Δ. Then, we define the *conclusions of* S, denoted $\mathsf{Concs}(S)$, to be $\mathsf{Concs}(S) = \{\mathsf{Conc}(A) | A \in S$ s.t. $\mathsf{As}(A) = \emptyset\}$.

4 Modelling explanations of semantic paradoxes in ASPIC-END

In this section, we discuss how ASPIC-END can be applied to modelling argumentation about explanations of semantic paradoxes, and illustrate this potential application with a simple example. We start by briefly motivating this application of structured argumentation theory.

Philosophy is an academic discipline in which good argumentative skills are a central part of every student's training. Philosophical texts are often much richer in explicit formulation of arguments than texts from other academic disciplines. For these reasons, we believe that modeling arguments from philosophical textbooks, monographs and papers can be an interesting test case for structured argumentation theory.

Different areas of philosophy vary with respect to how much logical rigor is commonly applied in the presentation of arguments. Even logically rigorous argumentation poses many interesting problems, as the rich literature on abstract and structured argumentation attests. In order to not confound these interesting problems with issues arising from the lack of logical rigor, it is a good idea to concentrate on the study of logically rigorous argumentation. Philosophical logic is an area of logic where logically rigorous arguments abound. One topic that has gained a lot of attention in philosophical logic is the study of semantic paradoxes such as the Liar paradox and Curry's paradox [2,10]. We therefore use the argumentation about the various explanations of the paradoxes that have been proposed in the philosophical literature as a test case for structured argumentation theory.

In our application of ASPIC-END to argumentation about explanations of semantic paradoxes, the explananda are the paradoxes (i.e. arguments that derive an absurdity under no assumption without using defeasible rules), which other arguments can explain by attacking the said derivation. So we instantiate the set $\mathcal{X}$ of explananda and criterion $\mathcal{C}$ for an explanation of an explanandum by an argument as specified in the following two definitions:

Definition 15. Let $\Sigma = (\mathcal{L}, \mathcal{R}, n, \leq)$ be an argumentation theory. For every argument A on the basis of Σ such that $\mathsf{DefRules}(A) = \emptyset, \mathsf{As}(A) = \emptyset$ and $\mathsf{Conc}(A) = \bot$, we stipulate an explanandum E_A, and say that $\mathsf{Source}(E_A) = A$.

We define the set $\mathcal{X}$ of explananda based on Σ to be the set of all explananda E_A that we have thus stipulated.

Definition 16. Let $\Sigma = (\mathcal{L}, \mathcal{R}, n, \leq)$ be an argumentation theory, A, B arguments and E an explanandum based on Σ. We say that A *explains* E iff A defeats $\mathsf{Source}(E)$.

The following example illustrates an application of ASPIC-END to a version of the Liar paradox and two very simple explanations of it:[1]

Example: Define L to be the sentence "L is false". If L is true, i.e. "L is false" is true, then L is false, which is a contradiction. So L is not true, i.e. L is false. So "L is false" is true, i.e. L is true. So we have the contradiction that L is both true and false from no assumption.

A truth-value gap explanation: L is neither true nor false. When concluding that L is false because L is not true, we are making the assumption that any sentence is either true or false. This assumption does not hold for problematically self-referential sentences such as L.

A paracomplete explanation: The reasoning that led to the conclusion that L is not true is a proof by contradiction that derives a contradiction from the assumption that L is true. However, a proof by contradiction based on assumption ϕ can only be accepted once one accepts that the law of excluded middle holds for ϕ, i.e. that $\phi \vee \neg\phi$. However, the law of excluded middle should not be accepted for problematically self-referential statements like L, and thus also not to the statement "L is true". So "L is true" cannot be assumed for a proof by contradiction.

We now proceed to the ASPIC-END model of the reasoning and argumentation involved in the paradox and the two explananda. We use T, F and Psr to mean *true*, *false* and *problematically self-referential* respectively. The rules in our model are $\mathcal{R}_{is} = \{T(L) \rightsquigarrow T(F(L)); T(F(L)) \rightsquigarrow F(L); T(L), F(L) \rightsquigarrow \bot; \neg T(L) \rightsquigarrow F(L); F(L) \rightsquigarrow T(F(L)); T(F(L)) \rightsquigarrow T(L)\}$ with $n(\neg T(L) \rightsquigarrow F(L)) = r_1$ and $\mathcal{R}_d = \{\Rightarrow \neg T(L) \wedge \neg F(L); \neg T(L) \wedge \neg F(L) \Rightarrow \neg r_1; \Rightarrow Psr(L); Psr(L) \Rightarrow \neg T(L) \wedge \neg F(L); \neg T(L) \wedge \neg F(L) \Rightarrow \neg Assumable(T(L))\}$. We also define the predicate $Expl$ to be: $Expl(A)$ iff $\mathsf{DefRules}(A) = \emptyset, \mathsf{As}(A) = \emptyset$ and $\mathsf{Conc}(A) = \bot$.

Infinitely many arguments can be constructed from this argumentation theory. However, the following set of arguments is the set of most relevant arguments, in the sense that other arguments will not defeat these arguments and will not add relevant new conclusions.

$$\begin{aligned}
A_1 &= \mathsf{ProofByContrad}(\neg T(L), (\mathsf{Assume}(T(L)),\\
&\quad ((\mathsf{Assume}(T(L)) \rightsquigarrow T(F(L))) \rightsquigarrow F(L)) \rightsquigarrow \bot)) \rightsquigarrow F(L)\\
A_2 &= ((A_1 \rightsquigarrow T(F(L))) \rightsquigarrow T(L)), A_1 \rightsquigarrow \bot\\
B_1 &= (\Rightarrow Psr(L)) \Rightarrow \neg T(L) \wedge \neg F(L)\\
B_2 &= (\Rightarrow \neg T(L) \wedge \neg F(L)) \Rightarrow \neg r_1\\
C &= ((\Rightarrow Psr(L)) \Rightarrow \neg T(L) \wedge \neg F(L)) \Rightarrow \neg Assumable(T(L))
\end{aligned}$$

[1] See [10] for comprehensive presentations of truth-value gap and paracomplete explanations, besides many others.

We get the explanandum E with $\mathsf{Source}(E) = A_2$. B_2 defeats A_2 on A_1 and C defeats A_2 on $\mathsf{Assume}(T(L))$, thus they both explain E. B_1 explains B_2 by non-trivially concluding $\neg T(L) \wedge \neg F(L)$. The AC-extension is $\{B_1, B_2, C\}$ and the EC-extensions are $\{B_1, B_2\}$ and $\{C\}$.[2]

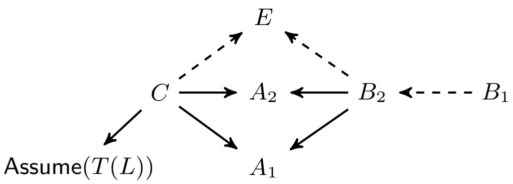

Fig. 1. The relevant arguments, explanandum, attacks and explanations from Example

5 Closure and rationality postulates

In this section, we show that ASPIC-END satisfies four rationality postulates analogous to the four postulates that Modgil and Prakken [11] have established for ASPIC+, as well as two new postulates motivated by the application of structured argumentation to the domain of philosophical logic.

The first postulate concerns the closure of the extensions under the sub-argument relation. The idea is that one cannot accept an argument while rejecting part of it.

Theorem 1. Let $\Sigma = (\mathcal{L}, \mathcal{R}, n, \leq)$ be an argumentation theory, $\Delta = \langle \mathcal{A}, \mathcal{X}, \rightarrow, \dashrightarrow \rangle$ be the EAF defined by Σ and S be an AC-extension of Δ. Then, for all $A \in S$, $\mathsf{Sub}(A) \subseteq S$.

The proof of Theorem 1 rests on the following lemma, which can be proven in a straightforward way as in the case of ASPIC+ (see Lemma 35 of [11]):

Lemma 1. Let $\Sigma = (\mathcal{L}, \mathcal{R}, n, \leq)$ be an argumentation theory, $\Delta = \langle \mathcal{A}, \mathcal{X}, \rightarrow, \dashrightarrow \rangle$ be the EAF defined by Σ, $S \subseteq \mathcal{A}$ and $A, B \in \mathcal{A}$. We have that:

1. If S defends A and $S \subseteq S'$, then S' defends A.
2. If A defeats B' and $B' \in \mathsf{Sub}(B)$, then A defeats B.
3. If S defends A and $A' \in \mathsf{Sub}(A)$, then S defends A'.

Proof of Theorem 1: Let $A \in S$ and $A' \in \mathsf{Sub}(A)$. Suppose $S \cup \{A'\}$ is not conflict-free. Then, either some $B \in S$ defeats A', or A' defeats some $B' \in S$. Since S defends itself, if A' defeats $B' \in S$, then there exists B which defeats A'. So in both cases there exists $B \in S$ which defeats A'. But then by Lemma 1.2,

[2] Notice that both solutions appear in the same AC-extension. This is only due to the brevity of our example. In a more comprehensive exposition of these explanations, arguments attacking other explanations would be included, and thus each AC-extension would contain no more than one solution.

B defeats A, so S is not conflict-free, which is a contradiction. So $S \cup \{A'\}$ is conflict-free. Also, since S defends A, by Lemma 1.3, S also defends A'. Hence, by maximality of the AC-extensions, $A' \in S$. □

Notice that this postulate does not hold for EC-extensions, as they are by definition minimal in their inclusion of arguments, and thus will often leave out low-level sub-arguments.

The second postulate concerns the closure of the conclusions under intuitively strict rules. In the case of ASPIC+, the corresponding postulate concerned the closure of the conclusions under all strict rules (see Theorem 13 in [11]. But since ASPIC-END allows for the rejection of intuitively strict rules, it is undesirable to consider the closure under all of them. Instead, we consider the closure under the accepted intuitively strict rule. The following two definitions define the set of *accepted* intuitively strict rules and the *closure* under a given set of intuitively strict rules:

Definition 17. Let $\Sigma = (\mathcal{L}, \mathcal{R}, n, \leq)$ be an argumentation theory, $\Delta = \langle \mathcal{A}, \mathcal{X}, \rightarrow, \dashrightarrow \rangle$ be the EAF defined by Σ and S be an extension of Δ. The *set of intuitively strict rules accepted by* S is $\mathcal{R}_{isa}(S) = \{r \in \mathcal{R}_{is} | \forall A \in \mathcal{A}$ s.t. $\mathsf{As}(A) = \emptyset$ and $\mathsf{Conc}(A) = \neg n(r)$ or $\neg \mathsf{Conc}(A) = n(r), \exists B \in S$ s.t. B defeats $A\}$.

Definition 18. Let $\Sigma = (\mathcal{L}, \mathcal{R}, n, \leq)$ be an argumentation theory, $P \subseteq \mathcal{L}$ and $R' \subseteq \mathcal{R}_{is}$. We define the *closure of* P *under the set of rules* R', denoted $Cl_{R'}(P)$, as the smallest set such that $P \subseteq Cl_{R'}(P)$, and when $(\varphi_1, ..., \varphi_n \rightsquigarrow \psi) \in R'$ and $\varphi_1, ..., \varphi_n \in Cl_{R'}(P)$, then $\psi \in Cl_{R'}(P)$.

Now the postulate on the closure under accepted intuitively strict rules can be formulated as follows:

Theorem 2. Let $\Sigma = (\mathcal{L}, \mathcal{R}, n, \leq)$ be an argumentation theory, $\Delta = \langle \mathcal{A}, \mathcal{X}, \rightarrow, \dashrightarrow \rangle$ be the EAF defined by Σ and S be an AC-extension of Δ. Then, $\mathsf{Conc}(S) = Cl_{\mathcal{R}_{isa}(S)}(\mathsf{Concs}(S))$.

Proof: We need to show that if $(\varphi_1, ..., \varphi_n \rightsquigarrow \psi) \in \mathcal{R}_{isa}(S)$ and $\varphi_1, ..., \varphi_n \in \mathsf{Concs}(S)$, then $\psi \in \mathsf{Concs}(S)$. Supposing these conditions are met, there exist arguments $A_1, ..., A_n$ with conclusions $\varphi_1, ..., \varphi_n$ respectively. We can then construct $A = A_1, ..., A_n \rightsquigarrow \psi$. Since $A_1, ..., A_n$ are defended by S and $\mathsf{TopRule}(A)$ is accepted by S, A is also defended by S, so $A \in S$. Hence, $\psi \in \mathsf{Concs}(S)$. □

The last two postulates presented in [11] are direct and indirect consistency, which state that when the set of strict rules is consistent, the set of conclusions and the closure of this set under strict rules are consistent.

We have three requirements for applying the consistency postulates. The first is that there cannot be non-defeasible arguments which contradict each other. The second requirement ensures that a formula and its negation are considered as contradictory and the third guarantees that no assumptions are prevented. The last two requirements are motivated by the consideration that in the applications of ASPIC-END not related to paradoxes, one would likely accept classical or intuitionistic logic, for both of which these requirements hold.

Definition 19. Let $\Sigma = (\mathcal{L}, \mathcal{R}, n, \leq)$ be an argumentation theory. We say that Σ is *consistency-inducing* iff:

1. there are no $A, B \in \mathcal{A}$ such that $\mathsf{DefRules}(A) = \mathsf{DefRules}(B) = \emptyset$ and $\mathsf{Conc}(A) = \neg\mathsf{Conc}(B)$,
2. for each $\varphi \in \mathcal{L}$ there is a rule r_φ of the form $\varphi, \neg\varphi \rightsquigarrow \bot \in \mathcal{R}_{is}$ such that $n(r_\varphi)$ is undefined,
3. there is no rule $r \in \mathcal{R}$ such that the conclusion of r is of the form $\neg \mathit{Assumable}(\varphi)$.

The following theorem establishes direct consistency for ASPIC-END:

Theorem 3. Let $\Sigma = (\mathcal{L}, \mathcal{R}, n, \leq)$ be a consistency-inducing argumentation theory, $\Delta = \langle \mathcal{A}, \mathcal{X}, \rightarrow, \dashrightarrow \rangle$ be the EAF defined by Σ and S be an AC or EC-extension of Δ. Then, there does not exist $\varphi \in \mathsf{Concs}(S)$ such that $\neg\varphi \in \mathsf{Concs}(S)$.

Proof: Suppose for a contradiction that there exists $\varphi \in \mathsf{Conc}(S)$ such that $\neg\varphi \in \mathsf{Conc}(S)$. Then, there exist two arguments $A, B \in S$ such that $\mathsf{Conc}(A) = \varphi$ and $\mathsf{Conc}(B) = \neg\varphi$. Since Σ is consistency-inducing, at least one of A and B has a defeasible sub-argument. For each maximal (w.r.t Sub) sub-argument C of A with a defeasible top rule, let A_C be the copy of A that has $\mathsf{Assume}(\mathsf{Conc}(C))$ instead of C (so $\mathsf{As}(A_C) = \{\mathsf{Conc}(C)\}$), and let D_C be $\mathsf{ProofByContrad}(\neg\mathsf{Conc}(C), A_C, B \rightsquigarrow \bot)$ (so D_C rebuts C). We can do this as well for every maximal sub-argument of B with a defeasible top rule. Then for at least one such sub-argument C of A or B, say of A, $A_C \not\prec C$ and $B \not\prec C$, hence $D_C \not\prec C$, and so D_C will defeat C. Then D_C defeats A on C. So some $F \in S$ defeats D_C. Since $B \in S$, F does not defeat B, so F defeats A_C. Since $\mathsf{Conc}(F) \neq \neg\mathit{Assumable}(\mathsf{Conc}(C))$ by item 3 of Definition 19 and F does not defeat A, $As(F) = \{C\}$. By Theorem 1, $C \in S$. Let F' be hte copy of F that has C instead of $\mathsf{Assume}(C)$. Then F' defeats A. So some argument $G \in S$ defeats F'. but then G defeats F or C, which is a contradiction. □

Indirect consistency of AC-extensions follows from closure under accepted intuitively strict rules together with direct consistency:

Theorem 4. Let $\Sigma = (\mathcal{L}, \mathcal{R}, n, \leq)$ be a consistency-inducing argumentation theory, $\Delta = \langle \mathcal{A}, \mathcal{X}, \rightarrow, \dashrightarrow \rangle$ be the EAF defined by Σ and S be an AC-extension of Δ. Then, there does not exist $\varphi \in Cl_{\mathcal{R}_{isa}(S)}(\mathsf{Concs}(S))$ such that $\neg\varphi \in Cl_{\mathcal{R}_{isa}(S)}(\mathsf{Concs}(S))$.

As explained in Section 2, we want ASPIC-END to be applicable to domains like philosophical logic, in which the correctness of logical rules can be up for debate. Among the proposals made by philosophers of how to handle the semantic paradoxes, there is paraconsistent dialetheism [14], which accepts some inconsistencies as true and uses a paraconsistent logic to avoid that everything can be derived. And in order to be able to show the internal structure of the paradox, we need to have an inconsistency arise from intuitively strict rules under no assumptions. For these reasons, the consistency postulates do not make sense for this kind of application of ASPIC-END.

However, there is a property similar to consistency that should still hold even when the intuitively strict rules lead to paradoxes and when the output extensions contain one that accepts paraconsistent dialetheism, namely that an extension should never be trivial, i.e. conclude everything.

For the non-triviality of the extensions, we require that rules are present in the framework which allow one to derive any formula from $\bot$.[3] We also require these rules of conjunction elimination from $\bot$ not to have a corresponding formula in $\mathcal{L}$ as a name, which prevents them from being attackable. Also, we require every other intuitively strict rule to have a name so that it can be attacked. We say that the argumentation theory is well-defined if it satisfies these requirements, and assume well-definedness in the non-triviality postulate stated in Theorem 5.

Theorem 5. Let $\Sigma = (\mathcal{L}, \mathcal{R}, n, \leq)$ be an argumentation theory, $\Delta = \langle \mathcal{A}, \mathcal{X}, \rightarrow, \dashrightarrow \rangle$ be the EAF defined by Σ, and S be an AC or EC-extension of Δ. Then, $\bot \notin \mathsf{Concs}(S)$.

Proof: Suppose for a contradiction that $\bot \in \mathsf{Concs}(S)$. Then there exists a minimal (under sub-argument relation) argument $A \in S$ such that $\mathsf{Conc}(A) = \bot$ and $\mathsf{As}(A) = \emptyset$. Let $r = \mathsf{TopRule}(A)$. If $r \in \mathcal{R}_{is}$, then from Definition 6, $n(r) \in \mathcal{L}$ and so let $B = A \rightsquigarrow \neg n(r)$. Otherwise, let $B = A \rightsquigarrow \neg\bot$. By Definition 9, $B \not\prec A$. Then B undercuts or successfully rebuts A on A, so B defeats A. Since S is an AC- or EC-extension of Δ, it defends itself, so there exists $C \in S$ such that C defeats B. Suppose for a contradiction that C defeats B on $B' \neq B$. Since $\mathsf{Sub}(B) = \mathsf{Sub}(A) \cup \{B\}$, $B' \in \mathsf{Sub}(A)$. Then, by Lemma 1.2, C defeats A on B'. But S is conflict-free, so we have a contradiction. Hence, C defeats B on B. Since $B = A \rightsquigarrow \neg n(r)$, B cannot be rebutted nor assumption-attacked. Hence, C undercuts B on B. But from Definition 6 and since $\mathsf{TopRule}(B) \in \mathcal{R}_{ce}$, $n(\mathsf{TopRule}(B))$ is undefined, i.e. no argument undercuts B on B, a contradiction. Hence, $\bot \notin \mathsf{Concs}(S)$. □

Indirect non-triviality of AC-extensions then follows from closure under accepted intuitively strict rules and direct non-triviality:

Theorem 6. Let $\Sigma = (\mathcal{L}, \mathcal{R}, n, \leq)$ be an argumentation theory, $\Delta = \langle \mathcal{A}, \mathcal{X}, \rightarrow, \dashrightarrow \rangle$ be the EAF defined by Σ and S be an AC-extension of Δ. Then, $\bot \notin Cl_{\mathcal{R}_{isa}(S)}(\mathsf{Concs}(S))$.

6 Conclusion and Future Work

We have proposed a modification of ASPIC+ called ASPIC-END, which incorporates a formal model of explanations, and features natural-deduction style arguments. We have shown how ASPIC-END can be instantiated for modelling argumentation about explanations of semantic paradoxes in ASPIC-END. Finally, we have shown that ASPIC-END satisfies rationality postulates analogous

[3] As noted earlier, we interpret $\bot$ as the conjunction of all formulas in $\mathcal{L}$, so these rules are in effect conjunction elimination rules.

to those satisfied by ASPIC+, as well as non-triviality postulates that are relevant in the application to semantic paradoxes.

One topic of our future work on ASPIC-END is to study possible ways of instantiating explananda and explanations in other scientific domains. For explanations from the natural sciences, this might require an instantiation of ASPIC-END with a language covering causal notions. Furthermore, we will study the possibility of integrating the new results of Beirlaen, Heyninck and Straßer [3] on reasoning by cases in structured argumentation with our work on ASPIC-END.

References

1. Pietro Baroni, Martin Caminada, and Massimiliano Giacomin. An introduction to argumentation semantics. *The Knowledge Engineering Review*, 26(4):365–410, 2011.
2. Jc Beall, Michael Glanzberg, and David Ripley. Liar Paradox. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition, 2016.
3. Mathieu Beirlaen, Jesse Heyninck, and Christian Straßer. Reasoning by Cases in Structured Argumentation. In *Proceedings of SAC/KRR 2017*, in press.
4. Philippe Besnard, Alejandro Garcia, Anthony Hunter, Sanjay Modgil, Henry Prakken, Guillermo Simari, and Francesca Toni. Introduction to structured argumentation. *Argument & Computation*, 5(1):1–4, 2014.
5. Martin Caminada. A formal account of socratic-style argumentation. *Journal of Applied Logic*, 6(1):109–132, 2008.
6. Martin Caminada and Leila Amgoud. On the evaluation of argumentation formalisms. *Artificial Intelligence*, 171(5-6):286–310, 2007.
7. Martin Caminada, Sanjay Modgil, and Nir Oren. Preferences and unrestricted rebut. In *Computational Models of Argument - Proceedings of COMMA 2014*, pages 209–220, 2014.
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(2):321–357, 1995.
9. Solomon Feferman. Reflecting on incompleteness. *The Journal of Symbolic Logic*, 56(01):1–49, 1991.
10. Hartry Field. *Saving Truth from Paradox*. Oxford University Press, 2008.
11. Sanjay Modgil and Henry Prakken. A general account of argumentation with preferences. *Artificial Intelligence*, 195:361–397, 2013.
12. Sanjay Modgil and Henry Prakken. The ASPIC+ framework for structured argumentation: a tutorial. *Argument & Computation*, 5(1):31–62, 2014.
13. Henry Prakken. An abstract framework for argumentation with structured arguments. *Argument & Computation*, 1(2):93–124, 2010.
14. Graham Priest. *In Contradiction: A Study of the Transconsistent*. Oxford University Press, 2006.
15. William N Reinhardt. Some remarks on extending and interpreting theories with a partial predicate for truth. *Journal of Philosophical Logic*, 15(2):219–251, 1986.
16. Dunja Šešelja and Christian Straßer. Abstract argumentation and explanation applied to scientific debates. *Synthese*, 190(12):2195–2217, 2013.