UNIVERSITÉ DU
LUXEMBOURG

Faculty of Science, Technology and Communication

Modeling Arguments about the Liar Paradox using Formal Argumentation Theory

Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master in Information and Computer Sciences

Author:
Jeremie Dauphin

Supervisor:
Prof. Leon van der Torre

Reviewer:
Prof. Pierre Kelsen

Advisor:
Marcos Cramer

August 2016

Abstract

The connection between the methods developed in formal argumentation theory and real human reasoning are not yet well-researched. Arguments about logical paradoxes are an interesting test-case for researching such connections, as they include conflicting information and as they can be formally modeled without modeling much world knowledge. We introduce two formalisms that model such arguments. We first introduce ASPIC-END, a structured argumentation system encompassing explanations and natural deduction style proofs by contradiction. We then introduce EEAFs, an extension of explanatory argumentation frameworks, providing a more manual approach to formalizing arguments, which allows for more expressive power. We use these two formalism to model and analyze various short texts containing arguments about the liar paradox.

Acknowledgements

First, I would like to express my deepest gratitude to my advisor Marcos Cramer for his constant support and guidance and for his precious comments on the writing of this thesis.
I would also like to thank my supervisor, Prof. Leon van der Torre, for his valuable feedback on the formalisms introduced in this thesis and for providing me with an office for the duration of the thesis so that I could work in the best conditions.
Lastly I would like to thank my family and friends for their support and encouragements.

Contents

Chapter 1

Introduction

1.1 Modeling reasoning about conflicting information in Artificial Intelligence

Reasoning about conflicting information is one of the current challenges in the field of Artificial Intelligence. Classical logic is a very good formalism to reason about making inferences from given information. It is however monotonic, meaning that any new information will not affect the validity of conclusions previously made from rules of classical logic. In this aspect, it fails to capture human reasoning which can be non-monotonic and defeasible, as previously made conclusions or inferences might turn out to be disproved or invalidated under the light of new evidence. This is due to the fact that humans have to reason with incomplete and conflicting information and use defeasible rules to make inferences.

As an answer to this, researchers in Artificial Intelligence have been developing not only non-monotonic logics, such as Reiter's default logic [12], but also defeasible argumentation systems. In 1995, Phan Minh Dung introduced a new approach to argumentation called Abstract Argumentation [5]. By abstracting away most of the elements, not only did it reduce the complexity of the system, which was one of the main issues with argumentation systems at the time, but it also provided a simplified system which could accommodate the different existing approaches.

In the study on the connections between argumentation-theoretical formalisms and human reasoning, we are faced with the obstacle that we humans usually make use of world-knowledge in a non-transparent way, making it harder to formalize our reasoning. For this reason, in an attempt to avoid this problem, we propose to study reasoning about conflicting information in the context of mathematics and logic. In particular, argumentation about logical paradoxes is an interesting topic to test argumentation systems, as these paradoxes feature mostly abstract knowledge while still involving conflicting defeasible arguments, unlike standard mathematics which revolve around absolute certainties and consistency. In this thesis, we propose to focus on the liar paradox.

1.2 The liar paradox

The first sentence in this paragraph is a lie. When wondering whether such a sentence is true or false, one discovers that in both cases something goes wrong. Supposing it is true, the first sentence of this paragraph must then be a lie. But if that is the case, then the sentence is merely stating the truth, which is contradictory. Is it false then? This would mean that the first sentence in this paragraph is no lie, and thus states the truth.

However, this truth is then that the first sentence in this paragraph is a lie, even though we have just assumed otherwise. So we cannot assign a truth value in a classical way to such a sentence without ending up with a contradiction.

This is called the liar paradox, and this kind of sentence is called a liar sentence. There are different ways one can formulate a liar sentence, but there is always some kind of circular referencing involved. For example, the following two sentences are together also liar sentences:

The next sentence is true.
The previous sentence is false.

The problem is the same here, trying to assign them a truth value results in a contradiction.

The discussion of this paradox has started around 600 BC with the Epimenides paradox, which was first discussed in a formal setting by T. Fowler in 1887 [7]. The paradox is as follows: a Cretan seer would have stated "all Cretans are liars". But being Cretan himself, this would mean he was a liar too and this his statement would be a lie. The discussion around this kind of sentence has given rise to stronger versions such as the first sentence in this chapter. Many solutions have been proposed for this kind of paradox, some of which deny the applicability of general principles of logic to this kind of problematically self-referencing sentences. For example, one solution is to deny that this kind of sentence is subject to the law of non-contradiction which says that every formula cannot be true and false at the same time. One then also has to deny that the explosion property, also called *ex falso ad quolibet*, applies in this kind of situation. This property allows one to derive anything from a contradiction, and hence saying that a liar sentence is both true and false would allow one to derive that the moon is made of cheese for example. This solution suggests that we should accept the inconsistent nature of some sentences but prevent it from deriving unrelated statements.

1.3 Thesis outline and objectives

In this thesis, we will attempt to model some of the solutions to the paradox using explanatory abstract argumentation. These different solutions all aim at explaining the paradox while also pointing out flaws in the other solutions as to strengthen theirs, hence it seems fitting to use an argumentation model. The ability to explain the paradox being central to the debate, it also seems natural for our model to include some kind of measure of explanatory properties.

This thesis will also contribute to the theoretical work of the interdisciplinary project Cognitive Aspects of Formal Argumentation, which will start in November 2016 and will be led by Prof. Leon Van Der Torre and Prof. Christine Schiltz. In this project, the relation between actual human reasoning and Formal Argumentation Theory will be investigated with the goal of making Formal Argumentation Theory cognitively more plausible, which will then strengthen our understanding of human reasoning. To do so, the project will conduct empirical studies which study hypotheses based on the models built in the course of this thesis.

We will first begin by describing the general notions of formal argumentation in chapter 2. We will then proceed with defining in chapter 3 a system for structured argumentation that extends the widely studied system ASPIC+ with the notion of explanation and with natural-deduction-style arguments. We will then use this method to propose a model for some of the solutions to the liar paradox in chapter 4. After this, we will examine a more abstract modeling method based on meta-argumentation and attempt to model some solutions from a different perspective in chapter 5. We will then summarize our results and list a few related open problems and possible future work in chapter 6.

Chapter 2

Basics of Formal Argumentation

2.1 Abstract argumentation

Abstract argumentation systems were introduced by P. M. Dung in 1995 [5]. They consist of a set of arguments and an attack relation between them. The abstraction is present in both of these elements. On the one hand, the internal structure of the arguments is ignored and hence they are all treated in a similar way. On the other hand, the attack relation between them is also abstracted away, and one does not distinguish between the different kinds of attacks on arguments or even the motivation for the attack.

Definition 2.1.1. Argumentation Frameworks:
An *argumentation framework* (AF) is a pair $\langle \mathcal{A}, \rightarrow \rangle$ where $\mathcal{A}$ is a set of arguments, and $\rightarrow \subseteq \mathcal{A} \times \mathcal{A}$ is a relation between them.

"$a \rightarrow b$" is read as "a attacks b" and $\rightarrow$ is called the attack relation.

Example 2.1.1. $AF_1 = (\{a, b, c\}, \{(a, b), (b, c)\})$ is an argumentation framework with 3 arguments, a, b and c, where a attacks b and b attacks c.

Graphically, we represent an argumentation framework as a graph where the arguments are the nodes and the attacks are the edges.

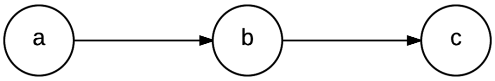

Figure 2.1: Graphical representation of AF1

Argumentation frameworks revolve around the acceptability of an argument. For a given framework $(\mathcal{A}, \rightarrow)$, we would like to say that a set of arguments $S \subseteq \mathcal{A}$ is acceptable or admissible. Let us look at our example with AF_1. The argument a should be acceptable since there is no attack on it. But what about b and c? Notice that by accepting a, we have an acceptable reason to reject b since a attacks b. Therefore, b would not be acceptable. This would also mean that there is no acceptable reason to reject c and so c would be acceptable too, on the condition that a is. This phenomenon is called defense. We say that a defends c because it attacks b which is c's only attacker.

Definition 2.1.2. Defense:
Let AF $= (\mathcal{A}, \rightarrow)$ be an argumentation framework. We say that an argument $a \in \mathcal{A}$

defends an argument $b \in \mathcal{A}$ if and only if for all $c \in \mathcal{A}$ such that $(c, b) \in \rightarrow$, we have $(a, c) \in \rightarrow$.

Similarly, we say that a set of arguments $S \subseteq \mathcal{A}$ *defends* an argument $b \in \mathcal{A}$ (or set of arguments $S' \subseteq \mathcal{A}$) if and only if for all $c \in \mathcal{A}$ such that $(c, b) \in \rightarrow$ (for some $b \in S'$), there exists $a \in S$ such that $(a, c) \in \rightarrow$.

We can now try to express our notion of acceptability in terms of defense, however that alone is not sufficient. We would also like to only consider sets of arguments which are consistent, meaning that if there is an attack between two arguments, it would make little sense to accept them both. Considering our running example, we would not want to accept all of $\{a, b, c\}$. Since a attacks b, the acceptability of a questions the acceptability of b of so it wouldn't make sense to accept both as we would then contradict ourselves. This is where the notion of being conflict-free comes in.

Definition 2.1.3. Conflict-free:
Let $(\mathcal{A}, \rightarrow)$ be an argumentation framework. A set of arguments $S \subseteq \mathcal{A}$ is said to be *conflict-free* if and only if there are no arguments $a, b \in S$ such that $(a, b) \in \rightarrow$.

We can now define what it means for a set of arguments to be admissible. It simply needs to be plausible and consistent, which formally means it must defend itself and be conflict-free.

Definition 2.1.4. Admissible:
Let $(\mathcal{A}, \rightarrow)$ be an argumentation framework. A set of arguments $S \subseteq \mathcal{A}$ is said to be *admissible* if and only if S is conflict-free and S defends S.

However this notion might be a bit incomplete. For any given argumentation framework, the empty set of arguments is always trivially admissible, yet in most cases we would rather have a non-empty set of arguments in the end. In our example of AF_1, even though the set $\{a\}$ is admissible, it seems a bit lacking as it defends the argument c yet does not include it. We would rather want to consider only the set $\{a, c\}$ as it seems to be more meaningful. There are however different ways of selecting the arguments to be accepted as not all frameworks are as straight-forward as AF_1. These are the semantics of abstract argumentation which select sets of arguments that we call *extensions*. We start with the complete extensions.

Definition 2.1.5. Complete extensions:
Let $(\mathcal{A}, \rightarrow)$ be an argumentation framework. A set of arguments $S \subseteq \mathcal{A}$ is called a *complete extension* of $(\mathcal{A}, \rightarrow)$ if and only if it is admissible and for each argument $a \in \mathcal{A}$, if S defends a, then $a \in S$.

In our running example, the empty set is not a complete extension, since it defends the argument a (an argument which is not attacked is trivially defended by all sets of arguments) but does not contain it. Similarly, the set $\{a\}$, even though admissible, is not complete, because it defends c but does not contain it. So the only complete extension of AF_1 is $\{a, c\}$.

Example 2.1.2. Consider the following argumentation framework:

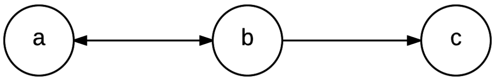

Figure 2.2: Graphical representation of AF2

Notice that AF_2 is similar to AF_1, however there is one additional attack from b to a. We now have again that $\{a, c\}$ is a complete extension, however this time it is not the only one. Notice that b defends itself from a this time by attacking it back. Hence, $\{b\}$ is also a complete extension. Also, notice that every argument is subject to some attack this time, and hence $\emptyset$ does not defend any arguments. Thus, $\emptyset$ is also a complete extension of AF_2.

From a skeptical point of view, one might not want to accept any argument when faced with two arguments attacking each other. For example, if person A says that person B is a liar and vice-versa, accepting the argument of any of them could be fine as they both also constitute a counter-argument to each other. However, one might be inclined not to accept any of their arguments. This skeptical point of view is represented by the grounded extension.

Definition 2.1.6. Grounded extension:

Let $(\mathcal{A}, \rightarrow)$ be an argumentation framework. A set of arguments $S \subseteq \mathcal{A}$ is called the *grounded extension* of $(\mathcal{A}, \rightarrow)$ if and only if it is the minimal (with respect to $\subseteq$) complete extension.

It can be shown that the grounded extension is always unique. In AF_2, the grounded extension is $\emptyset$ and in AF_1, it is $\{a, c\}$.

On a more credulous point of view, one might argue that as long as a set of arguments defends itself, it is valid. One might then want to extract the sets with as many valid arguments as possible, as it might seem unjustified not to consider these arguments. This point of view is reflected in the *preferred extension*.

Definition 2.1.7. Preferred extensions:

Let $(\mathcal{A}, \rightarrow)$ be an argumentation framework. A set of arguments $S \subseteq \mathcal{A}$ is called a *preferred extension* of $(\mathcal{A}, \rightarrow)$ if and only if it is a maximal (with respect to $\subseteq$) complete extension.

In AF_2, the preferred extensions are $\{a, c\}$ and $\{b\}$.

In ideal cases, a preferred extension manages to attack all arguments it does not contain and hence we can split the set of arguments into the ones we accept and the ones we reject. Sometimes however, preferred extensions do not includes some arguments which they do not attack either. For example, suppose there is an argument which attacks itself (for example a paradoxical argument) while not being subject to any other attack. Then, for any preferred extension, it cannot be included in it because it will prevent it from being conflict-free. Yet, no argument from the extension attacks it either since the only argument attacking it is itself. Hence, even though no preferred extension includes it, we cannot say that any extension fully rejects it either. Consider the following example framework:

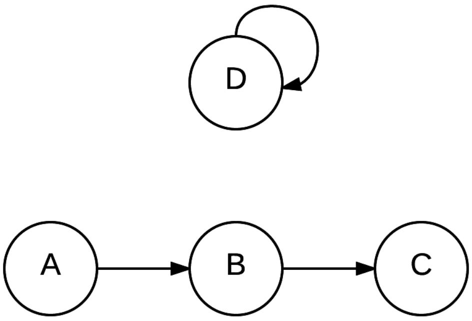

Figure 2.3: Example framework with a self-attacking argument.

The preferred extension is $\{A, C\}$, yet neither A nor C attack D.

The stable semantics capture this idea of ideal cases of preferred extensions and is defined as follows:

Definition 2.1.8. Stable extensions:

Let $(\mathcal{A}, \rightarrow)$ be an argumentation framework. A set of arguments $S \subseteq \mathcal{A}$ is called a *stable extension* of $(\mathcal{A}, \rightarrow)$ if and only if S is a complete extension and for every argument $a \in \mathcal{A} \setminus S$, there exists an argument $b \in S$ such that $b \rightarrow a$.

Note that all stable extensions are also preferred extensions, but not vice-versa. Also, the existence of a stable extension is not guaranteed, meaning that some frameworks do not have any stable extensions such as the one depicted in Figure 2.3.

Example 2.1.3. Let us now consider a slightly more complex example:

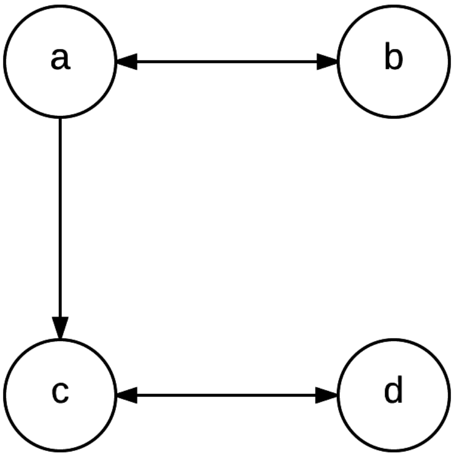

Figure 2.4: Graphical representation of AF3

Here we have two cycles of even size: a and b attack each other, and c and d attack each other. Also, a attacks c. The complete extensions are $\emptyset, \{a, d\}, \{b\}, \{b, c\}$, $\{b, d\}$ and $\{d\}$. The grounded extension is $\emptyset$ while the preferred and stable extensions are $\{a, d\}, \{b, c\}$ and $\{b, d\}$.

2.2 Explanatory Argumentation Frameworks

In scientific debates, the discussions are usually centered around some phenomenons or evidence and the different parties propose theories to explain them. With this idea in mind, D. Šešelja and C. Straßer have extended abstract argumentation framework with explanatory features [13]. In these frameworks, there are not only arguments but also explananda. These are scientific phenomenons of which, unlike arguments, the acceptability is not being questioned. These explananda are being explained by the arguments via an explanation relation which is represented by $\dashrightarrow$. $a \dashrightarrow e$ is read as "a explains e". The explanation relation may also occur between arguments. This corresponds to the fact that real world arguments may explain each other by either explaining one of the other argument's premises or the inference relation that argument is making between its premises and conclusion. The last element is an incompatibility relation between arguments. This relation is used to differentiate between the opposing theories.

Definition 2.2.1. Explanatory argumentation frameworks:

An *explanatory argumentation framework* (EAF) is a tuple $\langle \mathcal{A}, \mathcal{X}, \rightarrow, \dashrightarrow, \sim \rangle$, where $\mathcal{A}$ is a set of arguments, $\mathcal{X}$ is a set of explananda, $\rightarrow \subseteq \mathcal{A} \times \mathcal{A}$ is an attack relation, $\dashrightarrow \subseteq \mathcal{A} \times (\mathcal{A} \cup \mathcal{X})$ is an explanation relation from arguments to either explananda or other arguments, and $\sim \subseteq \mathcal{A} \times \mathcal{A}$ is a symmetric incompatibility relation.

The incompatibility relation might seem similar to a bi-directional attack at first, and acts as such in the selection of conflict-free sets, however there are differences. The main one is in that the incompatibility relation does not appear in the definition of defense, which is unchanged from standard abstract argumentation frameworks. Hence, being incompatible with an argument does not count as defending the arguments it attacks. The intuition behind the concept is that opposing theories usually do not logically exclude each other, however scientists deem it implausible that both hold at the same time. The notion of conflict-freeness must be accordingly revised:

Definition 2.2.2. Conflict-free:

Let $\langle \mathcal{A}, \mathcal{X}, \rightarrow, \dashrightarrow, \sim \rangle$ be an EAF. A set of arguments $S \subseteq \mathcal{A}$ is said to be *conflict-free* if and only if there are no arguments $a, b \in S$ such that $(a, b) \in \rightarrow \cup \sim$.

Note that the definition of admissible sets still stands but with the revised definition of conflict-freeness.

We now need a notion of what it means for a set of arguments to offer an explanation for a given explanandum.

Definition 2.2.3. Explanations:

An *explanation* $X[e]$ for $e \in \mathcal{X}$ offered by a set of arguments S is a subset S' of 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.

Example 2.2.1. Consider the following EAF:

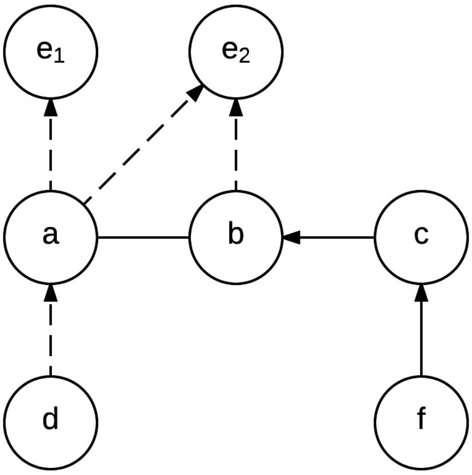

Figure 2.5: Example EAF1

Note that the incompatibility relation has been represented by a straight line with no arrow between a and b.

Here we have two explananda, e_1 and e_2. a explains both e_1 and e_2 while b explains only e_2. Consider the conflict-free set $\{a, d, f\}$. It contains two explanations for e_1, namely $X_1[e_1] = \{a\}$ and $X_2[e_2] = \{a, d\}$. Similarly, it offers two explanations for e_2. The conflict-free set $\{b, f\}$ however offers an explanation only for e_2.

Also, notice that since a and b are incompatible, any set containing both of them cannot be conflict-free. This also allows us to easily differentiate between the two theories when selecting the admissible sets of arguments.

For our goal of selecting the best theory from our model, we need a way to compare how much a given set of arguments is able to explain. We thus need to define a notion of explanatory power.

Definition 2.2.4. Explanatory power:

A set of arguments S_1 is *explanatory at least as powerful* as a set of arguments S_2 ($S_1 \geq_p S_2$) if and only if the set of explananda for which S_1 offers an explanation is a super-set of or equal to the set of explananda for which S_2 offers an explanation.

In our previous example, we have that $\{a, d\} \geq_p \{b\}$ since $\{a, d\}$ offers an explanation for $\{e_1, e_2\}$ while $\{b\}$ only offers an explanation for $\{e_2\}$. Notice however that we use the subset relation to define our notion of explanatory power, as to consider it in a qualitative way. One may also wish to consider it in a quantitative way and compare the sizes of the sets of explananda for which the two sets offer an explanation. However this relies on the assumption that all explananda are of equal importance which is usually false. This is why we stick with the qualitative notion of explanatory power.

In our example, we had that $\{a, d\}$ offered two explanations for e_1, namely $\{a\}$ and $\{a, d\}$. We need a notion of explanatory depth to be able to compare these two explanations.

Definition 2.2.5. Explanatory depth:

An explanation $X_1[e]$ is *explanatory deeper* than another explanation $X_2[e]$ ($X_1[e] \geq_d X_2[e]$) if and only if $X_2[e] \subseteq X_1[e]$.

So in our running example, we have that $\{a, d\} \geq_d \{a\}$. This also means that the explanations offered by $\{a, d, f\}$ are explanatory deeper than the explanations offered by $\{a, f\}$.

Šešelja and Straßer [13] then propose two procedures for the selection of the best sets of arguments with respect to these notions. The first one selects the argumentative core of the most explanatory powerful theories while the second one selects their explanatory core.

Definition 2.2.6. Procedure 1:

The idea of this procedure is to select the *argumentative core* of the most explanatory powerful theories together with arguments which attack rivaling theories. It consists of the following steps:

1. Select all the conflict-free sets.
2. From those, select the most explanatory powerful ones.
3. From those, select the most defended ones.
4. From those, select the maximal ones with respect to $\subseteq$.

Notice that the criterion of defense has been relaxed and this procedure as well as the next one do not require the sets to be fully defended. This is to reflect the fact that most real-world theories are imperfect and still suffer from some flaws yet we cannot simply disregard them.

Also, while the criterion of explanatory power takes higher priority than the criterion of defense here, they also state it is possible to reverse their order if one wishes to give higher priority to the criterion of defense. If steps 2 and 3 are reversed so that step 3 is done before step 2, then steps 1 and 3 collapse into "Select all the admissible sets". Note that later in the thesis we will use this modified version of the procedure.

In the example, the procedure would select the sets $\{a\}$, $\{a, d\}$, $\{a, d, f\}$ and $\{a, d, c\}$ after the first two steps. Note that the sets containing b will not be retained since b explains less than a and both are incompatible. The set $\{a, d, c\}$ suffers an attack from f on c from which it does not manage to defend itself, and hence step 3 will eliminate it as the other sets are fully defended. The last step will then select the maximal remaining set which is $\{a, d, f\}$.

Notice that in this procedure, the explanatory relation between arguments is ignored, i.e. only explanations from arguments to explananda are considered. Also, in a case where the set of explananda $\mathcal{X}$ is empty, this procedure is equivalent to the preferred semantics.

The other procedure focuses on the explanations provided and aims at selecting the promising theories to be further developed.

Definition 2.2.7. Procedure 2:

The idea of this procedure is to select the *explanatory core* of the most explanatory powerful theories. It consists of the following steps:

1. Select all the conflict-free sets.
2. From those, select the most explanatory powerful ones.
3. From those, select the most defended ones.
4. From those, select the explanatory deepest ones.

5. From those, select the minimal ones with respect to $\subseteq$.

Notice the first three steps are the same as in the first procedure and the difference lies in steps 4 and 5. Just as in Procedure 1, steps 2 and 3 can be reversed and the resulting modified version will be used later in the thesis.

In our example, after step 3 the remaining sets are $\{a\}$, $\{a, d\}$ and $\{a, d, f\}$. Steps 4 now eliminates $\{a\}$ as the explanation it offers is not as deep as the others which also contain d. The last step then selects the minimal remaining set which is $\{a, d\}$.

2.3 Argumentation Frameworks with Recursive Attacks

Abstract argumentation frameworks can be extended in multiple ways, allowing them to express real life arguments with more details. This is achieved for example in explanatory argumentation frameworks by adding explananda, an explanatory relation and an incompatibility relation. Let us explore more ways in which abstract argumentation frameworks can be extended.

To understand the motivation for the extension of abstract argumentation frameworks we will examine, consider the following example from Baroni et al. [1]: Bob wants to go on a vacation for Christmas and usually takes the best last minute deal. There are two offers currently available: one week to Cuba or one week to Gstaad. This would give us two arguments: the first one, A, is based on the premise that a one week deal to Cuba is available and concludes that Bob should go to Cuba, and the other one, B, is based on the premise that a one week deal to Gstaad is available and concludes that he should go to Gstaad. Now since Bob can obviously not go to both locations at the same time, these two arguments are attacking each other.

Now suppose Bob has a preference for skiing and knows that Gstaad is a ski resort. We could represent this as an argument C based on the premise that Bob likes skiing, and the conclusion would be that Bob prefers going to a ski resort if possible. Notice that this argument does not attack A directly since, by itself, it does not reject the fact that there is a last minute deal for Cuba nor the fact that he should go there. It does however give a reason not to choose Cuba over Gstaad and hence attacks the attack relation from A to B, attempting to render it ineffective much like it would with an argument.

Suppose Bob then learns that there have been no snowfalls in Gstaad for the past month, meaning it might not be possible for him to ski there. This would gives us another argument D based on the premise that there hasn't been any snowfall in the past month and concluding that Bob might not be able to ski there. Now again, D does not directly attack the argument C as it does not reject the fact that Bob has a preference for skiing nor that he would thus prefer going to a ski resort. It does however attack the attack originating from C, since it annuls the impact that Bob's preference for ski resorts has on the choice between Cuba and Gstaad.

Finally, suppose that Bob now learns that it is still possible to ski in Gstaad thanks to an abundant amount of artificial snow which makes up for the lack of snowfall. We would model this as an argument E based on the premise that there is an abundant amount of artificial snow which concludes that even though there is a lack of snowfall, it is still possible to ski in Gstaad. Here there is a direct attack from E to D since it attacks the fact that Bob might not be able to ski there due to lack of snowfall. We could model this informal analysis as follows:

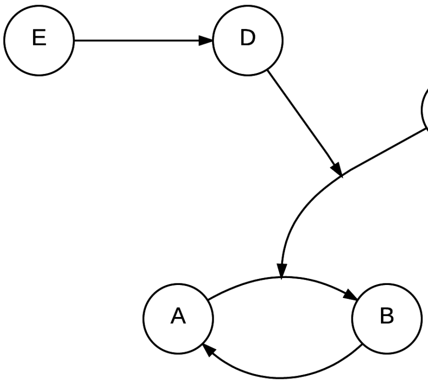

Figure 2.6: Extended argumentation framework representing our example

This motivates allowing attacks to target not only arguments but also other attack relations. This concept of representing a preference from an argument a over another argument b as an argument attacking the attack from b to a has been observed by Modgil in [9]. In that paper, he introduces a type of frameworks in which there is a second attack relation, representing the fact that arguments may attack elements of the first attack relation. We will however focus on Argumentation Frameworks with Recursive Attacks (AFRA), a formalism proposed by Baroni et al.[1] which allows for attacks to be nested on an unbounded number of levels. An AFRA, similarly to a classic abstract argumentation framework, is made of a set of argument and a set of attacks. The difference is that the attacks are pairs where the first member is an argument and the second member is either an argument (base case) or an attack (recursive case).

Definition 2.3.1. AFRA

An *Argumentation Framework with Recursive Attacks* (AFRA) is a pair $\langle \mathcal{A}, \rightarrow \rangle$ where:

- $\mathcal{A}$ is a set of arguments
- $\rightarrow \subseteq \mathcal{A} \times (\mathcal{A} \cup \rightarrow)$ is an attack relation from arguments to either arguments or attacks

For a given attack $\alpha = (A, X) \in \rightarrow$, we say that the source of α is $src(\alpha) = A$ and its target is $trg(\alpha) = X$.

Because of this notion of attacking an attack relation, we have to extend our notion of acceptability of arguments to also include attack relations. Indeed, in the same way that arguments could be reinstated by being defended from their attackers, attacks may or may not be accepted in a final extension. We will briefly examine the case of the complete semantics, starting by the notion of defeat.

Definition 2.3.2. Defeat in AFRAs

Let $F = \langle \mathcal{A}, \rightarrow \rangle$ be an AFRA and $\varphi \in \rightarrow, \psi \in (\mathcal{A} \cup \rightarrow)$. We say that φ *defeats* ψ if and only if either $\psi = trg(\varphi)$ or $src(\psi) = trg(\varphi)$.

A conflict free set of arguments and attacks is then simply a set where no two elements defeat each other.

Definition 2.3.3. Conflict-free in AFRAs

Let $F = \langle \mathcal{A}, \rightarrow \rangle$ be an AFRA and $S \subseteq (\mathcal{A} \cup \rightarrow)$. We say that S is *conflict-free* if there do not exist $\varphi, \psi \in S$ such that φ defeats ψ.

We then say that an element of the AFRA is acceptable with respect to a set of such elements if and only if this set manages to defend that element.

Definition 2.3.4. Acceptability in AFRAs
Let $F = \langle \mathcal{A}, \rightarrow \rangle$ be an AFRA, $\varphi \in (\mathcal{A} \cup \rightarrow)$ and $S \subseteq (\mathcal{A} \cup \rightarrow)$. We say that φ is *acceptable with respect to* S if and only if for every $\psi \in \rightarrow$ such that ψ defeats φ, there exists a $\delta \in S$ such that δ defeats ψ.

The notion of admissibility of a set then follows as the property for every element of that set to be acceptable with respect to it as well as the whole set being conflict-free.

Definition 2.3.5. Admissibility in AFRAs
Let $F = \langle \mathcal{A}, \rightarrow \rangle$ be an AFRA and $S \subseteq (\mathcal{A} \cup \rightarrow)$. We say that S is *admissible* if and only if it is conflict-free and for every $\varphi \in S$, φ is acceptable with respect to S.

The complete semantics then follows with a similar definition as in classical abstract argumentation but using the adapted notions just defined.

Definition 2.3.6. Complete semantics in AFRAs
Let $F = \langle \mathcal{A}, \rightarrow \rangle$ be an AFRA and $S \subseteq (\mathcal{A} \cup \rightarrow)$. We say that S is a *complete extension* of F if and only if S is admissible and for every $\varphi \in (\mathcal{A} \cup \rightarrow)$ such that φ is acceptable with respect to S, $\varphi \in S$.

We can also define another way to obtain semantics for this kind of framework by flattening it, i.e. reducing it to an equivalent classic abstract argumentation framework with no higher order attacks. This flattening procedure has been introduced in [2] and flattens second order attack frameworks, a subset of AFRAs where an argument may only attack an attack relation α if $trg(\alpha) \in \mathcal{A}$.

Definition 2.3.7. Second order AFRA:
A *second order AFRA* is an AFRA $\langle \mathcal{A}, \rightarrow \rangle$ where:

$$\text{for all } \alpha \in \rightarrow, \text{ if } trg(\alpha) = \psi \text{ for some } \psi \in \rightarrow, \text{ then } trg(\psi) \in \mathcal{A}.$$

In the rest of this section we will focus on second order AFRAs, but later on, in Section 5.1, we will introduce a flattening for AFRAs of any order, but with a small restriction which excludes AFRAs containing problematic cycles.

In the flattening, all arguments and higher order attacks are represented by *meta-arguments* which represent their semantic meaning. An argument A is flattened into the argument $accept(A)$, while attacks are flattened into a chain of arguments. For example, the AFRA in Figure 2.7 would be flattened into the meta-argumentation framework in Figure 2.8.

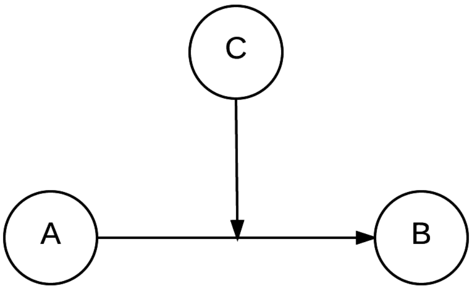

Figure 2.7: AFRA1: Simple example of a higher order attack in an AFRA

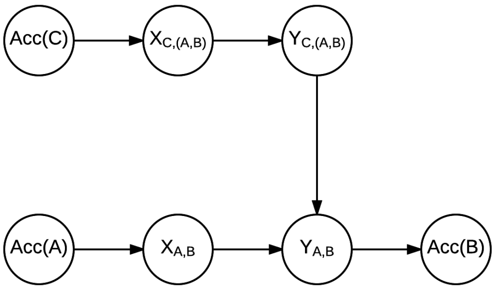

Figure 2.8: AFRA1 flattened into meta-arguments

As mentioned earlier, the arguments are flattened into meta-arguments which represent their acceptability. Each attack relation (A, B) is flattened into two arguments $X_{A,B}$ and $Y_{A,B}$ such that $acc(A)$ attacks $acc(B)$ through them. $Y_{A,B}$ directly attacks $acc(B)$ and is itself attacked by $X_{A,B}$. In turn, $acc(A)$ defends $Y_{A,B}$ by attacking $X_{A,B}$. An attack which targets this attack relation in the AFRA will be flattened into a chain of meta-arguments which will target the meta-argument $Y_{a,b}$ of the flattened target. In our example, C was attacking the attack form A to B in the AFRA. In the flattening, $Y_{C,(A,B)}$ now attacks $Y_{A,B}$. Notice that if $acc(C)$ is deemed acceptable, then $Y_{C,(A,B)}$ will be deemed acceptable too since it is defended from its only attacker $X_{C,(A,B)}$. Hence, $Y_{A,B}$ is now being targeted by an attack it cannot defend itself from and thus will not be deemed acceptable. This will reinstate $acc(B)$ which is being defended from its only attacker. Notice that in this case, $acc(A)$ will also be deemed acceptable. This follows our intuition from the AFRA as only the attack relation originating from A was being attacked and hence A remains unattacked. Formally, we have:

Definition 2.3.8. Flattening of second order AFRAs

Let $F = \langle \mathcal{A}, \rightarrow \rangle$ be a second order AFRA. The set of meta-arguments corresponding to F is $MA = \{acc(a) \mid a \in \mathcal{A}\} \cup \{X_{a,b}, Y_{a,b} \mid a, b \in \mathcal{A}\}$.

We define the *flattening function* f to be $f(F) = \langle MA, \rightarrow_2 \rangle$, where $\rightarrow_2 \subseteq MA \times MA$ is a binary relation on MA such that

- $acc(a) \rightarrow_2 X_{a,b}, X_{a,b} \rightarrow_2 Y_{a,b}, Y_{a,b} \rightarrow_2 acc(b)$ for all $a, b \in \mathcal{A}$ such that $a \rightarrow b$
- $acc(a) \rightarrow_2 Y_{b,c}$ for all $a, b, c \in \mathcal{A}$ such that $a \rightarrow (b, c)$

For a given semantics, we can now get the extensions of an AFRA by flattening it, retrieving the extensions according to the semantics for classical abstract argumentation frameworks and then translating the meta-arguments from the extension back into arguments from the AFRA. The arguments in an extension of the AFRA are the ones for which there is a meta-argument for their acceptability in the corresponding extension of the flattened AFRA. The "auxiliary arguments" of the form $X_{a,b}$ do not need to appear in the unflattened extension since they do not correspond to any element of the AFRA.

Let us consider our example of AFRA1 again. In the flattened framework, the only complete extension is $\{Acc(A), Acc(B), Acc(C), Y_{C,(A,B)}\}$. $Y_{C,(A,B)}$ Represents the attack from C to (A, B). The other three meta-arguments represent the argument A, B and C, so this gives us that the only complete extension of AFRA1 is $\{A, B, C, (C, (A, B))\}$.

2.4 Support in Abstract Argumentation

While classical abstract argumentation revolves around attacks, there has been research on extending it with a positive relation of support between arguments. We will first examine the formalism introduced by Cayrol and Lagasquie-Schiex called bipolar argumentation framework [4], as summarized by G. Boella et al. in [3].

Definition 2.4.1. Bipolar Argumentation Framework (BAF):
A *bipolar argumentation framework* is a triple $\langle \mathcal{A}, \rightarrow, \Rightarrow \rangle$ where $\mathcal{A}$ is a set of arguments, $\rightarrow \subseteq \mathcal{A} \times \mathcal{A}$ is an attack relation and $\Rightarrow \subseteq \mathcal{A} \times \mathcal{A}$ is a support relation.

Caylor and Lagasquie [4] attempt to give meaning to the support relation in their framework by translating each bipolar argumentation framework into a meta-argumentation framework. These meta-argumentation frameworks consist of meta-arguments and an attack relation between them, similarly to a classical abstract argumentation framework. The meta-arguments correspond to a set of arguments from the bipolar argumentation framework. The idea is that a meta-argument makes sense if the arguments it corresponds to form a chain of support. These meta-arguments are called elementary coalitions and are defined as follows:

Definition 2.4.2. Elementary coalition:
Let $\langle \mathcal{A}, \rightarrow, \Rightarrow \rangle$ be a bipolar argumentation framework. An *elementary coalition* is a set $EC = \{a_1, ..., a_n\} \subseteq \mathcal{A}$ such that:

1. there exists a permutation $\{i_1, ..., i_n\}$ of $\{1, ..., n\}$ such that there is a sequence of supports $a_{i_1} \Rightarrow ... \Rightarrow a_{i_n}$;

2. EC is conflict-free as defined in Definition 2.1.3;

3. EC is maximal with respect to $\subseteq$ among the subsets of $\mathcal{A}$ satisfying (1) and (2).

They call $EC(BAF)$ the set of all elementary coalitions in BAF and $ECAF = \langle EC(BAF), c - attacks \rangle$ the elementary coalition argumentation framework associated with the bipolar argumentation framework BAF. The conflict relation $c - attacks$ is defined as follows:

Definition 2.4.3. c-attacks relation
Let $BAF = \langle \mathcal{A}, \rightarrow, \Rightarrow \rangle$ be a bipolar argumentation framework and $EC_1, EC_2 \in EC(BAF)$ be two elementary coalitions of BAF. We have that EC_1 *c-attacks* EC_2 if and only if there exists arguments $a_1 \in EC_1$ and $a_2 \in EC_2$ such that $a_1 \rightarrow a_2$.

The acceptability semantics for BAFs are then defined in terms of the elementary coalitions in the framework.

Definition 2.4.4. Acceptability semantics
Let $BAF = \langle \mathcal{A}, \rightarrow, \Rightarrow \rangle$ be a bipolar argumentation framework and $S \subseteq \mathcal{A}$. We define that:

- S is a *ecp-extension* of BAF if and only if there exists a preferred extension $\{EC_1, ..., EC_n\}$ of ECAF as defined in Definition 2.1.7 such that $S = EC_1 \cup ... \cup EC_n$.

- S is a *ecs-extension* of BAF if and only if there exists a stable extension $\{EC_1, ..., EC_n\}$ of ECAF as defined in Definition 2.1.8 such that $S = EC_1 \cup ... \cup EC_n$.

- S is a *ecg-extension* of BAF if and only if there exists a grounded extension $\{EC_1, ..., EC_n\}$ of ECAF as defined in Definition 2.1.6 such that $S = EC_1 \cup ... \cup EC_n$.

More such extensions can be defined to match the other semantics of classical abstract argumentation frameworks. Note however that it is not possible to produce an equivalent definition of admissible sets as defined in Definition 2.1.4. Cayrol and Lagasquie [4] argue that this drawback is not exactly problematic as they consider the collective attacks emerging from the coalitions of arguments rather than the individual attacks between single arguments. Boella et al. [3] argue that using meta-argumentation should preserve all of Dung's properties and principles and propose to consider support in a deductive sense by introducing mediated attacks. The intuition behind these attacks is that if from a we can deduce b, then if we do not have b, we also cannot have a. Thus, any attack on b will result in an attack on a.

Definition 2.4.5. Mediated attacks:

Let $BAF = \langle \mathcal{A}, \rightarrow, \Rightarrow \rangle$ be a bipolar argumentation framework. For $a, b \in \mathcal{A}$, there is a *mediated attack* from a to b if and only if there is a sequence $a_1 \Rightarrow a_2, ..., a_{n-1} \Rightarrow a_n$ such that $n \geq 2$, $a = a_1$ and $b \rightarrow a_n$.

They then define the semantics of bipolar argumentation frameworks with respect to their flattening in [3]. Similarly as for AFRAs, the flattened framework will consist of meta-arguments and an attack relation, with the support relation present in the BAFs being flattened as a combination of auxiliary meta-arguments and attack relations. The idea behind the flattening of the support relation is that supposing a supports b, we have that b is a deductive consequence of a. Hence, if b is not accepted then neither is a. The flattening will introduce an auxiliary argument $Z_{a,b}$ which is attacked by b only. $Z_{a,b}$ in turn attacks a. Hence, b defends a from $Z_{a,b}$. However, as the only attacker of $Z_{a,b}$, if b not accepted then $Z_{a,b}$ will be accepted and hence a will not be accepted.

Definition 2.4.6. BAF flattening

Given a bipolar argumentation framework $BAF = \langle \mathcal{A}, \rightarrow, \Rightarrow \rangle$, the set of corresponding meta-arguments MA is $\{acc(a) \mid a \in \mathcal{A}\} \cup \{X_{a,b}, Y_{a,b} \mid a, b \in \mathcal{A}\} \cup \{Z_{a,b} \mid a, b \in \mathcal{A}\}$ and $\rightarrow_2 \subseteq MA \times MA$ is a binary relation on MA such that:

- For all $a, b \in \mathcal{A}$ such that $a \rightarrow b$, we have $acc(a) \rightarrow_2 X_{a,b}$, $X_{a,b} \rightarrow_2 Y_{a,b}$ and $Y_{a,b} \rightarrow_2 acc(b)$

- For all $a, b \in \mathcal{A}$ such that $a \Rightarrow b$, we have $acc(b) \rightarrow_2 Z_{a,b}$ and $Z_{a,b} \rightarrow_2 acc(a)$

Notice that in their definition of the flattening, Boella et al. [3] have included a flattening for second order attacks, as later in their paper, they study a framework which combines support and second order attacks. Later in the thesis, we will also present a framework with features including support and higher order attacks.

The example represented in Figure 2.9 is flattened in Figure 2.10:

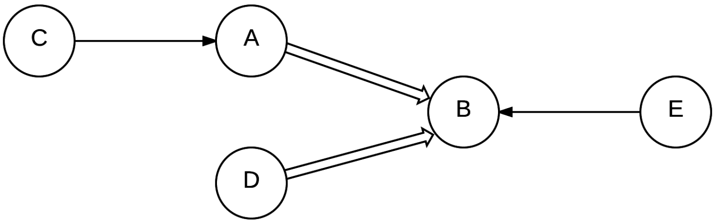

Figure 2.9: Argumentation framework including a support relation representing the wet grass example. The dotted line with a white arrow represents the support relation.

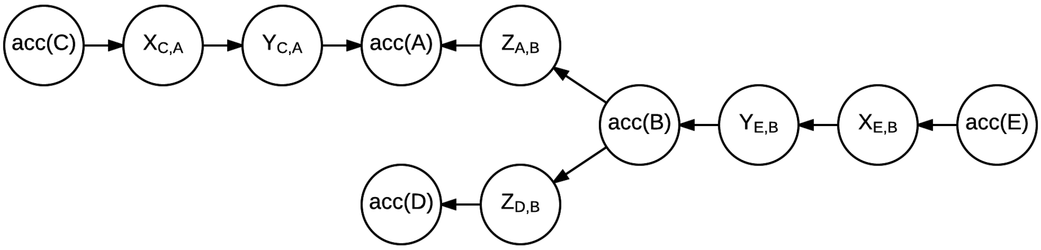

Figure 2.10: Flattened BAF from Figure 2.9

In the flattening, the mediated attacks are made apparent. By applying the semantics of classical abstract argumentation frameworks we can then retrieve the corresponding extensions of the BAF.

2.5 Joint attacks

The idea of coalitions from BAFs [4] can also be represented in a different way. In Cayrol and Lagasquie's formalism, coalitions are sets of arguments which support each other and attack arguments and coalitions which are targeted by at least one of the attacking coalition's argument. Another way to view coalitions would be as sets of arguments which join together in order to make an attack on some other argument possible.

Suppose for example that there has been a murder and the prime suspect is a man named John. Until proven guilty, John is however considered innocent and thus we have an argument A which states that John is innocent. Now the evidence shows a gun was found on the crime scene with John's fingerprints. This would be our second argument B. Note that this alone is not enough of an argument to warrant an attack on A.

After further investigation, it happens that this gun was the murder weapon. This would be our third argument, C. Again, C alone cannot attack A. However, by combining their information, we can conclude that John's fingerprints were found on the murder weapon. This is clearly an attack on his innocence. Hence, by joining forces, the two arguments B and C could form an attack on A. This means that we require both B and C to be accepted in order to reject A, since alone they are not enough to warrant an attack on A. The framework is represented below.

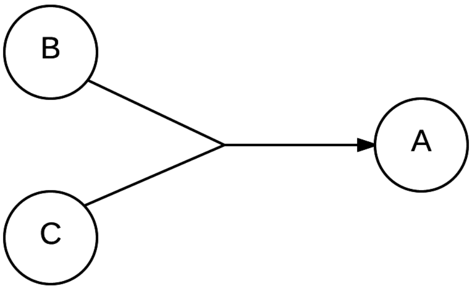

Figure 2.11: Higher level network corresponding to our example

D. Gabbay [8] calls this kind of relation a *joint attack*. He defines it as follows:

Definition 2.5.1. Higher level networks:
A *higher level argumentation framework* is a triple $(S, S^0, \rightarrow)$, where $S \neq \emptyset$ is a set of arguments, S^0 is the family of all finite non-empty subsets of S and $\rightarrow \subseteq S^0 \times S$ is an attack relation.

For simplicity of notation we will identify the singleton set $\{x\}$ with x.

Similarly as before, the semantics of higher level networks will be defined in terms of their flattening. We define the flattening as follows:

Definition 2.5.2. Higher level networks flattening:
Given a higher level network $(S, S^0, \rightarrow)$, the set of corresponding meta-arguments MA is $\{acc(a), rej(a) \mid a \in \mathcal{A}\} \cup \{e(X) \mid X \in S^0\}$ and $\rightarrow_2 \subseteq MA \times MA$ is a binary relation on MA such that:

- For all $a \in \mathcal{A}$, we have $acc(a) \rightarrow_2 rej(a)$
- For all $X \in S^0$, and every $b \in \mathcal{A}$ such that $X \rightarrow b$, we have that $e(X) \rightarrow_2 acc(b)$ and $rej(a) \rightarrow_2 e(X)$ for every $a \in X$.

The idea behind the flattening is that each argument a will have a meta-argument for its acceptance $acc(a)$ as well as a complementary meta-argument for its rejection $rej(a)$. By construction, $acc(a)$ is the only argument attacking $rej(a)$. Therefore, if $acc(a)$ is part of a complete extension, then $rej(a)$ will not be part of it, and if an argument from that complete extension attacks $acc(a)$, it will thus defend $rej(a)$ which will then be part of the complete extension.

Then, for each joint attack originating from a set of arguments X, there is a meta-argument $e(X)$ representing this coalition which attacks the meta-argument corresponding to the argument that X was attacking. This meta-argument $e(X)$ is then being attacked by the complementary meta-arguments $rej(a)$ for each argument $a \in X$. This way, if even one argument from the coalition is successfully attacked, then its complementary meta-argument will be accepted and will thus successfully defeat $e(X)$. Hence, the joint attack will succeed if and only if every single argument from the coalition is accepted.

The flattening of the framework from Figure 2.11 is as follows:

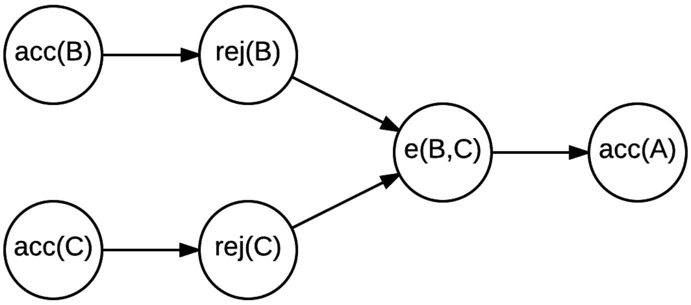

Figure 2.12: Flattened version of the framework from Figure 2.11

2.6 Structured argumentation

When designing an argumentation framework, two of the important design decisions which have to be made are the following: how can arguments be built and how can they be attacked? S. Modgil and H. Prakken proposed a system called the ASPIC+ framework [10] which attempts to ease the designing of argumentation models by answering those questions among others. There are two main ideas on which the ASPIC+ framework is based. The first idea is that conflicts are usually resolved with explicit preferences. The second idea is that arguments are built using either strict or defeasible inference rules. While strict rules guarantee the inference of a certain conclusion from given premises, defeasible rules only present a presumption in favor of their conclusion. The goal of the ASPIC+ framework is to provide a systematic way of constructing arguments and attacks from a knowledge base and a set of inference rules. These rules and knowledge base are more intuitive to mine from a text and easier to motivate.

In order to use the ASPIC+ system, one needs to provide some information. The first element is a logical language closed under negation. Then one has to provide two sets of (possibly empty) strict and defeasible inference rules. Additionally, one must provide a partial naming function which maps some of the defeasible rules to a formula from the chosen logical language. The collection of this information is called an argumentation system.

Definition 2.6.1. Argumentation systems:

An *argumentation system* is a triple $AS = (\mathcal{L}, \mathcal{R}, n)$ where:

- $\mathcal{L}$ is a logical language closed under negation ($\neg$).
- $\mathcal{R} = \mathcal{R}_s \cup \mathcal{R}_d$ is a set of strict ($\mathcal{R}_s$) and defeasible ($\mathcal{R}_d$) inference rules of the form $\phi_1, ..., \phi_n \rightarrow \phi$ and $\phi_1, ..., \phi_n \Rightarrow \phi$ respectively, where ϕ_i and ϕ are well-formed formulas in $\mathcal{L}$ and $\mathcal{R}_s \cap \mathcal{R}_d = \emptyset$.
- n is a partial function such that $n : \mathcal{R}_d \longrightarrow \mathcal{L}$.

We define a function $-$ such that $-\phi = \psi$ if $\phi = \neg\psi$, otherwise $-\phi = \neg\phi$.

The intuition is that the rules in $\mathcal{R}$ are on the meta-level compared to the language $\mathcal{L}$ and allow one to conclude the head of a rule if given the antecedents. The strict rules are rules of inference which are considered to hold in all cases. Hence, if one accepts its antecedents, then one must also accept its conclusion. Defeasible rules on the other hand are ones which are known to be generally true but which might fail in some cases and hence their inferences and conclusions are possible subjects of attacks.

For the rules to be of some use, one needs to define a set of premises which will serve as a knowledge base from which one can start building arguments by applying the rules.

Definition 2.6.2. Knowledge bases:
A *knowledge base* in an $AS = (\mathcal{L}, \mathcal{R}, n)$ is a set of $\mathcal{K} \subseteq \mathcal{L}$ consisting of two disjoint subsets $\mathcal{K}_n$ (the axioms) and $\mathcal{K}_p$ (the ordinary premises).

The axioms are formulas of which the truth value is indisputable, while the ordinary premises are formulas which can be used to make further inferences but might turn out to be defeated in the end. By joining these with an appropriate argumentation system, one gets an argumentation theory.

Definition 2.6.3. Argumentation theories:
An *argumentation theory* is a tuple $AT = (AS, \mathcal{K})$ where AS is an argumentation system and $\mathcal{K}$ is a knowledge in AS.

An argumentation theory now contains all the elements needed for building the arguments. ASPIC+ provides a few ways to construct these from the theory. An argument in ASPIC+ has a few properties which are given by the following functions:

- Prem returns the set of all ordinary premises of the argument.
- Conc returns the conclusion of the argument.
- Sub returns all its sub-arguments.
- $\mathsf{DefRules}$ returns the set of defeasible rules used in the argument.
- $\mathsf{TopRule}$ returns the last inference rule used, if applicable.

We now have three different ways to build an argument. Either it introduces one of the ordinary premises from the knowledge, or it makes an inference from a strict or defeasible rule.

Definition 2.6.4. Arguments:
An *argument* A on the basis of an argumentation theory with a knowledge base $\mathcal{K}$ and an argumentation system $(\mathcal{L}, \mathcal{R}, n)$ has one of the following forms:

1. φ, where $\varphi \in \mathcal{K}$ with:
   $\mathsf{Prem}(A) = \{\varphi\}$,
   $\mathsf{Conc}(A) = \varphi$,
   $\mathsf{Sub}(A) = \{\varphi\}$,,
   $\mathsf{DefRules}(A) = \emptyset$,,
   $\mathsf{TopRule}(A)$ is undefined.
2. $A_1, ..., A_n \rightarrow \varphi$, where $A_1, ..., A_n$ are arguments such that $\mathsf{Conc}(A_1), ..., \mathsf{Conc}(A_n) \rightarrow \varphi \in \mathcal{R}_s$ with:
   $\mathsf{Prem}(A) = \mathsf{Prem}(A_1) \cup ... \cup \mathsf{Prem}(A_n)$,
   $\mathsf{Conc}(A) = \varphi$,
   $\mathsf{Sub}(A) = \mathsf{Sub}(A_1) \cup ... \cup \mathsf{Sub}(A_n) \cup \{A\}$,
   $\mathsf{DefRules}(A) = \mathsf{DefRules}(A_1) \cup ... \cup \mathsf{DefRules}(A_n)$,
   $\mathsf{TopRule}(A) = \mathsf{Conc}(A_1), ..., \mathsf{Conc}(A_n) \rightarrow \varphi$.

3. $A_1, ..., A_n \Rightarrow \varphi$, where $A_1, ..., A_n$ are arguments such that $\mathsf{Conc}(A_1), ..., \mathsf{Conc}(A_n) \Rightarrow \varphi \in \mathcal{R}_d$ with:
   $\mathsf{Prem}(A) = \mathsf{Prem}(A_1) \cup ... \cup \mathsf{Prem}(A_n)$,
   $\mathsf{Conc}(A) = \varphi$,
   $\mathsf{Sub}(A) = \mathsf{Sub}(A_1) \cup ... \cup \mathsf{Sub}(A_n) \cup \{A\}$,
   $\mathsf{DefRules}(A) = \mathsf{DefRules}(A_1) \cup ... \cup \mathsf{DefRules}(A_n) \cup \{\mathsf{Conc}(A_1), ..., \mathsf{Conc}(A_n) \Rightarrow \varphi\}$,
   $\mathsf{TopRule}(A) = \mathsf{Conc}(A_1), ..., \mathsf{Conc}(A_n) \Rightarrow \varphi$.

Example 2.6.1. Consider a knowledge base in an argumentation system with language $\mathcal{L}$ consisting of p, q, r, s, t, d_1, d_2 and their negations, with $\mathcal{R}_d = \{d_1, d_2\}$ and $\mathcal{R}_s = \{s_1, s_2\}$, where the rules are defined as:

- d_1: $r \Rightarrow \neg q$
- d_2: $t \Rightarrow \neg p$
- s_1: $p \rightarrow q$
- s_2: $s \rightarrow \neg d_2$

Also, the knowledge base is formed by $\mathcal{K}_n = \{r, s\}$ and $\mathcal{K}_p = \{p, t\}$. Notice that we have defined rules by writing them in the form $n(r) : r$.

Two of the arguments we can construct are $A_1 = p$ and $A_2 = A_1 \rightarrow q$, where $\mathsf{Prem}(A_2) = \{p\}$, $\mathsf{Conc}(A_2) = q$, $\mathsf{Sub}(A_2) = \{A_1, A_2\}$, $\mathsf{DefRules}(A_2) = \emptyset$, $\mathsf{TopRule}(A_2) = s_1$.

Now that we have defined a way to construct the arguments from the knowledge base and inference rules, we can define how to build the other component of an abstract argumentation framework, namely the attacks. There are 3 ways for an argument to attack another one. It must attack it either on one of its premises, on the inference rule used or on the conclusion.

Definition 2.6.5. Attacks:

An argument A *attacks* an argument B if and only if A *undermines, undercuts* or *rebuts* B, where:

- A *undermines* B on φ if and only if $\mathsf{Conc}(A) = -\varphi$ for an ordinary premise $\varphi \in \mathsf{Prem}(B)$.
- A *undercuts* B (on B') if and only if $\mathsf{Conc}(A) = -n(r)$ for some $B' \in \mathsf{Sub}(B)$ such that $\mathsf{TopRule}(B') = r$.
- A *rebuts* B (on B') if and only if $\mathsf{Conc}(A) = -\varphi$ for some $B' \in \mathsf{Sub}(B)$ of the form $B''_1, ..., B''_n \Rightarrow \varphi$.

Example 2.6.2. In our previous example, we can also construct the arguments $B_1 = r$, $B_2 = B_1 \Rightarrow \neg q$, $C_1 = t$, $C_2 = C_1 \Rightarrow \neg p$, $D_1 = s$ and $D_2 = D_1 \rightarrow \neg d_2$. We then have that C_2 undermines A_1 and A_2 on p, A_2 rebuts B_2 and D_2 undercuts C_2.

Notice that rebuttal is usually symmetric, however this kind of duality might be resolved by having a preference over the arguments. This way, an argument may only attack another one if it is at least as preferred as the attacked one. Given a preference relation, we can then define what it means for an attack be successful, and in general we will say that an argument A *defeats* an argument B if the attack is successful.

Definition 2.6.6. Successful undermining, rebuttal and defeat: Given a preference relation $\preceq$ over the arguments, we say that:

- A *successfully undermines* B if and only if A undermines B on φ and $\varphi \preceq A$.
- A *successfully rebuts* B if and only if A rebuts B on B' and $B' \preceq A$.
- A *defeats* B if and only if it undercuts, successfully undermines or successfully rebuts B.

Notice that all undercuttings are considered as successful irrespective of preference as no such criteria is required for that kind of attack.

We can then define the procedure to generate an abstract argumentation framework from an argumentation theory and a preference relation.

Definition 2.6.7. Abstract argumentation frameworks:

An *abstract argumentation framework* (AF) *corresponding to* an argumentation theory $AT = (AS, \mathcal{K})$ and a preference relation over arguments $\preceq$ is a pair $(\mathcal{A}, D)$, such that:

- $\mathcal{A}$ is the smallest set of all finite arguments constructed from $\mathcal{K}$ in AS satisfying Definition 2.6.4;
- $(X, Y) \in D$ if and only if X defeats Y with respect to $\preceq$.

The preference relation is easier to motivate and understand if it is first defined on the set of defeasible rules and premises. We can then lift the preference relation from rules to arguments in one of several ways. One is the weakest-link principle, another is the last-link principle. In the weakest-link principle we compare two arguments A and B by comparing the least preferred rules in $\mathsf{DefRules}(A)$ and $\mathsf{DefRules}(B)$. If the least preferred rule in $\mathsf{DefRules}(A)$ is at least as preferred as the weakest rule in $\mathsf{DefRules}(B)$, then we say that A is at least as preferred to B. Formally, we get:

Definition 2.6.8. Weakest-link preference:

Let A and B be two arguments. We have that $A \preceq_w B$ if and only if:

1. If $\mathsf{DefRules}(A) = \mathsf{DefRules}(B) = \emptyset$, then there exists $p_a \in \mathsf{Prem}(A)$, such that for all $p_b \in \mathsf{Prem}(B)$, we have $p_a \leq p_b$, else;
2. If $\mathsf{Prem}(A) = \mathsf{Prem}(B) = \emptyset$, then there exists $r_a \in \mathsf{DefRules}(A)$, such that for all $r_b \in \mathsf{DefRules}(B)$, we have $r_a \leq r_b$, else;
3. There exists $r_a \in \mathsf{DefRules}(A)$ and $p_a \in \mathsf{Prem}(A)$, such that for all $r_b \in \mathsf{DefRules}(B)$ and $p_b \in \mathsf{Prem}(B)$, we have $r_a \leq r_b$ and $p_a \leq p_b$

We define a notion of strict preference $\prec_w$ by replacing $\leq$ with $<$ in the above definition.

The other way to lift a preference relation over rules to one over arguments is by using the last link principle. According to this principle, we compare the last defeasible rules used in the argument, which corresponds to the value given by applying the function $\mathsf{LastDefRules}$ to the argument. We define this function as follows:

Definition 2.6.9. Last defeasible rules:

Let A be an argument. We define the function $\mathsf{LastDefRules}$ as follows:

- If $\mathsf{DefRules}(\mathrm{A}) = \emptyset$, then $\mathsf{LastDefRules}(\mathrm{A}) = \emptyset$, else;
- If $A = A_1, ..., A_n \Rightarrow \varphi$, then $\mathsf{LastDefRules}(A) = \{\mathsf{Conc}(A_1), ..., \mathsf{Conc}(A_n)\}$, else;

- If $A = A_1, ..., A_n \rightarrow \varphi$, then $\mathsf{LastDefRules}(A) = \{\mathsf{LastDefRules}(A_1), ..., \mathsf{LastDefRules}(A_n)\}$

We then define the lifting of the preference from rules to arguments according to last link principle as:

Definition 2.6.10. Last link preference:
Let A and B be two arguments. We have that $A \preceq_l B$ if and only if:

- If $\mathsf{LastDefRules}(A) = \mathsf{LastDefRules}(B) = \emptyset$, then there exists $p_a \in \mathsf{Prem}(A)$ such that for all $p_b \in \mathsf{Prem}(B)$, we have $p_a \leq p_b$, else;
- There exists $r_a \in \mathsf{LastDefRules}(A)$ such that for all $r_b \in \mathsf{LastDefRules}(B)$, w have $r_a \leq r_b$.

Again, we define the strict preference relation $\prec_l$ by replacing $\leq$ with $<$ in the above definition.

Chapter 3

ASPIC-END: a framework for explanatory argumentation using natural deduction

In this chapter, we will introduce ASPIC-END, an extension of ASPIC+ which features explanations and natural-deduction-style arguments. We add explanations as we will later focus on arguments about the proposed solutions to the liar paradox, it is thus essential to be able to model how the solutions explain the paradox. We also allow argument to be built using natural deduction as arguments in the context of logical paradoxes often involve natural deduction and this thus allows for better representation of such arguments. The last notable modification we make is that we replace strict rules by intuitively strict rules which can sometimes be attacked, but are still stronger than defeasible rules.

3.1 Natural deduction arguments and explananda

In order to use the system ASPIC-END, one has to select a language $\mathcal{L}$ and a set of rules $\mathcal{R}$. $\mathcal{L}$ must be a logical language closed under negation. $\mathcal{R}$ (possibly empty) is made of two subsets of rules, intuitively strict rules and defeasible rules, such that each rule belongs to exactly one of these subsets. We need a way to assign a formula from $\mathcal{L}$ to some of the rules, so that these rules can later be attacked by an argument claiming its negation. One hence also has to specify a partial function n (for "naming") from $\mathcal{R}$ to $\mathcal{L}$. The intuition is that for some rules $r \in \mathcal{R}$, $n(r)$ is a well formed formula in $\mathcal{L}$ which says that r is applicable, and so an argument claiming $\neg n(r)$ attacks the inference step in r.

Formally, we have:

Definition 3.1.1. Argumentation Theories:
An *argumentation theory* is a triple $AT = (\mathcal{L}, \mathcal{R}, n)$, where:

- $\mathcal{L}$ is a logical language closed under negation ($\neg$).
- $\mathcal{R} = \mathcal{R}_{is} \cup \mathcal{R}_d$ 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 φ_i, φ are meta-variables ranging over well-formed formulae in $\mathcal{L}$ and intuitively strict rules may have no antecedents), and $\mathcal{R}_{is} \cap \mathcal{R}_d = \emptyset$.
- n is a partial function such that n: $\mathcal{R} \mapsto \mathcal{L}$, and for every rule $\mathcal{R}$ of the form $\rightsquigarrow \varphi$, we have $n(\rightsquigarrow \varphi) = \varphi$.

We define a negation function $-\colon \mathcal{L} \mapsto \mathcal{L}$ such that $-\varphi = \psi$ if $\varphi = \neg\psi$ and $-\varphi = \neg\varphi$ otherwise.

Also, the language $\mathcal{L}$ may contain some formulas of the form $\neg Assume(\varphi)$, where $\varphi \in \mathcal{L}$ does not contain the $Assume$ operator. These formulas will represent the fact that it does not make sense to assume φ and will be used later on as a point of attack for proofs by contradictions.

In our study of the liar paradox, we come to question the inference of rules which seemed strict to begin with. Hence, we introduce intuitively strict rules to replace strict rules in our system. For example, it seems fair to say that if a sentence is not true, then it must be false. However, Kripke's solution to the liar paradox 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. Here, the 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. However, these intuitively strict rules are still different from defeasible rules in the sense that if one accepts its application, then one cannot question its conclusion. Therefore, intuitively strict rules are not defeasible in the sense that they cannot be rebutted, only undercut.

With this idea in mind, it seems unjustified to allow for some rules to be completely strict, as it would be hard to justify being able to question the inference of some rules but not others. This also allows for new solutions, if properly motivated, to be modeled in our system.

Recall that in ASPIC+ we have premises as well as axioms. The difference between a premise and an axiom lies in the fact that an argument using a premise φ can be undermined on φ, while an argument using an axiom ψ cannot be undermined on ψ. Just as we do not want to have rules which cannot be undercut, we also do not wish to have axioms which cannot be undermined. However, this being the only difference between an axiom and a premise, we do not have a reason to distinguish them anymore, and keep only premises.

Also, note that contrary to ASPIC+, we do not have a separate set of premises representing the knowledge base. Instead, premises are represented as intuitively strict rules with no antecedent and are hence included in $\mathcal{R}$. We thus require that for each rule with no antecedent $(\rightsquigarrow \varphi) \in \mathcal{R}_d$, we have that $n(\rightsquigarrow \varphi) = \varphi$. By representing premises as intuitively strict rules instead of defeasible rules, we prevent rebuttal attacks from targeting arguments which introduce premises, and make a clear distinction between rebuttal and undermining attacks. We have found however that undermining functions closely to undercutting and hence make undermining a special case of undercutting. In this way, we view an undermining attack as attacking the rule which introduces the premise. Note that having premises as a separate entity in our system would not add to its expressive power.

Now that we have defined what an argumentation theory is, we must now provide the ways of constructing arguments. One should be able to build arguments from ordinary premises, using chain applications of rules of inference from the theory. For every argument, we defined five functions which allow us to retrieve some of their specific properties. Conc returns the conclusion of the argument. Sub returns the set of its sub-arguments. Rules returns the set of all rules which have been used in its construction. $\mathsf{TopRule}$ returns the last inference rule which has been used in the argument, and is undefined in the case of arguments without such a rule, i.e. assumption introduction and proof by contradiction arguments. As returns the set of assumptions under which the argument is working. This will allow us to construct arguments in a natural deduction manner, as we will see in an example later. Notice that do not require the Prem function which was present in ASPIC+, as the premises in ASPIC-END are counted as rules and hence part of the elements retrieved by the Rules function. Also, the ProofByContradiction construct is based on natural deduction, namely the ¬-introduction rule from natural-deduction.

Definition 3.1.2. Arguments:

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

1. $A_1, \dots A_n \rightsquigarrow \psi$ where $A_1, \dots A_n$ are arguments such that there exists an intuitively strict rule $\mathsf{Conc}(A_1), \dots, \mathsf{Conc}(A_n) \rightsquigarrow \psi$ in $\mathcal{R}_{is}$.
   $\mathsf{Conc}(A) = \psi$,
   $\mathsf{Sub}(A) = \mathsf{Sub}(A_1) \cup \dots \cup \mathsf{Sub}(A_n) \cup \{A\}$,
   $\mathsf{Rules}(A) = \mathsf{Rules}(A_1) \cup \dots \cup \mathsf{Rules}(A_n) \cup \{\mathsf{Conc}(A_1), \dots, \mathsf{Conc}(A_n) \rightsquigarrow \psi\}$,
   $\mathsf{TopRule}(A) = \mathsf{Conc}(A_1), \dots, \mathsf{Conc}(A_n) \rightsquigarrow \psi$,
   $\mathsf{As}(A) = \mathsf{As}(A_1) \cup \dots \cup \mathsf{As}(A_n)$.

2. $A_1, \dots A_n \Rightarrow \psi$ where $A_1, \dots A_n$ are arguments such that there exists a defeasible rule $\mathsf{Conc}(A_1), \dots, \mathsf{Conc}(A_n) \Rightarrow \psi$ in $\mathcal{R}_d$ and $\mathsf{As}(A_1) \cup \dots \cup \mathsf{As}(A_n) = \emptyset$.
   $\mathsf{Conc}(A) = \psi$,
   $\mathsf{Sub}(A) = \mathsf{Sub}(A_1) \cup \dots \cup \mathsf{Sub}(A_n) \cup \{A\}$,
   $\mathsf{Rules}(A) = \mathsf{Rules}(A_1) \cup \dots \cup \mathsf{Rules}(A_n) \cup \{\mathsf{Conc}(A_1), \dots, \mathsf{Conc}(A_n) \Rightarrow \psi\}$,
   $\mathsf{TopRule}(A) = \mathsf{Conc}(A_1), \dots, \mathsf{Conc}(A_n) \Rightarrow \psi$,
   $\mathsf{As}(A) = \emptyset$.

3. $\mathsf{Assume}(\varphi)$ where $\varphi \in \mathcal{L}$ with $\mathsf{Conc}(A) = \varphi$, $\mathsf{Sub}(A) = \{\mathsf{Assume}(\varphi)\}$, $\mathsf{Rules}(A) = \emptyset$, $\mathsf{TopRule}(A)$ is undefined and $\mathsf{As}(A) = \{\varphi\}$.

4. $\mathsf{ProofByContrad}(\neg\varphi, A')$ where A' is an argument such that $\varphi \in \mathsf{As}(A')$ and $\mathsf{Conc}(A') = \bot$.
   $\mathsf{Conc}(A) = \neg\varphi$,
   $\mathsf{Sub}(A) = \mathsf{Sub}(A') \cup \{\mathsf{ProofByContrad}(\neg\varphi, A')\}$,
   $\mathsf{Rules}(A) = \mathsf{Rules}(A')$,
   $\mathsf{TopRule}(A)$ is undefined,
   $\mathsf{As}(A) = \mathsf{As}(A') \setminus \{\varphi\}$.

In all examples, we present the rules in the form $n(r) : r$, so that the label in front of the rule represents the result of applying the naming function to the rule. Also, for the sake of notational convenience, we will write $p_i :\rightsquigarrow \varphi$ when introducing rules with no antecedents (recall that formally, $n(\rightsquigarrow \varphi) = \varphi$) and refer to them as p_i (with the appropriate i).

Example 3.1.1. Consider an argumentation theory $AT_1 = (\mathcal{L}, \mathcal{R}, n)$, where $\mathcal{L}$ consists of p, q, r, s, u and their negations, $\mathcal{R}_{is} = \{t_1, t_2, p_1\}$ and $\mathcal{R}_d = \{d_1, d_2\}$, where:

- t_1: $p \rightsquigarrow q$
- t_2: $q \rightsquigarrow \bot$
- p_1: $\rightsquigarrow r$
- d_1: $\neg p, r \Rightarrow s$
- d_2: $u \Rightarrow q$

We can then construct an argument for s as follows:

- A_1: $\mathsf{Assume}(p)$, $\mathsf{As}(A_1) = \{p\}$, $\mathsf{Conc}(A_1) = p$
- A_2: $A_1 \rightsquigarrow q$, $\mathsf{As}(A_2) = \{p\}$, $\mathsf{Conc}(A_2) = q$
- A_3: $A_2 \rightsquigarrow \bot$, $\mathsf{As}(A_3) = \{p\}$, $\mathsf{Conc}(A_3) = \bot$
- A_4: $\mathsf{ProofByContrad}(\neg p, A_3)$, $\mathsf{As}(A_4) = \emptyset$, $\mathsf{Conc}(A_4) = \neg p$
- A_5: $\rightsquigarrow r$, $\mathsf{As}(A_5) = \emptyset$, $\mathsf{Conc}(A_5) = r$
- A_6: $A_4, A_5 \Rightarrow s$, $\mathsf{As}(A_6) = \emptyset$, $\mathsf{Conc}(A_6) = s$

We can see that A_1 introduces the assumption p, and from there the arguments A_2 and A_3 manage to derive a contradiction, which allows the construction of argument A_4 with conclusion $\neg p$ under no assumption. We can then use this together with the premise r to form an argument for s. Note however that we cannot form an argument for $\neg u$ using a proof by contradiction, because to derive an inconsistency from u we would have to use d_2. However, defeasible rules can only be applied under no assumption, hence we would be unable to apply it in the proof by contradiction for $\neg u$.

The aim of our system is to generate graphs belonging to an extension of Dung-Style argumentation frameworks, namely Explanatory Argumentation Frameworks, which we will refer to as EAFs. In this extension, there are not only arguments but also explananda. Hence, we need to define a way to construct them:

Definition 3.1.3. Explananda:
There is an *explanandum* E such that $\mathsf{Source}(E) = A$ if and only if there exists an argument A such that:

$$\mathsf{Conc}(A) = \bot,\ \mathsf{As}(A) = \emptyset \text{ and } \mathsf{Rules}(A) \subseteq \mathcal{R}_{is}.$$

Now this definition might be adapted to the application you are making of ASPIC-END, and in our case was made for the purpose of explaining inconsistencies arising from studying the truth value of liar sentences when using informal reasoning which appears intuitive. For this reason, we require that the source for the explananda uses only intuitively strict rules to reach a contradiction under no assumption. Notice however that we do not want every self-contradicting argument to give rise to an explanandum, only ones which do not use any defeasible rules but only intuitively strict ones. Argumentation theory already gives us a way to handle the conflicting arguments which use defeasible rules, namely by allowing one to attack and reject the conclusion of an argument which uses defeasible rules. Rejecting such an argument does not warrant an explanation, because some of the rules used are defeasible and hence known to not always hold.

Example 3.1.2. Consider an argumentation theory $AT_2 = (\mathcal{L}, \mathcal{R}, n)$ where $\mathcal{L}$ consists of p, q, r and their negations, $\mathcal{R}_{is} = \{t_1, t_2, t_3, p_1\}$ and $\mathcal{R}_d = \emptyset$, where:

- t_1: $p \rightsquigarrow q$
- t_2: $q \rightsquigarrow r$
- t_3: $r \rightsquigarrow \bot$
- p_1: $\rightsquigarrow p$

We can construct an explanandum as follows:

- A_1: $\rightsquigarrow p$, $\mathsf{Conc}(A_1) = p$
- A_2: $A_1 \rightsquigarrow q$, $\mathsf{Conc}(A_2) = q$
- A_3: $A_2 \rightsquigarrow r$, $\mathsf{Conc}(A_3) = r$
- A_4: $A_3 \rightsquigarrow \bot$, $\mathsf{Conc}(A_4) = \bot$
- E: $\mathsf{Source}(E) = A_4$

3.2 Attacks and explanations

We now need to define attack and explanation relations in our framework. In general, we distinguish three kinds of attacks in argumentation: undermining, undercutting and rebuttal. Intuitively, for one argument to undermine another, it needs to conclude the negation of one of its premises. In the case of undercut, it needs to claim that one of the rules the other argument has used is not applicable and so is attacking one of the inferences that the argument is making. In the case of rebuttal, one has to provide an argument for accepting the negation of the proposed conclusion. Hence, rebuttals are often symmetrical. As mentioned earlier however, we do not allow for rebuttal of an intuitively strict rule's conclusion, only for undercutting attacks where an argument would motivate the non-applicability of the rule. In the case of defeasible rules however, the inference steps are weaker and so providing an argument which concludes the opposite should be enough to warrant an attack.

In ASPIC-END, we also allow for an argument A to attack an argument B which makes an assumption φ if A concludes that it makes no sense to assume φ. 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 introducing this assumption and any argument making an inference from this assumption.

Definition 3.2.1. Attacks:
A *attacks* B if and only if A *rebuts*, *undercuts* or *assumption-attacks* B, where:

- A *rebuts* argument B (on B') iff $\mathsf{Conc}(A) = -\varphi$ 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 $\mathsf{Conc}(A) = -n(r)$ for some $B' \in \mathsf{Sub}(B)$ such that $\mathsf{TopRule}(B') = r$, $\mathsf{As}(A) \subseteq (\mathsf{As}(B) \cup \mathsf{As}(B'))$ and there is no $\varphi \in \mathsf{As}(B')$ such that $-\varphi = \mathsf{Conc}(A')$ for some $A' \in \mathsf{Sub}(A)$.
- A *assumption-attacks* B (on B') iff for some $B' \in \mathsf{Sub}(B)$ such that $B' = \text{Assume}(\varphi)$, $\mathsf{Conc}(A) = \neg Assume(\varphi)$ and $\mathsf{As}(A) = \emptyset$.

We require that the attacking argument is making less assumptions than the one it attacks, as to prevent arguments from attacking outside of their scope of assumption. In the case of rebuttal, notice that since the attacked argument is using a defeasible rule, it has an empty set of assumptions, as defeasible rules cannot be applied under any assumption. Hence, the attacking argument must also have an empty set of assumption. Also, notice that we introduce assumptions in the hope of deriving an inconsistency, and so every successful proof by contradiction would rebut every other argument due to *ex* *falso quolibet*, which allows one to derive any formula from $\bot$. This is another reason to allow only arguments with an empty set of assumptions to rebut other arguments.

Regarding undercutting, we have again the requirement that no argument may attack out of its scope of assumption. However, in this case, it may very well be possible that the argument (or sub-argument) whose top-rule the attack is being directed at is working under a non-empty set of assumptions. Notice that we consider the union of the assumption sets of both the super-argument being attacked and the sub-argument introducing the inference attacked. This is due to the fact that we have to consider the argument using the said inference in the context it was introduced as well as in the context of the argument being attacked. For example, an argument B with $\mathsf{As}(B)=\{\varphi\}$ might be undercut on a sub-argument $B' \in \mathsf{Sub}(B)$ with $\mathsf{As}(B')=\{\psi\}$ by an argument A with $\mathsf{As}(A)=\{\varphi, \psi\}$. While B' would not be subjected to the same undercutting attack because A would not stand in the assumption scope of $\{\psi\}$, it is important to consider that when supporting B, B' must be considered in the union of its own as well as B's assumption scopes, which is $\{\varphi, \psi\}$. In this context, A stands and thus is able to undercut B on B'.

Additionally, we have the requirement that the attacking argument does not use the negation of any assumptions made by the attacked argument in any of its inferences, because in the scope of the attacked argument, the attack would not stand. See Example 3.2.2 below for an illustration of why this is needed.

Example 3.2.1. Consider an argumentation theory $AT_3 = (\mathcal{L}, \mathcal{R}, n)$ where $\mathcal{L}$ consists of p, q, r, s, u, v, w and their negations, $\mathcal{R}_{is} = \{t_1, t_2, t_3, t_4, p_1, p_2\}$ and $\mathcal{R}_d = \{d_1, d_2, d_3, d_4\}$, where:

- t_1: $p \leadsto q$
- t_2: $q \leadsto u$
- t_3: $q, u \leadsto \bot$
- t_4: $s \leadsto \neg w$
- p_1: $\leadsto r$
- p_2: $\leadsto v$
- d_1: $r \Rightarrow s$
- d_2: $v \Rightarrow w$
- d_3: $w \Rightarrow \neg r$
- d_4: $\neg r \Rightarrow \neg t_1$

We can then construct the following arguments (when not specified, $\mathsf{As}(A) = \emptyset$ for each argument):

- A_1: $\mathsf{Assume}(p)$, $\mathsf{As}(A_1) = \{p\}$, $\mathsf{Conc}(A_1) = p$
- A_2: $A_1 \leadsto q$, $\mathsf{As}(A_2) = \{p\}$, $\mathsf{Conc}(A_2) = q$
- A_3: $A_2 \leadsto u$, $\mathsf{As}(A_3) = \{p\}$, $\mathsf{Conc}(A_3) = u$
- A_4: $A_2, A_3 \leadsto \bot$, $\mathsf{As}(A_4) = \{p\}$, $\mathsf{Conc}(A_4) = \bot$

- A_5: $\mathsf{ProofByContrad}(\neg p, A_4)$, $\mathsf{Conc}(A_5) = \neg p$
- B_1: $\rightsquigarrow v$, $\mathsf{Conc}(B_1) = v$
- B_2: $B_1 \Rightarrow w$, $\mathsf{Conc}(B_2) = w$
- B_3: $B_2 \Rightarrow \neg r$, $\mathsf{Conc}(B_3) = \neg r$
- B_4: $B_3 \Rightarrow \neg t_1$, $\mathsf{Conc}(B_4) = \neg t_1$
- C_1: $\rightsquigarrow r$, $\mathsf{Conc}(C_1) = r$
- C_2: $C_1 \Rightarrow s$, $\mathsf{Conc}(C_2) = s$
- C_3: $C_2 \rightsquigarrow \neg w$, $\mathsf{Conc}(C_3) = \neg w$

In this example, we have that B_4 undercuts A_2, A_3, A_4 and A_5 on A_2, C_3 rebuts B_2, B_3 and B_4 on B_2, C_1 rebuts B_3 and B_4 on B_3 and B_3 undermines C_1, C_2 and C_3 on C_1.

Example 3.2.2. Consider an argumentation theory $AT_4 = (\mathcal{L}, \mathcal{R}, n)$ where $\mathcal{L}$ consists of p, q, r, s and their negations, $\mathcal{R}_{is} = \{t_1, t_2, p_1\}$ and $\mathcal{R}_d = \{d_1, d_2\}$, where:

- t_1: $s \rightsquigarrow p$
- t_2: $p \rightsquigarrow \bot$
- p_1: $\rightsquigarrow r$
- d_1: $r \Rightarrow \neg s$
- d_2: $\neg s \Rightarrow \neg t_2$

We can construct the following arguments:

- A_1: $\rightsquigarrow r$, $\mathsf{Conc}(A_1) = r$
- A_2: $A_1 \Rightarrow \neg s$
- A_3: $A_2 \Rightarrow \neg t_2$
- B_1: $\mathsf{Assume}(s)$, $\mathsf{As}(B_1) = \{s\}$, $\mathsf{Conc}(B_1) = s$
- B_2: $B_1 \rightsquigarrow p$, $\mathrm{As}(B_2) = \{s\}$, $\mathsf{Conc}(B_2) = p$
- B_3: $B_2 \rightsquigarrow \bot$, $\mathsf{As}(B_3) = \{s\}$, $\mathsf{Conc}(B_3) = \bot$
- B_4: $\mathsf{ProofByContrad}(\neg s, B_3)$, $\mathsf{Conc}(B_4) = \neg s$

Notice that we do not have an undercutting attack from A_3 to B_4 on B_3, since $A_2 \in Sub(A_3)$, $\mathsf{Conc}(A_2) = \neg s$ and $s \in \mathsf{As}(B_3)$. The intuition is that since A_3 is using $\neg s$ in its inferences, it would not stand as a valid argument under the assumption s under which B_3 is working. Hence, there is no undercutting attack from A_3 to B_4 nor B_3.

Let us now define how explanations arise in our system. One type of explanatory relation is from arguments to explananda, where the argument proposes an explanation of the phenomenon described by the explananda. Recall that in our case, explananda are generated by arguments which derive an inconsistency from no assumption. Hence, for an argument to explain it, we want it to prevent the derivation of the inconsistency by attacking one of the inferences it made. This way, the inconsistency from no assumption would not be derivable anymore. However, note that the attack must also be properly motivated, otherwise the explanation would easily be attacked by other arguments and would not stand a chance as an interesting explanation in the later stages of the modeling process.

The other type of explanatory relation is between arguments themselves, which allows for explanations to be deepened. An argument A which explains an explanandum E may be further explained by another argument B which explains one of A's premises or the link between A's premises and conclusion. Notice that in our system, premises are merely intuitively strict rules with no antecedents, and thus both of these intuitive cases combine into one formal case: B explains A if B non-trivially concludes one of the inference rules r used in A. The case where B explains one of A's premises corresponds to the case where r has no antecedent while the case where B explains the link between A's premises and conclusion corresponds to the case where r has at least one antecedent. By non-trivial, we mean that B must have some sub-argument which is not itself. This way, an argument which merely concludes r from no antecedent will not deepen A's explanation. This is in line with our intuition as such an argument would realistically not provide any further details on the origin of the premise.

Definition 3.2.2. Explanatory relations:

- An argument A *explains* an explanandum E if and only if A attacks **Source**(E).
- An argument B *explains* another argument A (on A') if and only if $\mathsf{Conc}(B) = \mathsf{TopRule}(A')$ for some $A' \in \mathsf{Sub}(A)$ such that $\mathsf{As}(B) \subseteq \mathsf{As}(A')$ and $\mathsf{Sub}(B) \setminus B \neq \emptyset$.

Notice that we require B to be under less assumptions than A', as it would reasonably not count as an explanation if B was working under more assumptions than A'. It is however possible for B to provide more details on a specific step in a proof by contradiction leading to the explanation of an explanandum, hence why we allow A' and B to have non-empty sets of assumptions.

Example 3.2.3. Consider an argumentation theory $AT_5 = (\mathcal{L}, \mathcal{R}, n)$ where $\mathcal{L}$ consists of p, q, r, s, t, u and their negations, $\mathcal{R}_{is} = \{p_1, p_2, p_3, p_4, p_5, t_1, t_2, t_3, t_4, t_5, t_6, t_7\}$ and $\mathcal{R}_d = \emptyset$, where:

- t_1: $p \rightsquigarrow q$
- t_2: $q \rightsquigarrow r$
- t_3: $r \rightsquigarrow \bot$
- t_4: $s \rightsquigarrow \neg p$
- t_5: $t \rightsquigarrow u$
- t_6: $u \rightsquigarrow \neg t_2$
- t_7: $v \rightsquigarrow t_8$

- p_1: $\rightsquigarrow p$
- p_2: $\rightsquigarrow s$
- p_3: $\rightsquigarrow t$
- p_4: $\rightsquigarrow u$
- p_5: $\rightsquigarrow v$

We can then construct the following arguments and explanandum:

- A_1: $\rightsquigarrow p$, $\mathsf{Conc}(A_1) = p$
- A_2: $p \rightsquigarrow q$, $\mathsf{Conc}(A_2) = q$
- A_3: $q \rightsquigarrow r$, $\mathsf{Conc}(A_3) = r$
- A_4: $r \rightsquigarrow \bot$, $\mathsf{Conc}(A_4) = \bot$
- E: $\textbf{Source}(E) = A_4$
- B_1: $\rightsquigarrow s$, $\mathsf{Conc}(B_1) = q$
- B_2: $s \rightsquigarrow \neg p$, $\mathsf{Conc}(B_2) = \neg p$
- C_1: $\rightsquigarrow t$, $\mathsf{Conc}(C_1) = t$
- C_2: $t \rightsquigarrow u$, $\mathsf{Conc}(C_2) = u$
- D_1: $\rightsquigarrow u$, $\mathsf{Conc}(D_1) = u$
- D_2: $u \rightsquigarrow \neg t_2$, $\mathsf{Conc}(D_2) = \neg t_2$
- F_1: $\rightsquigarrow v$, $\mathsf{Conc}(F_1) = v$
- F_2: $v \rightsquigarrow t_8$, $\mathsf{Conc}(F_2) = t_8$

The explanandum E arises from A_4's derivation of $\bot$ from no assumption. By undercutting A_4 on A_1, B_2 explains E. Similarly, by undercutting A_4 on A_3, D_2 also explains E and is further explained by C_2, which non-trivially concludes u. In turn, C_2 is further explained by F_2, which non-trivially concludes $t_8 \in \mathsf{Rules}(C_2)$.

3.3 Defeats and successful explanations

Similarly as in ASPIC+, one can also define a notion of successful and unsuccessful attack. As mentioned before, rebuttal is symmetrical and hence if one argument rebuts another one, the second also rebuts the first. This might lead to situations where two arguments A and B are in conflict, yet even supposing A seems more plausible than B, B would still be able to defend itself from A and come out as an acceptable argument in the end. To prevent this kind of scenario, we use the notion of preference over the rules used in ASPIC+, which will translate into a similar relation over the arguments. One should provide a preference relation $\leq$ over the rules, where $t_1 \leq t_2$ means that t_2 is preferred to t_1. We also define a strict preference relation (denoted $<$), where $t_1 < t_2$ if and only if $t_1 \leq t_2$ and $t_2 \not\leq t_1$. Note that, contrary to ASPIC+, we do not restrict the preference relation to the defeasible rules, and also allow comparison of the intuitively strict rules.

We can then lift the preference relation from rules to arguments using the weakest-link principle, which we define for ASPIC-END as follows:

Definition 3.3.1. Weakest-link preference:

Let A and B be two arguments. We have that $A \preceq_w B$ if and only if $\mathsf{Rules}(A) \neq \emptyset$ and:

$$\text{There exists } r_a \in \mathsf{Rules}(A)\text{, such that for all } r_b \in \mathsf{Rules}(B)\text{, we have } r_a \leq r_b$$

Notice that this definition of weakest-link preference differs from the one presented in Section 2.6 in the fact that we do not have a set of premises in ASPIC-END, and thus it is only needed to compare the preference of the rules. We do however compare the preference of all rules, including the intuitively strict ones, as we allow for a preference to be defined between different intuitively strict rules and also between intuitively strict rules and defeasible ones.

Example 3.3.1. Consider an argumentation theory $AT_6 = (\mathcal{L}, \mathcal{R}, n)$ where $\mathcal{L}$ consists of p, q, r, s and their negations, $\mathcal{R}_{is} = \{p_1, p_2\}$ and $\mathcal{R}_d = \{d_1, d_2, d_3\}$, where:

- p_1: $\rightsquigarrow p$
- p_2: $\rightsquigarrow r$
- d_1: $p \Rightarrow q$
- d_2: $s \Rightarrow \neg q$
- d_3: $r \Rightarrow s$

We also define a preference relation $\leq$ such that $d_3 < d_1$ and $d_1 < d_2$. We can then construct the following arguments:

- A_1: $\rightsquigarrow p$, $\text{Conc}(A_1) = p$
- A_2: $A_1 \Rightarrow q$, $\text{Conc}(A_2) = q$
- B_1: $\rightsquigarrow r$, $\text{Conc}(B_1) = r$
- B_2: $r \Rightarrow s$, $\text{Conc}(B_2) = s$
- B_3: $B_2 \Rightarrow \neg q$, $\text{Conc}(B_3) = \neg q$

Notice that by the weakest-link principle, $B_3 \preceq_w A_2$, but we do not have that $A_2 \preceq_w B_3$.

We can then use this notion of preference over arguments to define defeat and successful explanations.

Definition 3.3.2. Defeats:

An argument A *defeats* an argument B with respect to a preference relation $\preceq$ if and only if:

$$A \text{ rebuts, undercuts or assumption-attacks } B \text{ on } B' \text{ and } B' \preceq A$$

We now have that conflicts can be resolved through preferences. This will prevent symmetrical attacks from surviving between arguments with different preferences. Also, note that we also consider the preference of the argument in the case of undercutting. In our study of the liar paradox, we will often come across low preference arguments which derive an inconsistency from no assumption and thus will be able to derive any formula in the language, including the negation of any rule. This kind of argument would then be able to undercut every argument, even ones with higher preference. Also, note that since we consider premises as intuitive rules with no antecedent and thus regard undermining a special case of undercutting, not considering preferences in the cases of undercutting means we would also not take preferences into account in the cases of undermining. Yet, we want to be able to block undermining attacks from low preference arguments to high preference ones.

Let us now define what it means for an explanation to be successful, using this new notion of defeat.

Definition 3.3.3. Successful explanations:
Let A be an argument and E an explanandum. *A successfully explains B* with respect to a preference relation $\preceq$ if and only if:

$$A \text{ defeats } \mathsf{Source}(B) \text{ with respect to } \preceq$$

By using the notion of "defeat" instead of "attack", this definition of successful explanations takes into account preferences.

3.4 Explanatory argumentation frameworks and argument selection procedures

We now have the tools to not only generate arguments and explananda from a set of rules, but we also have defined what it means for an argument to be preferred to another and how this affects the defeat and successful explanation relations. We can now use this to build *explanatory argumentation frameworks.* After this, we will define the procedures which allow one to extract the most relevant explanations from such a framework.

Definition 3.4.1. Explanatory argumentation frameworks:
Let $AT = (\mathcal{L}, \mathcal{R}, n)$ be an argumentation theory and $\leq$ a preference relation defined over $\mathcal{R}$. An *explanatory argumentation framework* (EAF) defined by AT and $\leq$, 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 $\mathcal{R}$ satisfying Definition 3.1.2;
- $\mathcal{X}$ is the set of all explananda that can be constructed from the arguments in $\mathcal{A}$ satisfying Definition 3.1.3;
- $\preceq$ is the preference relation over arguments obtained from lifting $\leq$ according to the weakest-link principle;
- $(X, Y) \in \rightarrow$ if and only if X defeats Y with respect to $\preceq$, where $X, Y \in \mathcal{A}$;
- $(X, E) \in \dashrightarrow$ if and only if X successfully explains E with respect to $\preceq$, where $X \in \mathcal{A}$ and $E \in \mathcal{X}$;
- $(X, Y) \in \dashrightarrow$ if and only if X explains Y, where $X, Y \in \mathcal{A}$.

Notice that for the explanation between arguments, we have not used any notion of preference, as it is mostly used to resolve attacks. While explanations from arguments to explananda are generate from attacking the explanandum's source, explaining another argument does not involve any attack and thus does not require one to consider preferences.

Once such a framework has been generated, we want to be able to extract the most interesting set of arguments from the graph. Such a set should be able to explain as many explananda as possible while providing as many details as possible, as well as being self-consistent and plausible. To do so, we consider conflict-freeness, defense, explanatory power and explanatory depth as defined in chapter 2. As a reminder, the explanatory power is defined by how much of the explananda a given explanation can explain, while the explanatory depth is defined by how deep an explanation is, i.e. how long the longest chain of explanation is.

We define two selection procedures similar to the ones introduced by Šešelja and Straßer [13]. As they have suggested in their paper, we chose to reverse the ordering of the steps 2 and 3 and thus 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 fall prey to many attacks from beating plausible theories which fail to explain some of the explananda but are sound and fully defended. Also, notice that once the conflict-free sets of arguments have been selected, selecting the most defended sets is equivalent to selecting the self-defending ones. Indeed, the empty set is always conflict-free and self-defended and hence if no other conflict-free set of arguments is fully self-defended, the empty set will be the only remaining set after this step. This is why we rephrase the defense criterion step as the selection of the self-defending sets.

Definition 3.4.2. Argument selection procedures:

Procedure 1:

In order to select the argumentative core of the most explanatory theories, proceed as follows:

1. Select all the conflict-free sets of arguments.
2. Out of these, select the self-defending ones.
3. Out of these, select the most explanatory powerful ones.
4. Out of these, select the maximal ones with respect to set inclusion.

Procedure 2:

In order to select the explanatory core of the most explanatory theories, proceed as follows:

1. Select all the conflict-free sets of arguments.
2. Out of these, select the self-defending ones.
3. Out of these, select the most explanatory powerful ones.
4. Out of these, select the explanatory deepest ones.
5. Out of these, select the minimal ones with respect to set inclusion.

Notice again that in both cases, steps 1 and 2 can be merged into the step “Select all admissible sets of arguments”.

Chapter 4

Applying ASPIC-END to the liar paradox

In the previous chapter, we have defined a system, ASPIC-END, which was designed for building a formal argumentative model of the liar paradox and its proposed solutions. We will now apply ASPIC-END to model the arguments from a few short texts, which will each focus on a specific solution of the liar paradox. These texts were provided by my collaborator Marcos Cramer for the purpose of this research. Afterwards, we will put these frameworks together in one bigger model and analyze it.

4.1 The paracomplete solutions

We will start by observing a general description of the paracomplete solutions. There exist many different versions, however they all share this in common: they reject the law of excluded middle in some way, which says that every formula is either true or false. Let us first look at the following short text which describes the main idea of the paracomplete solutions:

Excerpt 1

> 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 absurd.
> 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 absurdity that L is both true and false from no assumption.
> One possible solution: 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. Even though applicable in many situations, this principle is not applicable to problematically self-referential sentences like L.

Let us try to construct an argumentation theory which reflects the reasoning present in the text. For this, we need to define a language, a set of intuitively strict and defeasible rules and a naming function n. We will then need to define a preference relation over these rules. Let us start by the construction of the rule set. Once this is done, we can define the name of each rule and then simply take the closure under negation of all the symbols we used as our language. The goal of our set of rules is to be able to capture the author's reasoning from the text, but also any implicit inferences the author might have made which does not appear explicitly in the text. To do so, we will analyze the text sentence by sentence.

We will represent our rules in the same format as we did in the examples from the previous chapter, where the label in front of the rule represents its name given by the function n.

From the first sentence, *Define L to be the sentence "L is false"*, we can create the rule:

$$t_1\colon\ Ltrue \rightsquigarrow \text{``}Lfalse\text{''}true$$

Here, the formula *Ltrue* represents the proposition 'L is true' and the formula "*Lfalse*"*true* represents the proposition '"L is false" is true'.

This would correspond to one direction of an instance of the rule of intersubstitutivity of equivalents. We can create another rule which represents the equivalence in the other direction:

$$t_2\colon\ \text{``}Lfalse\text{''}true \rightsquigarrow Ltrue$$

Notice that we make these rules intuitively strict, as the intersubstitutivity of equivalents is not defeasible in the sense that if applicable, its conclusion cannot be trivially rejected. One can however question its applicability in the given context.

Let us now look at the second sentence: *If L is true, i.e. "L is false" is true, then L is false, which is absurd.* This sentence contains a number of reasoning steps. Let us break the sentence down further and first focus on *If L is true, i.e. "L is false" is true.* Notice that the *If* marks the introduction of an assumption, namely *Ltrue.* This will be important during the construction of the arguments so let us keep it in mind, even though it is less relevant for our current task of creating rules. The use of the word *i.e.* represents the usage of rule t_1, the inference from *Ltrue* to "*Lfalse*"*true.* We now have *then L is false*, where the author is making an inference from "*Lfalse*"*true* to *Lfalse.* This can be represented as the following rule:

$$t_3\colon\ \text{``}Lfalse\text{''}true \rightsquigarrow Lfalse$$

Here, the formula *Lfalse* represents the proposition 'L is false'.

This would be an instance of the truth schema which says that for every sentence S, 'S' is true if and only if S. This allows one to infer the sentence S from a statement that S is true. Notice that the truth schema also allows the inference to be made in the other direction, and hence we should also include the following rule:

$$t_4\colon\ Lfalse \rightsquigarrow \text{``}Lfalse\text{''}true$$

Notice that again we represent them as intuitively strict rules, as unless one has an argument against the applicability of the truth schema in this context, one must accept any conclusion derived from it.

We then have the last segment of this sentence: *which is absurd.* In the whole sentence, the author has made three sub-conclusions: 'L is true', '"L is false" is true' and 'L is false'. Two of them are in obvious contradiction, namely 'L is true' and 'L is false'. This gives us the following rule:

$$t_5\colon\ Ltrue, Lfalse \rightsquigarrow \bot$$

We can now move on to the next sentence: *So L is not true, i.e. L is false.* First notice that the author has now retracted the assumption 'L is true', and because it led to a contradiction, has concluded 'L is not true'. There is now an inference from 'L is not true' to 'L is false', which we will represent as the following rule:

$$t_6\text{: } \neg Ltrue \leadsto Lfalse$$

Here, the formula $\neg Ltrue$ represents the proposition 'L is not true'.

After this, the author writes: *So "L is false" is true, i.e. L is true.* There we have the use of rule t_4 to infer '"L is false" is true' from 'L is false'. This is followed by the use of rule t_2 which derives 'L is true' from '"L is false" is true'. Hence no new rules need to be added here.

We then have the sentence: *So we have the absurdity that L is both true and false from no assumption.* This is the same reasoning step that was used previously in the first paragraph and that we modeled as rule t_5. Notice that this time the author explicitly mentioned there being no assumptions at this point.

We then move on to the solution. The first sentence *L is neither true nor false* seems to introduce the following premise:

$$p_1 :\leadsto \neg LEitherTrueOrFalse$$

Here, the formula $\neg LEitherTrueOrFalse$ represents the proposition 'L is neither true nor false'.

From this we should be able to infer separately that 'L is not true' and 'L is not false':

$$t_7\text{: } \neg LEitherTrueOrFalse \leadsto \neg Ltrue$$
$$t_8\text{: } \neg LEitherTrueOrFalse \leadsto \neg Lfalse$$

We then focus on the next sentence: *When concluding that L is false because L is not true, we are making the assumption that any sentence is either true of false.* Here, the author seems to question the applicability of the rule t_6 by saying that for it to be applicable, we need to have that L is either true or false. By claiming that the sentence is neither true nor false, there is motivation to attack the applicability of the rule t_6:

$$t_9\text{: } \neg LEitherTrueOrFalse \leadsto \neg t_6$$

Notice that here we use again an intuitively strict rule, as to defend the applicability of the rule t_6, one would need to either attack the premise 'L is neither true nor false' or attack the rule t_9 itself, a rebuttal does not intuitively seem enough to warrant an attack.

The author then proceeds to further explain the solution by giving motivations for the premise $\neg LEitherTrueOrFalse$ in the last sentence: *Even though applicable in many situations, this principle is not applicable to problematically self-referential sentences like L.* This can be modeled as introducing a premise about L's self-reference and a rule which links it to the formula $\neg LEitherTrueOrFalse$.

$$p_2\text{: } \leadsto LProblSelfRef$$
$$t_{10}\text{: } LProblSelfRef \leadsto \neg LEitherTrueOrFalse$$

We now have a set of 12 intuitively strict rules $\mathcal{R}_is = \{t_1, ..., t_{10}, p_1, p_2\}$ and a language $\mathcal{L}$ which consists of $Ltrue$, $Lfalse$, "$Lfalse$"$true$, $\perp$, $LEitherTrueOrFalse$, $LProblSelfRef$, $t_1, ..., t_{10}$, and their negations. Also, we have a preference relation over the rules. We consider the rules on the meta-level of higher preference than the ones on the object level, and hence get the following relation:

$$t_1, ..., t_6 < t_7, ..., t_{10}, p_1, p_2$$

Meaning that each rule $t_1, ..., t_6$ is strictly less preferred than each of $t_7, ..., t_{10}, d_1, d_2$.

We can now construct the following arguments and explanandum:

- A_1: $\mathsf{Assume}(Ltrue)$, TopRule is undefined, As $= Ltrue$, Conc $= Ltrue$

- A_2: $A_1 \rightsquigarrow$ "$Lfalse$"$true$, TopRule $= t_1$, As $= Ltrue$, Conc $=$ "$Lfalse$"$true$
- A_3: $A_2 \rightsquigarrow Lfalse$, TopRule $= t_3$, As $= Ltrue$, Conc $= Lfalse$
- A_4: $A_1, A_3 \rightsquigarrow \perp$, TopRule $= t_5$, As $= Ltrue$, Conc $= \perp$
- A_5: $\mathsf{ProofByContrad}(\neg Ltrue,\ A_4)$, TopRule is undefined, As $= \emptyset$, Conc $= \neg Ltrue$
- A_6: $A_5 \rightsquigarrow Lfalse$, TopRule $= t_6$, As $= \emptyset$, Conc $= Lfalse$
- A_7: $A_6 \rightsquigarrow$ "$Lfalse$"$true$, TopRule $= t_4$, As $= \emptyset$, Conc $=$ "$Lfalse$"$true$
- A_8: $A_7 \rightsquigarrow Ltrue$, TopRule $= t_2$, As $= \emptyset$, Conc $= Ltrue$
- A_9: $A_8, A_6 \rightsquigarrow \perp$, TopRule $= t_5$, As $= \emptyset$, Conc $= \perp$
- E_1: Source $= A_9$
- B_1: $\rightsquigarrow \neg LEitherTrueOrFalse$, TopRule $= p_1$, As $= \emptyset$, Conc $= \neg LEitherTrueOrFalse$
- B_2: $B_1 \rightsquigarrow \neg t_6$, TopRule $= t_9$, As $= \emptyset$, Conc $= \neg t_6$
- C_1: $\rightsquigarrow LProblSelfRef$, TopRule $= p_2$, As $= \emptyset$, Conc $= LProblSelfRef$
- C_2: $C_1 \rightsquigarrow \neg LEitherTrueOrFalse$, TopRule $= t_{10}$, As $= \emptyset$, Conc $= \neg LEitherTrueOrFalse$
- D_1: $B_1 \rightsquigarrow \neg Ltrue$, TopRule $= t_7$, As $= \emptyset$, Conc $= \neg Ltrue$
- D_2: $B_1 \rightsquigarrow \neg Lfalse$, TopRule $= t_8$, As $= \emptyset$, Conc $= \neg Lfalse$

Notice that there is an explanandum E_1 with source A_9 as it is an argument with conclusion $\perp$ under no assumption. We can see that there is an undercutting attack from B_2 to A_9 on its sub-argument A_6. Hence, B_2 provides an explanation for E_1. Also, this explanation is further explained by C_2 as it concludes $\neg LEitherTrueOrFalse$, one of B_2's premises, in a non-trivial way.

We then get the following explanatory argumentation framework. Notice that most solitary arguments have been omitted for clarity as they are of little relevance to the framework. Also, A_7 and A_8 are not represented, because even though they are also subject to an undercutting on A_6, the impact on the rest of the framework is irrelevant.

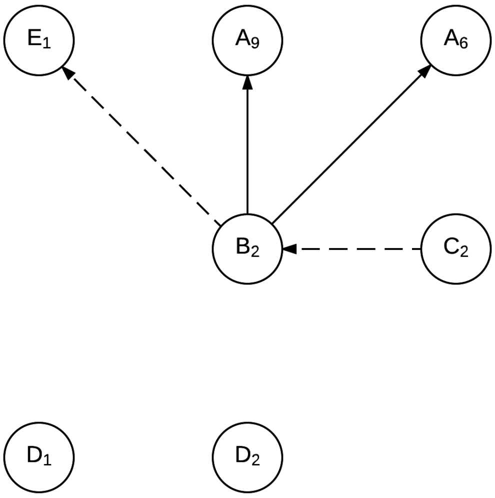

Figure 4.1: EAF representing the arguments in the first excerpt

By applying the selection procedure 1, we get that the argumentative core of the most explanatory powerful theory is $S_1 = \{B_2, C_2, D_1, D_2\}$. With the selection procedure 2, we get that the explanatory core is $S_2 = \{B_2, C_2\}$.

4.2 The presupposition solution

One of the main problems in defending a solution to the liar paradox is that the solutions are easily subject to so called *revenge paradoxes*, which rephrase the paradox in a specific way so that the solution fails to explain this new paradox. However, for each revenge paradox we usually find another solution to explain it. Let us examine the following short text which illustrates such a case:

Excerpt 2

> Define L to be the sentence “L is not true”. If L is true, i.e. “L is not true” is true, then L is not true, which is absurd.
> Therefore, L is not true. So “L is not true” is true, i.e. L is true. So we have an absurdity from no assumption.
> One possible solution: When we say that a sentence is true, what we really mean is that the proposition expressed by the sentence is true. But some grammatically well-formed sentences do not express a proposition. For example, “the present king of France is bald” does not express a proposition as France does not currently have a king. Both when saying that a sentence is true or when saying that a sentence is not true, we are presupposing that it expresses a proposition. L does not express a proposition, because interpreting L requires finding a proposition expressed by the sentence called L, which requires interpreting L etc. ad infinitum, so that it can never be fully interpreted. So already the step where we assume L to be true is problematic.

We will proceed in the same fashion as we did for the previous text and start by extracting rules from the text sentence by sentence. Again, we will present the rules in the format *name: rule.*

Let us start with *Define L to be the sentence “L is not true”.* Again we have a definition here of the sentence L. Since we wish to combine the frameworks later, let us call this sentence L_2 in our argumentation theory. We can then create the following rules:

$$\begin{aligned} t_1\colon&\ L_2\mathit{true} \rightsquigarrow \text{“}\neg L_2\mathit{true}\text{”}\mathit{true}\\ t_2\colon&\ \text{“}\neg L_2\mathit{true}\text{”}\mathit{true} \rightsquigarrow L_2\mathit{true}\end{aligned}$$

Here, the formulas $L_2\mathit{true}$ and “$\neg L_2\mathit{true}$”true stand for the propositions ‘L_2 is true’ and ‘“L_2 is not true” is true’ respectively.

These rules are similar to the first ones we created in the previous text and the few next ones will also have some similarity, as the process to derive a contradiction from a liar sentence usually follows the same pattern. After defining the sentence, one assumes for a contradiction that it is true, and then using the principle of intersubstitutivity of equivalents and some instance of the truth schema, one concludes that it is both true and not true, which allows one to retract the assumption that this liar sentence is true and conclude under no assumption that it is not true. From there one can derive a contradiction from no assumption with a similar process. To represent this reasoning process for the sentence L_2, we create the following rules:

$$\begin{aligned} t_3\colon&\ \text{“}\neg L_2\mathit{true}\text{”}\mathit{true} \rightsquigarrow \neg L_2\mathit{true}\\ t_4\colon&\ \neg L_2\mathit{true} \rightsquigarrow \text{“}\neg L_2\mathit{true}\text{”}\mathit{true}\\ t_5\colon&\ L_2\mathit{true}, \neg L_2\mathit{true} \rightsquigarrow \bot\end{aligned}$$

Here we have again an instance of the truth schema in both directions, as well as a rule representing the contradiction which arises from L_2 being true and not true at the same time.

Let us now move on to modeling the solution to this version of the paradox. Instead of trying to build a set of rules sentence by sentence as we have done before, we will this time try to analyze the main points of the author's explicit reasoning, but also the implicit steps which do not appear directly in the text.

In direct relation to the paradox, there is the statement that L_2 does not express a proposition, followed by the explanation that trying to interpret it requires an infinite loop of interpreting L_2 so that the interpretation can never be completed. We can model this as such:

$$p_1\colon \rightsquigarrow InterpretL_2AdInf$$
$$d_1\colon InterpretL_2AdInf \Rightarrow \neg L_2ExpressProp$$

Here, the formulas $InterpretL_2AdInf$ and $\neg L_2ExpressProp$ represent the propositions 'L_2 requires interpreting L_2 ad infinitum' and 'L_2 does not express a property' respectively.

However, one could say that is seems far-fetched to say that L_2 does not express a proposition when every other sentence does. And this is where the beginning of the paragraph comes into play. The author first states that some sentences, even though well-formed, do not express a proposition, and defends this point of view with the example that the sentence *"the present king of France is bald"*, even though it is well-formed, does not express a proposition because France currently does not have a king. We can model this as follows:

$$p_2\colon \rightsquigarrow EverySentExpressProp$$
$$d_2\colon EverySentExpressProp \Rightarrow L_2ExpressProp$$
$$p_3\colon \rightsquigarrow FranceHasNoKing$$
$$d_3\colon FranceHasNoKing \Rightarrow \neg KingBaldExpressProp$$
$$d_4\colon \neg KingBaldExpressProp \Rightarrow \neg EverySentExpressProp$$

Here, the formulas $EverySentExpressProp$, $FranceHasNoKing$ and $\neg KingBaldExpressProp$ represent the propositions 'every sentence expresses a proposition', 'France currently has no king' and '"the present king of France is bald" does not express a proposition' respectively.

Finally, the last part of this reasoning is present in the first and also the last sentences of the paragraph. When the author says *So already the step where we assume L to be true is problematic*, he is referring to the fact that *When we say a sentence is true, what we really mean is that the proposition expressed by the sentence is true.* Since L_2 does not express a proposition however, making a statement about the truth value of this non-existing proposition expressed by L_2 is problematic. Hence, making an assumption on the truth value of L_2 would not make sense since, in this train of thought, L_2 does not have a truth value. We can represent this as follows:

$$d_5\colon \neg L_2ExpressProp \Rightarrow \neg L_2HasTruthValue$$
$$d_6\colon \neg L_2HasTruthValue \Rightarrow \neg Assume(L_2true)$$

Here, the formula $\neg L_2HasTruthValue$ represents the proposition 'L_2 does not have a truth value'.

We now have a language, a set of rules and a naming function defined. We also define the following preference ordering on the set of rules, representing once again the fact that meta-level reasoning is preferred to object-level reasoning:

$$t_1, ..., t_5 < d_1, ..., d_6, p_1, p_2, p_3$$

Additionally, we have the following preference:

$$p_2 < p_3, d_3, d_4$$

Since even though p_2 and d_4 have opposite conclusion, showing a counter-example is a much stronger argument than simply making a statement that all sentences satisfy some property. Since p_3 and d_3 are also contributing to the construction of the counter-example, they are also preferred to p_2.

We can then construct the following arguments and explanandum:

- D_1: $\mathsf{Assume}(L_2 true)$, TopRule is undefined, As $= L_2 true$, Conc $= L_2 true$
- D_2: $D_1 \leadsto$ "$\neg L_2 true$"$true$, TopRule $= t_1$, As $= L_2 true$, Conc $=$ "$\neg L_2 true$"$true$
- D_3: $D_2 \leadsto \neg L_2 true$, TopRule $= t_3$, As $= L_2 true$, Conc $= \neg L_2 true$
- D_4: $D_1, D_3 \leadsto \bot$, TopRule $= t_5$, As $= L_2 true$, Conc $= \bot$
- D_5: $\mathsf{ProofByContrad}(\neg L_2 true, D_4)$, TopRule is undefined, As $= \emptyset$, Conc $= \neg L_2 true$
- D_6: $D_5 \leadsto$ "$\neg L_2 true$"$true$, TopRule $= t_4$, As $= \emptyset$, Conc $= \neg$"$L_2 true$"$true$
- D_7: $D_6 \leadsto L_2 true$, TopRule $= t_2$, As $= \emptyset$, Conc $= L_2 true$
- D_8: $D_7, D_5 \leadsto \bot$, TopRule $= t_5$, As $= \emptyset$, Conc $= \bot$
- E_2: Source $= D_8$
- F_1: $\leadsto InterpretL_2AdInf$, TopRule $= p_1$, As $= \emptyset$, Conc $= InterpretL_2AdInf$
- F_2: $F_1 \Rightarrow \neg L_2 ExpressProp$, TopRule $= d_1$, As $= \emptyset$, Conc $= \neg L_2 ExpressProp$
- F_3: $F_2 \Rightarrow \neg L_2 HasTruthValue$, TopRule $= d_5$, As $= \emptyset$, Conc $= \neg L_2 HasTruthValue$
- F_4: $F_3 \Rightarrow \neg Assume(L_2 true)$, TopRule $= d_6$, As $= \emptyset$, Conc $= \neg Assume(L_2 true)$
- G_1: $\leadsto EverySentExpressProp$, TopRule $= p_2$, As $= \emptyset$, Conc $= EverySentExpressProp$
- G_2: $G_1 \Rightarrow L_2 ExpressProp$, TopRule$= d_2$, As $= \emptyset$, Conc $= L_2 ExpressProp$
- H_1: $\leadsto FranceHasNoKing$, TopRule $= p_3$, As $= \emptyset$, Conc $= FranceHasNoKing$
- H_2: $H_1 \Rightarrow \neg KingBaldExpressProp$, TopRule $= d_3$, As $= \emptyset$, Conc $= \neg KingBaldExpressProp$
- H_3: $H_2 \Rightarrow \neg EverySentExpressProp$, TopRule $= d_4$, As $= \emptyset$, Conc $= \neg EverySentExpressProp$

We have an explanandum E_2 which arises from the argument D_8. By attacking D_8 on D_1, F_4 provides an explanation for E_2. However, there is an attack from G_2 to F_4 on F_2 on the sub-conclusion $\neg L_2 ExpressProp$. Notice that this attack is symmetrical as F_2 also rebuts G_2. This conflict is however resolved by H_3's undercutting of G_2 on G_1. Note that this attack is not symmetrical since all rules used in H_3, namely p_3, d_3, d_4, are strictly preferred to the rule used in G_1, which is p_2. The resulting graph is illustrated below:

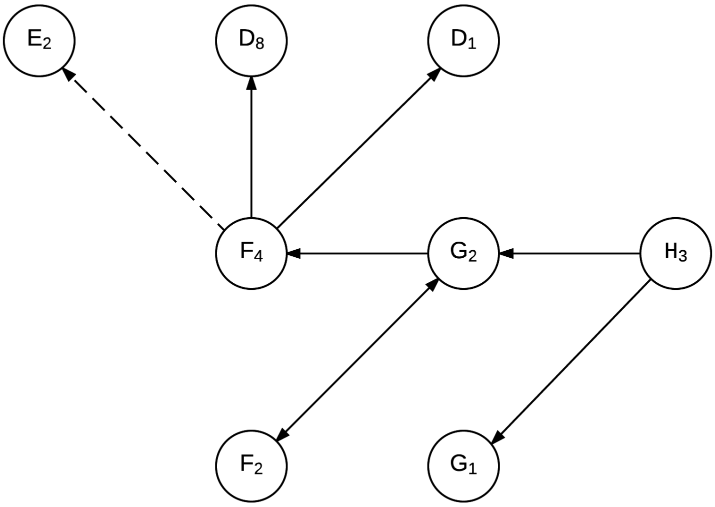

Figure 4.2: EAF representing the arguments in the second excerpt

By applying the selection procedures, we get that the argumentative core of the most explanatory powerful theory is $S_3 = \{F_2, F_4, H_3\}$, whereas the argumentative core is $\{F_4, H_3\}$.

4.3 A. Prior's solution

The third solution we will attempt to formalize is one which was proposed by A. Prior in 1961 [11]. This solution focuses on implicit assertions, and the main idea is described in our third excerpt:

Excerpt 3

> We consider the case where L is defined as "L is false". Every statement includes an implicit assertion of its own truth. E.g. when asserting "Snow is white", we implicitly assert that it is true that snow is white. So L says explicitly that L is false, but also implicitly that L is true. So it is just a contradictory sentence and thus false. We cannot conclude from "L is false" that L is true, because "L is false" contradicts the implicit assertion of L and thus actually makes L false. So we cannot derive a contradiction about the truth-value of L.

Since this excerpt refers to the same version of the liar paradox as the first one, it does not repeat how the contradiction is reached from no assumption. Hence, we will simply reuse the rules we created for the first excerpt:

$$\begin{gathered} t_1\text{: } Ltrue \rightsquigarrow \text{``}Lfalse\text{''}true \\ t_2\text{: } \text{``}Lfalse\text{''}true \rightsquigarrow Ltrue \\ t_3\text{: } Lfalse \rightsquigarrow \text{``}Lfalse\text{''}true \\ t_4\text{: } \text{``}Lfalse\text{''}true \rightsquigarrow Lfalse \\ t_5\text{: } Ltrue, Lfalse \rightsquigarrow \bot \\ t_6\text{: } \neg Ltrue \rightsquigarrow Lfalse \end{gathered}$$

Let us now try to model this new solution. In the first sentence, we can identify the introduction of a new premise. This would give us the following rule:

$$p_1\colon \rightsquigarrow EveryStatAssertTrue$$

Here, the formula $EveryStatAssertTrue$ represents the proposition 'Every statement includes an implicit assertion of its own truth'.

This is followed by an example which illustrates this principle. We could imagine that someone would doubt the principle just introduced and even come to claim that the opposite is true, namely that no statement does follow this principle. One could then introduce this claim as another premise:

$$p_2\colon \rightsquigarrow NoStatAssertTrue$$

Here, the formula $NoStatAssertTrue$ represents the proposition 'No statement includes an implicit assertion of its own truth'.

Notice that the formula $NoStatAssertTrue$ is not equivalent to $\neg EveryStatAssertTrue$ as this simply translates into the proposition 'There exists some statement which does not include an assertion of its own truth'. The two formulas are however clearly incompatible and thus we need the rules:

$$t_7\colon NoStatAssertTrue \rightsquigarrow \neg EveryStatAssertTrue$$
$$t_8\colon EveryStatAssertTrue \rightsquigarrow \neg NoStatAssertTrue$$

The examples comes now into play as it defends the principle introduced in the excerpt from this potential attack. By being an example for the principle that all statements contain an assertion of their own truth, it is a counter-example for the principle that no statement does so.

$$p_3\colon \rightsquigarrow SnowWhiteAssertTrue$$
$$t_9\colon SnowWhiteAssertTrue \rightsquigarrow \neg NoStatAssertTrue$$

Here, the formula $SnowWhiteAssertTrue$ represents the proposition '"Snow is white" contains an assertion of its own truth'.

Notice that once again we will give preference to the counter-example compared to the general principle that no statement contains an assertion of its own truth. This way, the example provides indirect support to the principle that every statement contains an assertion of its own truth by defending it against this potential attack. Note that this does not qualify however as a deepening of the explanation.

We now come back to the liar sentence L and apply the newly introduced principle. From this principle, we can derive that L states about itself that it is false, but also implicitly that it is true:

$$t_{10}\colon EveryStatAssertTrue \rightsquigarrow LAssertTrueAndFalse$$

Here, the formula $LAssertTrueAndFalse$ represents the proposition 'L contains an assertion of its own truth and of its own falsehood'.

The author then states that L is contradictory and thus false. We can represent this as such:

$$t_{11}\colon LAssertTrueAndFalse \rightsquigarrow LContradictory$$
$$t_{12}\colon LContradictory \rightsquigarrow Lfalse$$

Here, the formula $LContradictory$ represents the proposition ‘L is contradictory’.

In the derivation to a contradiction from no assumption from the liar sentence, one step is that because it is true that “L is false”, we can say that L is true. However, according to the principle just introduced, having “L is false” contradicts L’s assertion of its own truth and thus does not make L true but actually false. There is now an attack on this inference step from the contradictory nature of L.

$$d_1\colon\ LContradictory \Rightarrow \neg t_2$$

We now have a full set of rules which represent the reasoning in the text, a language and a naming function. We define the preference relation as usual by preferring the meta-level rules to the object-level ones:

$$t_1, ..., t_6 < t_7, ..., t_{12}, d_1, p_1, p_2, p_3$$

Also, as mentioned earlier, we prefer the counter-example to the principle which is targeted:

$$p_2 < p_3, t_9$$

We can then construct the following arguments and explanandum:

- A_1: $\mathsf{Assume}(Ltrue)$, TopRule is undefined, As = $Ltrue$, Conc = $Ltrue$
- A_2: $A_1 \rightsquigarrow$ “$Lfalse$”$true$, TopRule = t_1, As = $Ltrue$, Conc = “$Lfalse$”$true$
- A_3: $A_2 \rightsquigarrow Lfalse$, TopRule = t_3, As = $Ltrue$, Conc = $Lfalse$
- A_4: $A_1, A_3 \rightsquigarrow \bot$, TopRule = t_5, As = $Ltrue$, Conc = $\bot$
- A_5: $\mathsf{ProofByContrad}(\neg Ltrue,\ A_4)$, TopRule is undefined, As = $\emptyset$, Conc = $\neg Ltrue$
- A_6: $A_5 \rightsquigarrow Lfalse$, TopRule = t_6, As = $\emptyset$, Conc = $Lfalse$
- A_7: $A_6 \rightsquigarrow$ “$Lfalse$”$true$, TopRule = t_4, As = $\emptyset$, Conc = “$Lfalse$”$true$
- A_8: $A_7 \rightsquigarrow Ltrue$, TopRule = t_2, As = $\emptyset$, Conc = $Ltrue$
- A_9: $A_8, A_6 \rightsquigarrow \bot$, TopRule = t_5, As = $\emptyset$, Conc = $\bot$
- E_1: Source = A_9
- I_1: $\rightsquigarrow EveryStatAssertTrue$, TopRule = p_1, As = $\emptyset$, Conc = $EveryStatAssertTrue$
- I_2: $I_1 \rightsquigarrow \neg NoStatAssertTrue$, TopRule = t_8, As = $\emptyset$, Conc = $\neg NoStatAssertTrue$
- I_3: $I_1 \Rightarrow LAssertTrueAndFalse$, TopRule = t_{10}, As = $\emptyset$, Conc = $LAssertTrueAndFalse$
- I_4: $I_3 \Rightarrow LContradictory$, TopRule = t_{11}, As = $\emptyset$, Conc = $LContradictory$
- I_5: $I_4 \Rightarrow \neg t_2$, TopRule = d_1, As = $\emptyset$, Conc = $\neg t_2$
- I_6: $I_4 \Rightarrow Lfalse$, TopRule = t_{12}, As = $\emptyset$, Conc = $Lfalse$
- J_1: $\rightsquigarrow NoStatsAssertTrue$, TopRule = p_2, As = $\emptyset$, Conc = $NoStatsAssertTrue$
- J_2: $J_1 \rightsquigarrow \neg EveryStatAssertTrue$, TopRule = t_7, As = $\emptyset$, Conc = $\neg EveryStatAssertTrue$

- K_1: $\rightsquigarrow SnowWhiteAssertTrue$, TopRule $= p_3$, As $= \emptyset$, Conc $= SnowWhiteAssertTrue$
- K_2: $K_1 \Rightarrow \neg NoStatAssertTrue$, TopRule $= t_9$, As $= \emptyset$, Conc $= \neg NoStatAssertTrue$

Notice that the argument to the contradiction and the explanandum are the same as in the model of the first text, since these two texts are focusing on the same version of the paradox.

This time, we have that I_5 explains E_1 by attacking A_9 on A_8. I_5 is then attacked by both J_1 and J_2 on I_2 and I_1 respectively, with rebuttals. These rebuttals also translate in attacks from I_1 and I_2 to J_2 and J_1 respectively. Note that we also have another symmetrical rebuttal from I_2 to J_2 on J_1 and from J_2 to I_2 on I_1. This conflict is however resolved by K_2 attacking J_2 on J_1. This attack is unidirectional since the rules in K_2 are all strictly preferred to the rules in J_1. We get the following graph:

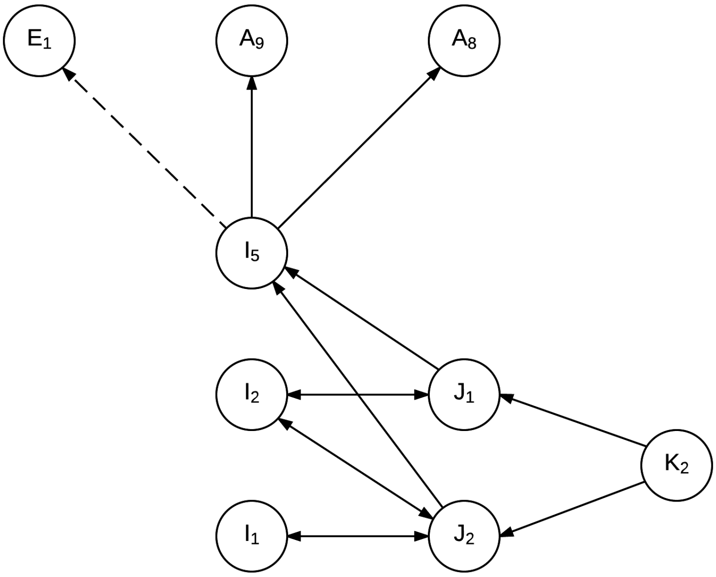

Figure 4.3: EAF representing the arguments in the third excerpt

By applying the argument selection procedures, we get that the argumentative core of the most explanatory theory is $\{I_1, I_2, I_5, K_2\}$, while the explanatory core is $\{I_5, K_2\}$.

4.4 Global analysis

We will now group all the arguments we have constructed from the three texts and analyze the resulting framework. Notice that the explanatory argument F_4 we had in the second excerpt could be slightly modified to also explain E_1 by attacking A_9 on A_1. Indeed, if the sentence L_2 does not express a proposition, neither does the sentence L. Hence, making the assumption that L is true makes as little sense as making the assumption that L_2 is true. Therefore, F_4 also explains E_1.

Similarly, I_5 could be slightly modified to also explain E_2. If L includes an assertion of its own truth, then so does L_2. Hence L_2 explicitly states its own untruth, but also implicitly states its own truth. Therefore L_2 is contradictory in the same way that L is, and one cannot infer from “L_2 is not true” that L_2 is true, and thus cannot derive a contradiction from no assumption. So it is fair to say that I_5 also explains E_2.

On the other hand, as was mentioned earlier, the first solution does not explain E_2 as the step it attacks in the first version of the liar paradox does not appear in any form in the second version. Indeed, it is a “revenge liar” on the first solution which aims at deriving a contradiction from no assumption in a way which that solution does not reject.

Let us look at the framework with all the main arguments from the three previous models:

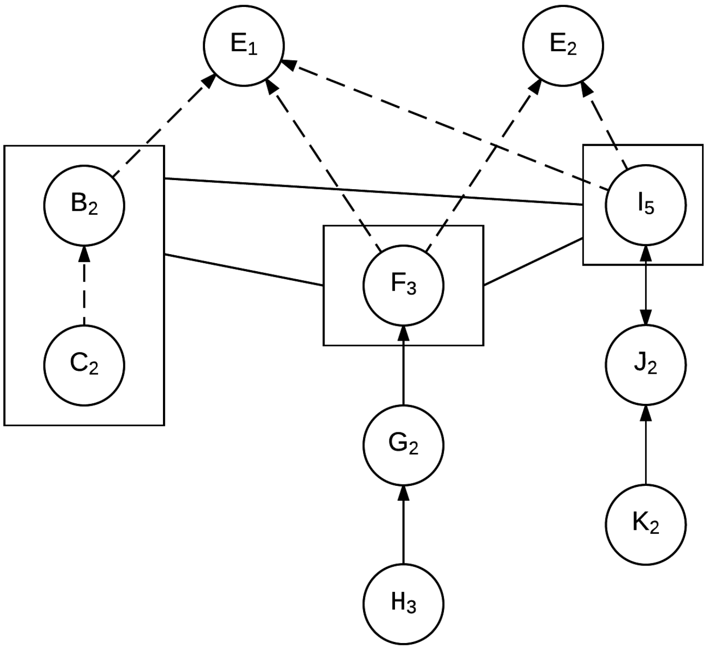

Figure 4.4: EAF representing the combined arguments in all three excerpts

Notice that we have collapsed all arguments from a given chain of arguments into one. For example, I_1, I_2 and I_5 have been merged into I_5.

In the EAFs, there is an incompatibility relation which allows to distinguish between the rivaling theories. We will define it here so as to separate the different solutions. Even though the solutions are technically not mutually exclusive, once a solution has been accepted, most paradoxes will be solved and hence the need for an explanation will be diminished. At that point, one is much less likely to accept another solution in conjunction with the previous one, even though this might be logically possible. We represent the different theories by grouping them in boxes on the graph. This means that we have three sets $Sol_1 = \{B_2, C_2\}$, $Sol_2 = \{F_4\}$ and $Sol_3 = \{I_5\}$, and each element of one set is incompatible with the elements of the other sets. For example, B_2 is incompatible with the elements of Sol_2 and Sol_3, F_4 is incompatible with the elements of Sol_1 and Sol_3, etc. Recall that the incompatibility relation does not represent a bi-directional attack but only the fact that they cannot be in the same conflict-free set and hence extension.

All three theories are fully defended and so the most relevant self-defending conflict-free sets are $S_1 = \{B_2, C_2, H_3, K_2\}$, $S_2 = \{B_2, C_2, H_3\}$, $S_3 = \{B_2, C_2, K_2\}$, $S_4 = \{B_2, C_2\}$, $S_5 = \{F_4, H_3, K_2\}$, $S_6 = \{F_4, H_3\}$, $S_7 = \{I_5, K_2, H_3\}$, $S_8 = \{I_5, K_2\}$. Let us now select the most explanatory ones. This removes all supersets of $\{B_2, C_2\}$ as they only provide an explanation for E_1 while the others also explain E_2. We have now completed all the common steps between the two selection procedures and end up with S_5, S_6, S_7, S_8 as extension candidates. Let us investigate the results of the first procedure. We then have to select the maximal sets. We then get S_5 and S_7. These would be the two extensions to consider from an argumentative point of view.

Continuing with the second procedure, we will now focus on the explanatory cores. After the common steps between the two procedures we get the sets S_5, S_6, S_7, S_8. We then have to select the explanatory deepest sets. All four of the candidate sets are either of the same depth or incomparable, hence we retain them all. Finally, we select the minimal sets with respect to set inclusion, which leaves us with S_6 and S_8 as the two extensions from the second procedure. These are the explanatory cores which contain only arguments which either form the explanations or directly defend them.

Chapter 5

Modeling arguments about the liar paradox using an extension of EAFs

In this chapter, we will introduce Extended Explanatory Argumentation Frameworks (EEAFs), an extension of EAFs from Section 2.2 with meta-argumentation features such as higher order attacks, support and joint attacks. We will then apply this EEAF approach to excerpts from the introduction of the book *Saving truth from paradox* by H. Field [6]. It discusses several families of solutions for Grelling's and Russell's paradoxes, which are very similar to the liar paradox. Since these excerpts feature reasoning on a more complex level, we believe using EEAFs to be more suited for this task than ASPIC-END. In this approach we will not be using structured argumentation and will instead be attempting to mine arguments directly from the text.

In order to motivate the semantics of EEAFs based on a flattening function, we will start by suggesting a flattening for a subset of AFRAs which we will call **grounded AFRAs**. We will prove that this flattening leads to the same extensions as the AFRA semantics defined by Baroni et al. [1].

5.1 Grounded AFRAs

In the definition of an AFRA in Section 2.3, notice that it is possible in an AFRA to have an attack relation which attacks itself, or two attack relations which attack each other. These two situations are depicted in Figure 5.1 and Figure 5.2 respectively.

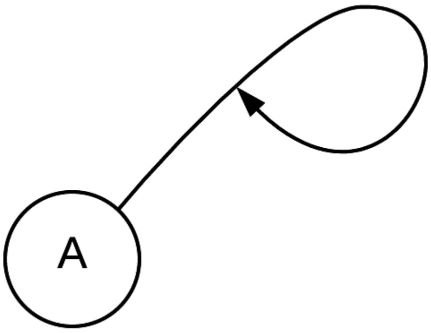

Figure 5.1: Ungrounded self-attacking attack relation

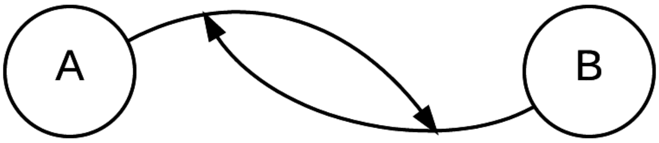

Figure 5.2: Ungrounded attacks attacking each other

These cases are not only unintuitive but also lead to problematic loops with regard to the flattening function we will define. Thus we introduce the notion of a **grounded AFRA**.

Definition 5.1.1. Grounded AFRA:
Let $F = \langle \mathcal{A}, \rightarrow \rangle$ be an argumentation framework with recursive attacks. We inductively define $(\varphi, \psi) \in \rightarrow$ to be *grounded* if and only if either $\psi \in \mathcal{A}$ or $\psi \in \rightarrow$ is a grounded attack. We say that F is *grounded* if and only if for all $\psi \in \rightarrow$, ψ is grounded.

Let us now focus on a flattening function for grounded AFRAs. Notice that for second order AFRAs, the names of the auxiliary arguments X and Y were always subscripted with the source and target arguments. Notice that for AFRAs of order higher than two, attacks can be nested an unbounded number of times and thus the names of the auxiliary arguments become more complex. The problem of loops that we mentioned earlier is that the arguments are the only elements with names, and in order to generate the names for auxiliary arguments representing the attacks, these attacks must lead to an argument at some point.

For the flattening function, we will define a function m which will associate each argument and each attack relation to the corresponding meta-argument. For an argument a, it will be the meta-argument $acc(a)$, while for an attack, it will be the Y auxiliary argument, since its acceptability is synonym of success of the attack.

Definition 5.1.2. Flattening of grounded AFRAs:
Let $F = \langle \mathcal{A}, \rightarrow \rangle$ be a grounded AFRA. The set of corresponding meta-arguments is $MA = \{acc(a) \mid a \in \mathcal{A}\} \cup \{X_{a,\psi}, Y_{a,\psi}\} \mid a \in \mathcal{A},\ \psi \in (\mathcal{A} \cup \rightarrow)\}$. We define a partial function m which assigns for each element of the framework a corresponding meta-argument.

$$m\colon (\mathcal{A} \cup \rightarrow) \mapsto MA.$$

such that:

- if $\varphi \in \mathcal{A}$, then $m(\varphi) = acc(\varphi)$.
- if $\varphi \in \rightarrow$ such that for some $\psi \in \mathcal{A}$ and some $\delta \in (\mathcal{A} \cup \rightarrow)$, $\varphi = (\psi, \delta)$, then $m(\varphi) = Y_{\psi,\delta}$.

We define the *flattening function* f to be $f(F) = \langle MA, \rightarrow_2 \rangle$, where $\rightarrow_2 \subseteq MA \times MA$ is a binary relations on MA such that

$$acc(a) \rightarrow_2 X_{a,\psi}, X_{a,\psi} \rightarrow_2 Y_{a,\psi} \text{ and } Y_{a,\psi} \rightarrow_2 m(\psi) \text{ for all } a \in \mathcal{A}, \psi \in (\mathcal{A} \cup \rightarrow)$$

One can then apply the classical abstract argumentation semantics such as complete, stable, preferred and grounded. We then need to define a function which can transform a meta-extension from the flattened grounded AFRA to an extension for the original grounded AFRA. The unneeded meta-arguments will need to be filtered out while the relevant ones will have to be converted back into regular arguments or attacks. A similar unflattening function has been introduced in [2], and has been slightly modified here to also unflatten attacks.

Definition 5.1.3. Unflattening function
Given a set of meta-arguments $B \subseteq MA$, we define the *unflattening function* g to be:

$$g(B) = \{a \mid acc(a) \in B\} \cup \{(a, \psi) \mid Y_{a,\psi} \in B\}$$

We also define a function $\bar{f}$ which provides a correspondence between a set of arguments and attacks from a grounded AFRA and a set of meta-arguments from its flattened version.

Definition 5.1.4. Correspondence function $\bar{f}$:
Let $F = \langle \mathcal{A}, \rightarrow \rangle$ be a grounded AFRA and $f(F) = \langle MA, \rightarrow_2 \rangle$ its flattening. We define the *correspondence function* $\bar{f}$ as follows:

$$\bar{f}:\ \mathcal{P}(\mathcal{A} \cup \rightarrow) \mapsto \mathcal{P}(MA)$$
$$\bar{f}(S) = \{acc(a) \mid a \in S \cap \mathcal{A}\} \cup \{Y_{a,\psi} \mid (a, \psi) \in S \cap \rightarrow\} \cup \{X_{b,\psi} \mid (a, b) \in S \cap \rightarrow,\ \psi \in \rightarrow\}$$

Notice that $g(\bar{f}(S)) = S$. We add the extra $X_{i,j}$ meta-arguments in order to represent the indirect attacks which the arguments in S might carry out, i.e. the attacks which are indirectly attacked by arguments in S due to them attacking the source of these attacks.

In [1], Baroni et al. define the semantics of AFRAs without having recourse to flattening. We will show that the process of flattening, applying complete semantics on the flattened frameworks and then unflattening it is equivalent to the directly applying the semantics they define for the complete semantics. We will show this gradually by first stating and proving three lemmas:

Lemma 5.1.1. Let $F = \langle \mathcal{A}, \rightarrow \rangle$ be a grounded AFRA, $f(F) = \langle MA, \rightarrow_2 \rangle$ and $S \subseteq (\mathcal{A} \cup \rightarrow)$. S is conflict-free in F if and only if $\bar{f}(S)$ is conflict-free in $f(F)$.

Proof:
Let $F = \langle \mathcal{A}, \rightarrow \rangle$ be a grounded AFRA, $f(F) = \langle MA, \rightarrow_2 \rangle$ and $S \subseteq (\mathcal{A} \cup \rightarrow)$.

1. We will first show that if S is conflict-free in F, then $\bar{f}(S)$ is conflict-free in $f(F)$. Assume that S is conflict-free in F. Then, there is no $\varphi, \psi \in S$ such that $trg(\varphi) = \psi$ or $trg(\varphi) = src(\psi)$. Suppose for a contradiction that $\bar{f}(S)$ is not conflict-free in $f(F)$. This means that there exists two arguments $p, q \in \bar{f}(S)$ such that $p \rightarrow_2 q$. By the construction of $\rightarrow_2$ defined by the flattening function, there are only four possible cases:
   (a) $p = acc(a)$ and $q = X_{a,\psi}$ for some $a \in \mathcal{A}$ and $\psi \in (\mathcal{A} \cup \rightarrow)$. Then, by the definition of $\bar{f}$, since $X_{a,\psi} \in \bar{f}(S)$, we have $(b, a) \in (S \cap \rightarrow)$ for some $b \in \mathcal{A}$. Also, since $acc(a) \in \bar{f}(S)$, we have $a \in S$. But (b, a) defeats a, and so S is not conflict-free.
   (b) $p = X_{a,\psi}$ and $q = Y_{a,\psi}$ for some $a \in \mathcal{A}$ and $\psi \in (\mathcal{A} \cup \rightarrow)$. Then, by the definition of $\bar{f}$, since $X_{a,\psi} \in \bar{f}(S)$, we have $(b, a) \in (S \cap \rightarrow)$ for some $b \in \mathcal{A}$. Also, since $Y_{a,\psi} \in \bar{f}(S)$, we have $(a, \psi) \in S$. But (b, a) defeats (a, ψ), and so S is not conflict-free.

(c) $p = Y_{a,b}$ and $q = acc(b)$ for some $a, b \in \mathcal{A}$. Then, by the definition of $\bar{f}$, $(a, b) \in S$ and $b \in S$. Thus, S is not conflict-free since (a, b) defeats b.

(d) $p = Y_{a,(b,\psi)}$ and $q = Y_{b,\psi}$ for some $a, b \in \mathcal{A}$ and $\psi \in (\mathcal{A} \cup \rightarrow)$. Then, by the definition of $\bar{f}$, $(a, (b, \psi)) \in S$ and $(b, \psi) \in S$. Since $trg(a, (b, \psi)) = (b, \psi)$, S is not conflict-free.

In each case, we get that S is not conflict-free. However, S being conflict-free is one of our assumptions, so we have a contradiction. Therefore $\bar{f}(S)$ is conflict-free.

2. We will now show that if $\bar{f}(S)$ is conflict-free in $f(F)$, then S is conflict-free in F. Suppose $\bar{f}(S)$ is conflict-free. Then, there is no $p, q \in MA$ such that $p \rightarrow_2 q$. Suppose for a contradiction that S is not conflict-free. Then, there exists $(a, \varphi), (b, \psi) \in S$ such that $\varphi = (b, \psi)$ or $\varphi = b$. Let us consider the two cases individually:

   (a) Suppose $\varphi = (b, \psi)$. Then, we have that $(a, (b, \psi)), (b, \psi) \in S$. By the definition of $\bar{f}$, this means that $Y_{a,(b,\psi)}, Y_{b,\psi} \in \bar{f}(S)$. However, by the construction of $\rightarrow_2$ defined by the flattening function, $Y_{a,(b,\psi)} \rightarrow_2 Y_{b,\psi}$. Hence, $\bar{f}(S)$ is not conflict-free. Since $\bar{f}(S)$ being conflict-free is one of our assumptions, we have a contradiction.

   (b) Suppose $\varphi = b$. Then, we have that $(a, b), (b, \psi) \in S$. By the definition of $\bar{f}$, this means that $Y_{a,b}, Y_{b,\psi} \in \bar{f}(S)$, but also that $X_{b,\psi} \in \bar{f}(S)$. However, by the construction of $\rightarrow_2$ defined by the flattening function, $X_{b,\psi} \rightarrow_2 Y_{b,\psi}$. Hence, $\bar{f}(S)$ is not conflict-free. Since $\bar{f}(S)$ being conflict-free is one of our assumptions, we have a contradiction.

   In both cases we have a contradiction, therefore S is conflict-free.

Hence, we can conclude that S is conflict-free in F if and only if $\bar{f}(S)$ is conflict-free in $f(F)$. □

Lemma 5.1.2. Let $F = \langle \mathcal{A}, \rightarrow \rangle$ be a grounded AFRA, $f(F) = \langle MA, \rightarrow_2 \rangle$, $\varphi \in (\mathcal{A} \cup \rightarrow)$ and $S \subseteq (\mathcal{A} \cup \rightarrow)$. We have that:

φ is acceptable with respect to S in F and if $\varphi = (a, \psi) \in \rightarrow$, we have $a \in S$
if and only if
$m(\varphi)$ is defended by $\bar{f}(S)$ in $f(F)$ and if $\varphi \in \rightarrow$, then $acc(src(\varphi))$ is also defended by $\bar{f}(S)$.

Proof:
Let $F = \langle \mathcal{A}, \rightarrow \rangle$ be a grounded AFRA, $f(F) = \langle MA, \rightarrow_2 \rangle$, $\varphi \in (\mathcal{A} \cup \rightarrow)$ and $S \subseteq (\mathcal{A} \cup \rightarrow)$.

1. We will first show that if φ is acceptable with respect to S in F and if $\varphi = (a, \psi) \in \rightarrow$, we have $a \in S$, then $m(\varphi)$ is defended by $\bar{f}(S)$ in $f(F)$ and if $\varphi \in \rightarrow$, then $acc(src(\varphi))$ is also defended by $\bar{f}(S)$.
   Suppose that φ is acceptable with respect to S in F and if $\varphi = (a, \psi) \in \rightarrow$, we have $a \in S$. Then, for all $\psi \in \rightarrow$ such that ψ defeats φ, there exists some $\delta \in S$ such that δ defeats ψ. Now consider $m(\varphi)$ in $f(F)$. Suppose for some $p \in MA$, $p \rightarrow_2 m(\varphi)$. By the construction of $\rightarrow_2$ defined by the flattening function, either $p = Y_{a,\varphi}$ for some $a \in \mathcal{A}$ or $p = X_{src(\varphi),trg(\varphi)}$. The second case is possible only if $\varphi \in \rightarrow$. Let us examine the two cases separately:

(a) Suppose $p = Y_{a,\varphi}$ for some $a \in \mathcal{A}$. We have that (a, φ) defeats φ in F and thus there exists some $\delta \in S$ such that δ defeats (a, φ). We distinguish two cases:

   i. Suppose $\delta = (b, a)$ for some $b \in \mathcal{A}$. This means that $Y_{b,a} \in \bar{f}(S)$ and that $X_{a,\varphi} \in \bar{f}(S)$. By the construction of $\rightarrow_2$ defined by the flattening function, $X_{a,\varphi} \rightarrow_2 Y_{a,\varphi}$ and thus $m(\varphi)$ is defended by $\bar{f}(S)$.

   ii. Suppose $\delta = (b, (a, \varphi))$ for some $b \in \mathcal{A}$. This means that $Y_{b,(a,\varphi)} \in \bar{f}(S)$. By the construction of $\rightarrow_2$ defined by the flattening function, $Y_{b,(a,\varphi)} \rightarrow_2 Y_{a,\varphi}$ and thus $m(\varphi)$ is defended by $\bar{f}(S)$.

(b) Now suppose $\varphi \in \rightarrow$ and $p = X_{src(\varphi),trg(\varphi)}$. Then, $src(\varphi) \in S$ and thus $acc(src(\varphi)) \in \bar{f}(S)$. By the construction of $\rightarrow_2$ defined by the flattening function, $acc(src(\varphi)) \rightarrow_2 X_{src(\varphi),trg(\varphi)}$ and thus $m(\varphi) = Y_{src(\varphi),trg(\varphi)}$ is defended by $\bar{f}(S)$.

In both cases, $m(\varphi)$ is defended by $\bar{f}(S)$. Hence, if φ is acceptable with respect to S in F, then $m(\varphi)$ is defended by $\bar{f}(S)$ in $f(F)$. We now have to show that if $\varphi \in \rightarrow$, then $acc(src(\varphi))$ is also defended by $\bar{f}(S)$.
Suppose $\varphi \in \rightarrow$ and $p \in MA$ such that $p \rightarrow_2 acc(src(\varphi))$. Then, p must be of the form $Y_{a,src(\varphi)}$ for some $a \in \mathcal{A}$, and hence there exists $(a, src(\varphi)) \in \rightarrow$. Since $(a, src(\varphi))$ defeats φ, there exists some $\delta \in S$ such that δ defeats $(a, src(\varphi))$. We distinguish two cases:

- Suppose $\delta = (b, a)$ for some $b \in \mathcal{A}$. This means that $Y_{b,a} \in \bar{f}(S)$ and that $X_{a,src(\varphi)} \in \bar{f}(S)$. By the construction of $\rightarrow_2$ defined by the flattening function, $X_{a,src(\varphi)} \rightarrow_2 Y_{a,src(\varphi)}$ and thus $acc(src(\varphi))$ is defended by $\bar{f}(S)$.
- Suppose $\delta = (b, (a, src(\varphi)))$ for some $b \in \mathcal{A}$. This means that $Y_{b,(a,src(\varphi))} \in \bar{f}(S)$. By the construction of $\rightarrow_2$ defined by the flattening function, $Y_{b,(a,src(\varphi))} \rightarrow_2 Y_{a,src(\varphi)}$ and thus $acc(src(\varphi))$ is defended by $\bar{f}(S)$.

Hence, $acc(src(\varphi))$ is also defended by $\bar{f}(S)$.
Therefore, if φ is acceptable with respect to S in F and if $\varphi = (a, \psi) \in \rightarrow$, we have $a \in S$, then $m(\varphi)$ is defended by $\bar{f}(S)$ in $f(F)$ and if $\varphi \in \rightarrow$, then $acc(src(\varphi))$ is also defended by $\bar{f}(S)$.

2. We will now show that if $m(\varphi)$ is defended by $\bar{f}(S)$ in $f(F)$ and if $\varphi \in \rightarrow$, then $acc(src(\varphi))$ is defended by $\bar{f}(S)$ also, then φ is acceptable with respect to S in F and if $\varphi = (a, \psi) \in \rightarrow$, we have $a \in S$.
Suppose $m(\varphi)$ is defended by $\bar{f}(S)$ in $f(F)$ and if $\varphi \in \rightarrow$, then $acc(src(\varphi))$ is also defended by $\bar{f}(S)$. So, for each $p \in MA$ such that $p \rightarrow_2 m(\varphi)$, we have that there exists a $q \in \bar{f}(S)$ such that $q \rightarrow_2 p$. Now consider φ in F. Suppose that for some $\psi \in \rightarrow$, ψ defeats φ. This means that either $\psi = (a, \varphi)$ or $\psi = (a, src(\varphi))$ for some $a \in \mathcal{A}$. Let us consider both cases individually:

(a) Assume $\psi = (a, \varphi)$. Suppose for a contradiction that there is no $\delta \in S$ such that δ defeats ψ. Then, by the construction of $\rightarrow_2$ defined by the flattening function, there exists no $Y_{src(\delta),(a,\varphi)} \in \bar{f}(S)$ nor $X_{a,\varphi} \in \bar{f}(S)$. Hence, there is no $s \in \bar{f}(S)$ such that $s \rightarrow_2 Y_{a,\varphi}$ and thus $m(\varphi)$ is not defended by $\bar{f}(S)$. However, one of our assumptions is that $m(\varphi)$ is defended by $\bar{f}(S)$ and thus we have a contradiction. Hence, there exists a $\delta \in S$ such that δ defeats ψ.

(b) Assume $\psi = (a, src(\varphi))$. Suppose for a contradiction that there is no $\delta \in S$ such that δ defeats ψ. Then, by the construction of $\rightarrow_2$ defined by the flattening function, there exists no $Y_{src(\delta),(a,src(\varphi))} \in \bar{f}(S)$ nor $X_{a,src(\varphi)} \in \bar{f}(S)$. Hence, $acc(src(\varphi))$ is not defended by $\bar{f}(S)$. However, we have that $acc(src(\varphi))$ is defended by $\bar{f}(S)$. Hence, we have a contradiction, and therefore there exists a $\delta \in S$ such that δ defeats ψ.

Hence, we can conclude that there exists a $\delta \in S$ such that δ defeats ψ. Therefore, φ is acceptable with respect to S.

We now have to show that if $\varphi = (a, \psi) \in \rightarrow$, we have $a \in S$, still under the assumption that $m(\varphi)$ is defended by $\bar{f}(S)$ in $f(F)$ and if $\varphi \in \rightarrow$, then $acc(src(\varphi))$ is also defended by $\bar{f}(S)$.
Suppose that $\varphi = (a, \psi) \in \rightarrow$. Then, by the construction of $\rightarrow_2$ defined by the flattening function, we have $X_{a,\psi} \rightarrow_2 Y_{a,\psi}$. Since $m(\varphi) = Y_{a,\psi}$ is defended by $\bar{f}(S)$, there exists $p \in \bar{f}(S)$ such that $p \rightarrow_2 X_{a,\psi}$. By the construction of $\rightarrow_2$, the only possibility is $p = acc(a)$. Hence, $acc(a) \in \bar{f}(S)$. Therefore, we have $a \in S$.

Thus, we can conclude that φ is acceptable with respect to S in F and if $\varphi = (a, \psi) \in \rightarrow$, we have $a \in S$, if and only if $m(\varphi)$ is defended by $\bar{f}(S)$ in $f(F)$ and if $\varphi \in \rightarrow$, then $acc(src(\varphi))$ is also defended by $\bar{f}(S)$.□

Lemma 5.1.3. Let $F = \langle \mathcal{A}, \rightarrow \rangle$ be a grounded AFRA, $f(F) = \langle MA, \rightarrow_2 \rangle$ and $S \subseteq (\mathcal{A} \cup \rightarrow)$. We have that:

$$\begin{gathered} S \text{ is admissible in } F \text{ and for every } (a, \psi) \in (S \cap \rightarrow), \text{ we have that } a \in S \\ \text{if and only if} \\ \bar{f}(S) \text{ is admissible in } f(F). \end{gathered}$$

Proof: Let $F = \langle \mathcal{A}, \rightarrow \rangle$ be a grounded AFRA, $f(F) = \langle MA, \rightarrow_2 \rangle$ and $S \subseteq (\mathcal{A} \cup \rightarrow)$.

1. We will first show that if $\bar{f}(S)$ is admissible in $f(F)$, then S is admissible in F and for every $(a, \psi) \in (S \cap \rightarrow)$, we have that $a \in S$.
Suppose $\bar{f}(S)$ is admissible in $f(F)$. Then, $\bar{f}(S)$ is conflict-free. Hence, according to Lemma 5.1.1, S is also conflict-free.
Let $\varphi \in S$. We need to show that φ is acceptable with respect to S. We do this by applying Lemma 5.1.2, i.e. by establishing that $m(\varphi)$ is defended by $\bar{f}(S)$ in $f(F)$ and if $\varphi \in \rightarrow$, then $acc(src(\varphi))$ is also defended by $\bar{f}(S)$. We have $m(\varphi) \in \bar{f}(S)$ and $m(\varphi)$ is defended by $\bar{f}(S)$ since $\bar{f}(S)$ is admissible. By the definition of $\bar{f}$, for every $(a, \psi) \in (S \cap \rightarrow)$, we have $Y_{a,\psi} \in \bar{f}(S)$. Therefore, $acc(a) \in \bar{f}(S)$, since it is the only argument which can defend $Y_{a,\psi}$ from $X_{a,\psi}$'s attack and $\bar{f}(S)$ is admissible. This means that $acc(a)$ is defended by $\bar{f}(S)$. Thus, according to Lemma 5.1.2, every $\varphi \in S$ is acceptable with respect to S, which means that S is admissible, and for every $(a, \psi) \in (S \cap \rightarrow)$, we have that $a \in S$.

2. We now have to show that if S is admissible in F and for every $(a, \psi) \in (S \cap \rightarrow)$, we have that $a \in S$, then $\bar{f}(S)$ is admissible in $f(F)$.
Suppose S is admissible in F and for every $(a, \psi) \in (S \cap \rightarrow)$, we have that $a \in S$. Then, S is conflict-free and so, according to Lemma 5.1.1, $\bar{f}(S)$ is also conflict-free. To conclude that $\bar{f}(S)$ is admissible, we still need to show that for every $p \in \bar{f}(S)$, $\bar{f}(S)$ defends p. So let $p \in \bar{f}(S)$. p is either of the form $m(\varphi)$ for some $\varphi \in S$, or of the form $X_{a,b}$ for some $a, b \in MA$ and $(\psi, a) \in S$. We treat these two cases separately.

(a) Suppose $p = m(\varphi)$ for some $\varphi \in S$. φ is acceptable with respect to S and if $\varphi = (a, \psi)$ for some $a \in \mathcal{A}$ and some $\psi \in \rightarrow$, then $a \in S$. Hence, according to Lemma 5.1.2, $p = m(\varphi)$ is defended by $\bar{f}(S)$.

(b) Suppose $p = X_{a,b}$ for some $a, b \in MA$ and $(\psi, a) \in S$. By construction, $X_{a,b} \in \bar{f}(S)$ means that $(c, a) \in S \cap \rightarrow$ for some $c \in \mathcal{A}$. Therefore, $Y_{c,a} \in \bar{f}(S)$ and $Y_{c,a} \rightarrow_2 acc(a)$. Since by construction, the only argument attacking $X_{a,b}$ is $acc(a)$, $Y_{c,a} \in \bar{f}(S)$ defends $X_{a,b}$. Therefore, $p = X_{a,b}$ is defended by $\bar{f}(S)$.

In both cases, p is defended by $\bar{f}(S)$. Hence, $\bar{f}(S)$ is admissible in $f(F)$.

Therefore, S is admissible in F and for every $(a, \psi) \in (S \cap \rightarrow)$, we have that $a \in S$, if and only if $\bar{f}(S)$ is admissible in $f(F)$.□

Theorem 5.1.1. Let $F = \langle \mathcal{A}, \rightarrow \rangle$ be a grounded AFRA, $f(F) = \langle MA, \rightarrow_2 \rangle$ and $S \subseteq (\mathcal{A} \cup \rightarrow)$. S is a complete extension of F if and only if $\bar{f}(S)$ is a complete extension of $f(F)$.

Proof:
Let $F = \langle \mathcal{A}, \rightarrow \rangle$ be a grounded AFRA, $f(F) = \langle MA, \rightarrow_2 \rangle$ and $S \subseteq (\mathcal{A} \cup \rightarrow)$.

1. We will first show that if S is a complete extension of F, then $\bar{f}(S)$ is a complete extension of $f(F)$.
Suppose S is a complete extension of F. Then, S is admissible and for every $a \in \mathcal{A}$ such that a is acceptable with respect to S, we have $a \in S$. For every $(a, \psi) \in (S \cap \rightarrow)$, if $\delta \in \rightarrow$ defeats a, then δ defeats (a, ψ). Hence, since (a, ψ) is acceptable with respect to S, we have that a is acceptable with respect to S. Hence, since S is a complete extension, we have $a \in S$. So we have that S is admissible and for all $(a, \psi) \in (S \cap \rightarrow)$, we have that $a \in S$. Therefore, by Lemma 5.1.3, $\bar{f}(S)$ is admissible.

   Take some arbitrary $p \in MA$ and suppose that p is defended by $\bar{f}(S)$. We have two cases, either $p = m(\varphi)$ for some $\varphi \in (\mathcal{A} \cap \rightarrow)$, or $p = X_{a,b}$ for some $a, b \in \mathcal{A}$. Let us consider both cases individually:

   (a) Suppose that $p = m(\varphi)$ for some $\varphi \in (\mathcal{A} \cap \rightarrow)$. Now assume that $\varphi \in \rightarrow$. Then, $m(\varphi) = Y_{src(\varphi),trg(\varphi)}$. By construction of $\rightarrow_2$, we have that $X_{src(\varphi),trg(\varphi)} \rightarrow_2 Y_{src(\varphi),trg(\varphi)}$. The only argument which can defend $Y_{src(\varphi),trg(\varphi)}$ from $X_{src(\varphi),trg(\varphi)}$ is $acc(src(\varphi))$. Since $\bar{f}(S)$ defends $Y_{src(\varphi),trg(\varphi)}$, we have that $acc(src(\varphi)) \in \bar{f}(S)$. As $\bar{f}(S)$ is admissible, $acc(src(\varphi))$ is defended by $\bar{f}(S)$. Hence, if $\varphi \in \rightarrow$, then $acc(src(\varphi))$ is defended by $\bar{f}(S)$.
   Therefore, by Lemma 5.1.2, φ is acceptable with respect to S. Since S is a complete extension, this means that $\varphi \in S$. Therefore, $p = m(\varphi) \in \bar{f}(S)$.

   (b) Now suppose that $p = X_{a,b}$ for some $a, b \in \mathcal{A}$. According to our assumptions, $\bar{f}(S)$ defends $X_{a,b}$. By construction of $\Rightarrow_2$, the only argument which attacks $X_{a,b}$ is $acc(a)$. Hence, there exists $Y_{c,a} \in \bar{f}(S)$ for some $c \in \mathcal{A}$. So, by definition of $\bar{f}$, we have that $p = X_{a,b} \in \bar{f}(S)$.

   In either case, we have that $p \in \bar{f}(S)$. Hence, $\bar{f}(S)$ contains all arguments it defends. Since it is also admissible, $\bar{f}(S)$ is a complete extension of $f(F)$.

2. We will now show that if $\bar{f}(S)$ is a complete extension of $f(F)$, then S is a complete extension of F.
Suppose that $\bar{f}(S)$ is a complete extension of $f(F)$. Then, $\bar{f}(S)$ is admissible and

contains all arguments it defends. According to Lemma 5.1.3, we have that S is admissible and for every $(a, \psi) \in (S \cap \rightarrow)$, we have that $a \in S$. Suppose that for some $\varphi \in (\mathcal{A} \cup \rightarrow)$, φ is acceptable with respect to S. Hence, by Lemma 5.1.2, $m(\varphi)$ is defended by $\bar{f}(S)$. Since $\bar{f}(S)$ is a complete extension of $f(F)$, $m(\varphi) \in \bar{f}(S)$. Hence, by construction of $\bar{f}(S)$, we have that $\varphi \in S$. Therefore, for any $\varphi \in (\mathcal{A} \cup \rightarrow)$ such that φ is acceptable with respect to S, we have $\varphi \in S$. Since S is also admissible, S is a complete extension of F.

Hence, S is a complete extension of F if and only if $\bar{f}(S)$ is a complete extension of $f(F)$.□

5.2 Extended Explanatory Argumentation Frameworks

We will now extend EAFs, as seen in Section 2.2, by integrating it with multiple meta-argumentation techniques. In order to produce a model of the reasoning present in the excerpts which captures as closely as possible the author's reasoning, we require that our formalism have enough expressive power. We need the explanatory relation as the reasoning is about different solutions which aim to explain a given paradox and hence we need our formalism to feature some measure of explanatory power and depth.

Also, we require higher-order attacks in order to be able to express some subtleties in the reasoning. For example, we might wish to be able to represent the fact that some argument A actually fails to explain a certain explanandum E. In such a case, the argument A might be sound and fine on its own, and thus attacking it would not capture the idea of failure to explain E. Hence why we require that in our formalism some argument may attack an explanation or even attack relation.

We also require a relation of support separate from the explanatory relation between arguments. An argument explains another argument by adding depth to the its explanation. Hence, the explanatory relation between arguments only makes sense if it forms a chain which leads to an explanandum. On the other hand, an argument supports another by deductively concluding it. The relation of support is not tied in any way to the explananda and the measures of explanatory power and depth. Also, notice that if a explains b, c attacking b has no influence on a, while if a supports b, then c attacking b will result in an indirect attack on a. This will be apparent in the flattening.

Definition 5.2.1. EEAF:

An *extended explanatory argumentation framework* (EEAF) is a tuple $\langle \mathcal{A}, \mathcal{X}, \rightarrow, \dashrightarrow, \sim, \Rightarrow \rangle$, where $\mathcal{A}$ is a set of arguments, $\mathcal{X}$ is a set of explananda, $\dashrightarrow \subseteq (\mathcal{A} \times \mathcal{A}) \cup (\mathcal{A} \times \mathcal{X})$ is an explanatory relation, $\rightarrow \subseteq (\mathbb{P}(\mathcal{A}) \cup \dashrightarrow \cup \rightarrow) \times (\mathcal{A} \cup \dashrightarrow \cup \rightarrow \cup \Rightarrow)$ is a grounded higher-order attack relation, $\sim \subseteq \mathcal{A} \times \mathcal{A}$ is an incompatibility relation and $\Rightarrow \subseteq \mathcal{A} \times \mathcal{A}$ is a support relation. $\rightarrow$ is a grounded higher-order attack relation means that every element of $\rightarrow$ is grounded, which we define inductively as follows: $(\varphi, \psi) \in \rightarrow$ is grounded if and only if either $\psi \in (\mathcal{P}(\mathcal{A}) \cup \dashrightarrow)$ or $\psi \in \rightarrow$ is grounded.

We then define the semantics of *EEAFs* in terms of their flattening.

Definition 5.2.2. Flattening of EEAFs:

Let $F = \langle \mathcal{A}, \mathcal{X}, \rightarrow, \dashrightarrow, \sim, \Rightarrow \rangle$ be an EEAF. The set of meta-arguments corresponding to F is $MA = \{acc(a), rej(a) \mid a \in \mathcal{A}\} \cup \{X_{m(\varphi),m(\psi)}, Y_{m(\varphi),m(\psi)} \mid \varphi \in (\mathcal{A} \cup \rightarrow \cup \dashrightarrow), \psi \in (\mathcal{A} \cup \rightarrow \cup \dashrightarrow \cup \Rightarrow)\} \cup \{e(S) \mid S \subseteq \mathcal{A}$ with at least two elements$\} \cup \{P_{a,\psi}, Q_{a,\psi} \mid a \in \mathcal{A},\ \psi \in (\mathcal{A} \cup \mathcal{X})\} \cup \{Z_{a,b} \mid a, b \in \mathcal{A}\}$ and the set of meta-explananda is $MX = \mathcal{X}$. We define a partial function m which assigns for each element of the framework a corresponding meta-argument.

$$m\colon (\mathcal{A} \cup \rightarrow \cup \dashrightarrow \cup \Rightarrow) \mapsto MA.$$

such that:

- if $\varphi \in \mathcal{U}$, then $m(\varphi) = acc(\varphi)$.
- if $\varphi \in \Rightarrow$ such that for some $a, b \in \mathcal{A}$, $\varphi = (a \Rightarrow b)$, then $m(\varphi) = Z_{a,b}$.
- if $\varphi \in \rightarrow$ such that for some $S \subseteq \mathcal{A}$ with at least two elements and some $\psi \in (\mathcal{A} \cup \dashrightarrow \cup \rightarrow \cup \Rightarrow)$, $\varphi = (S \rightarrow \psi)$, then $m(\varphi) = e(S)$.
- if $\varphi \in \rightarrow$ such that for some $\psi \in (\mathcal{A} \cup \dashrightarrow \cup \rightarrow)$ and some $\delta \in (\mathcal{A} \cup \dashrightarrow \cup \rightarrow \cup \Rightarrow)$, $\varphi = (\psi \rightarrow \delta)$, then $m(\varphi) = Y_{\psi,\delta}$.
- if $\varphi \in \dashrightarrow$ such that for some $a \in \mathcal{A}$ and $\psi \in (\mathcal{A} \cup \mathcal{X})$, $\varphi = (a \dashrightarrow \psi)$, then $m(\varphi) = P_{a,\psi}$.

We define the *flattening function* f to be $f(F) = \langle MA, \mathcal{X}, \rightarrow_2, \dashrightarrow_2, \sim \rangle$, where $\rightarrow_2, \dashrightarrow_2 \subseteq MA \times MA$ are binary relations on MA such that

- $X_{m(\varphi),m(\psi)} \rightarrow_2 Y_{m(\varphi),m(\psi)}, Y_{m(\varphi),m(\psi)} \rightarrow_2 m(\psi)$ for all $\varphi, \psi \in (\mathcal{A} \cup \rightarrow \cup \dashrightarrow \cup \Rightarrow)$
- $m(\varphi) \rightarrow_2 X_{m(\varphi),m(\psi)}$ if and only if $\varphi \rightarrow \psi$ and φ is not a set of arguments with at least two elements
- $acc(a) \rightarrow_2 rej(a)$ for all $a \in \mathcal{A}$
- $e(S) \rightarrow_2 m(\varphi)$ if and only if $S \rightarrow \varphi$ for $S \subseteq \mathcal{A}$ with at least 2 elements
- $rej(a) \rightarrow_2 e(S)$ if and only if $a \in S$
- $Z_{a,b} \rightarrow_2 acc(a)$ for all $a, b \in \mathcal{A}$
- $acc(b) \rightarrow_2 Z_{a,b}$ if and only if $a \Rightarrow b$
- $acc(a) \dashrightarrow_2 P_{a,\varphi}$, $P_{a,\varphi} \dashrightarrow_2 m(\varphi)$, $acc(a) \rightarrow_2 Q_{a,\varphi}$ and $Q_{a,\varphi} \rightarrow_2 P_{a,\varphi}$ if and only if $a \dashrightarrow \varphi$.

Notice that the set of meta-arguments MA and the correspondence function m are defined through a simultaneous inductive definition which terminates because $\rightarrow$ is grounded.

Note that we do not fully flatten the explanatory relation and flatten EEAFs into EAFs and not classical abstract argumentation frameworks. This is due to the fact that the explanatory relation is not easily flattened, and extensions can still be extracted from explanatory argumentation frameworks via the two selection procedures which are well-suited for our task. In order to do this, we need to define an unflattening function which will map a set of meta-arguments from a flattened EEAF to the corresponding set of arguments from the original EEAF.

Definition 5.2.3. Unflattening function for EEAFs:
Given an EEAF F and a set of meta-arguments $B \subseteq MA$ such that MA corresponds to F, we define the *unflattening function* g to be:

$$g(B) = \{a \mid acc(a) \in B\}$$

Notice that in the unflattening, we only care about the arguments and do not unflatten the meta-arguments which represent the other elements of EEAFs. This is due to the fact that we are only interested in selecting the arguments of the EEAF, which make up the argumentative and explanatory cores. Also taking into consideration other elements the EEAF would not add much information while adding possible unnecessary confusions.

We then describe the two argument selection procedures as follows:

Procedure 1:

1. Flatten the framework F into $M = f(F)$.
2. Select the conflict-free sets in M.
3. Out of those, select the defended ones.
4. Out of those, select the most explanatory powerful ones.
5. Out of those, select the maximal ones with respect to $\subseteq$ and call the resulting set AC.
6. For each $S \in AC$, compute and output $g(S)$.

This first procedure selects the argumentative cores of the EEAF while the next procedure will select the explanatory cores of the EEAF.

Procedure 2:

1. Flatten the framework F into $M = f(F)$.
2. Select the conflict-free sets in M.
3. Out of those, select the defended ones.
4. Out of those, select the most explanatory powerful ones.
5. Out of those, select the most explanatory deep ones.
6. Out of those, select the minimal ones with respect to $\subseteq$ and call the resulting set AC.
7. For each $S \in AC$, compute and output $g(S)$.

5.3 Applying EEAFs to the liar paradox

We can now proceed to formalizing the excerpts using *EEAFs*.

The first extract we will examine is taken from the introduction of [6] and focuses on the Russell property. There is a principle which the author refers to as (INST), where F is some intelligible predicate:

(INST) "The property of being F is instantiated by all and only those things that are F."

Now the Russell property is the property of not instantiating itself. By plugging the Russell property in INST, we get:

"The Russell property is instantiated by all and only those things that don't instantiate themselves."

Which can also be rephrased as:

"The Russell property instantiates itself if and only if it does not instantiate itself."

This is the Russell paradox, which is similar in many ways to the liar paradox as it has the form $B \leftrightarrow \neg B$. Let us now examine excerpt 4, which can be found in the appendix.

We have the following arguments:

1. E_p: this is an explanandum which represents Russell's paradox.
2. A: The Non-existence solution, which suggests that there is no such property as the Russell property.
3. B: One could argue that it would violate the raison d'être of properties to suppose that for an intelligible predicate such as 'doesn't instantiate itself', there is no corresponding property of not instantiating itself.
4. C: As an answer to this, one could deny that the Russell property is intelligible.
5. D: It seems odd to say that the property of not instantiating itself is not intelligible as all parts of it are intelligible.
6. E: By defining intelligible as "expresses a property", one can deny that the Russell property is intelligible.

These arguments give us the following framework:

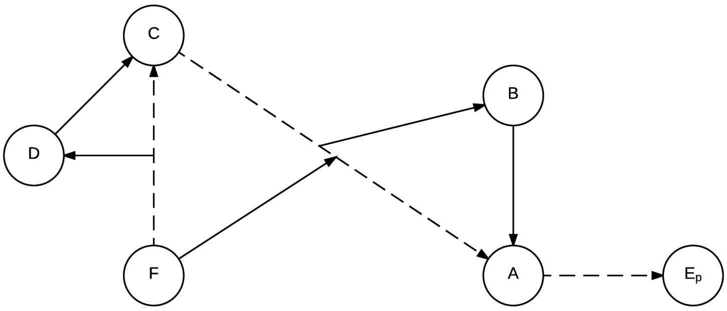

Figure 5.3: EEAF representing the reasoning behind excerpt 4

The non-existence solution A explains the paradox E_p and is attacked by the argument B that it violates the raison dêtre of properties to suppose that such a property does not exist. B is in turn attacked by the argument C that the property is not -intelligible, which also deepens A's explanation. The argument D then states that all parts of 'does not instantiates itself' are intelligible and thus attacks C. We then have the argument F that 'intelligible should be read as 'expresses a property". This attacks D and also adds to the explanatory depth of C as it explains the term 'intelligible' used in C. However, notice that F also attacks the explanatory relation from C to A as the definition of 'intelligible' and 'express a property' given by the solution are now defined in terms of each other and can thus never be fully settled. This means that the solution has not explained the failure of the property (COMP). As a consequence, F also attacks the attack from C to

B as the argument of non-intelligibility does now not seem solid enough to warrant an attack on B.

The framework can be flattened as follows:

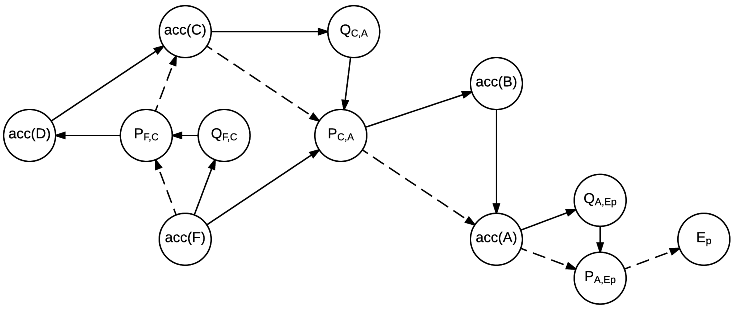

Figure 5.4: Flattened EEAF representing the reasoning behind Excerpt 4

Notice that we have simplified the flattened framework for the purpose of visibility. We have omitted the auxiliary arguments $X_{a,b}$ and $Y_{a,b}$ in the cases where (a, b) being not attacked or the origin of an attack, as including them would then have no impact on the final extensions.

We now apply the first selection procedure. The first step, the flattening, has already been done. Steps 2 and 3 give us the conflict-free and self-defended sets, of which the most relevant are $\emptyset$, $\{acc(F)\}$, $\{acc(F), P_{F,C}, acc(C)\}$, $\{acc(F), acc(B)\}$ and $\{acc(F), P_{F,C}, acc(C), acc(B)\}$. Since none of them contain an explanation for E_p, they are all equally as explanatory powerful and hence step 3 changes nothing. Step 4 makes us select the maximal ones, and there is one set which is a superset of all the others, the set $\{acc(F), P_{F,C}, acc(C), acc(B)\}$. $g(\{acc(F), P_{F,C}, acc(C), acc(B)\}) = \{F, C, B\}$ and thus the argumentative core is $\{F, C, B\}$.

The second procedure gives us that the explanatory core is $\emptyset$. This is due to the fact that the only argument explaining the explanandum E_p is A, yet A is attacked by B which is defended by the unattacked argument F. Hence, A can never be defended and thus we can extract no relevant explanation from this framework.

Let us now examine excerpt 5. Here, the author is focusing on the paracomplete solutions. The paracomplete solution reject the principle of excluded middle which states that for every formula φ, it always holds that $\varphi \vee \neg\varphi$. We have the following arguments:

- E_p: This explanandum represents once again the paradox.
- A: The paracomplete solution explains the paradox by rejecting the law of excluded middle.
- B: Why would one reject the law of excluded middle when it seems sound in mathematics, physics etc.
- C: The paracomplete solution only question its applicability to certain circular predicates such as this paradox.
- D: An interesting paracomplete theory in which the Naive Property Theory is consistent might not even be possible since intuitionist logic invalidates the central argument from equivalence to contradiction but still allows for contradictions from a formula such as $B \leftrightarrow \neg B$.

- F: In deMorgan logics without LEM, $B \leftrightarrow \neg B$ is not contradictory.

- G: $B \leftrightarrow \neg B$ not being contradictory is not enough, we also need to maintain Naive Property Theory and include intersubstitutivity of equivalents.

- H: Intersubstitutivity of equivalents follows from (INST) in classical logic.

- I: We are considering logics weaker than classical logic in which it may not follow from (INST).

- J: In the reasonably strong deMorgan logic advocated later in the book, (INST) holds.

We get the following framework:

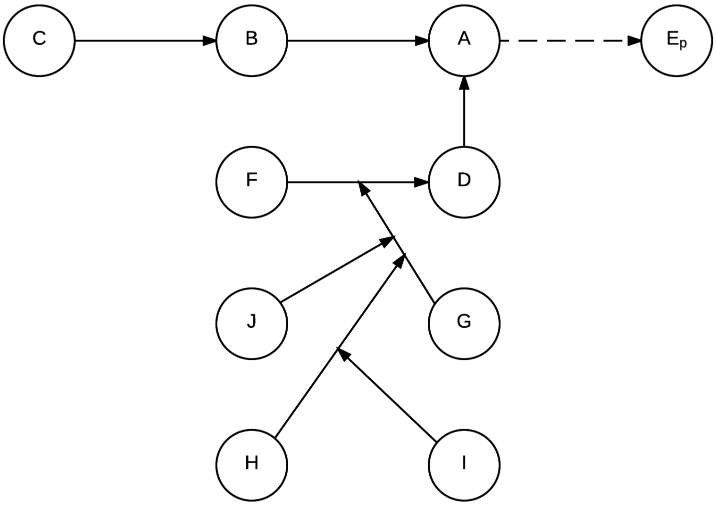

Figure 5.5: EEAF representing the reasoning behind Excerpt 5

The framework gets flattened into:

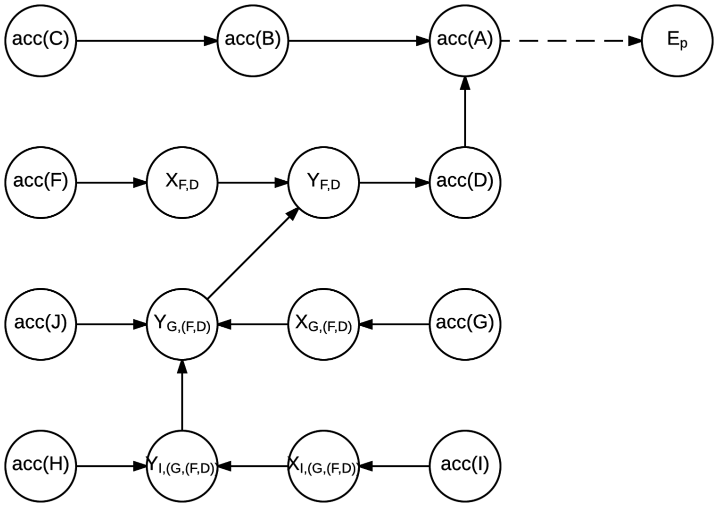

Figure 5.6: Flattened EEAF representing the reasoning behind Excerpt 5

Notice that once again, for the sake of visibility, we have omitted unnecessary auxiliary arguments $X_{a,b}, Y_{a,b}$ when $(a, b) \in \rightarrow$ is not being attacked or attacking another element of the framework, and similarly for $P_{a,e}, Q_{a,e}$ when $(a, e) \in \dashrightarrow$ is not being attacked or attacking.

By applying the two selection procedures, we get that the argumentative core is $\{A, C, F, G, H, I, J\}$ while the explanatory core is $\{A, C, F, J\}$. This means that the solution modeled in the excerpt seems to be consistent with the proposed arguments F, G, H and I, even if they do not directly contribute to the explanation of the solution. In the end, the solution A is defended by C from B and by F from D, which is then defended by J from G. Hence, the four arguments A, C, F, J are essential and sufficient to defend the solution in this model. Note that no arguments are explaining each other, hence we cannot measure explanatory depth in this model.

Let us now move on to excerpt 6, which focuses on two groups of solutions. The first group is the solutions which weaken classical logic, namely the paracomplete, paraconsistent and semi-classical solutions. The second group is comprised of the underspill and overspill solutions.

We have the following arguments:

- E_p: Again, this explanandum represents the paradox.
- A: The paracomplete, paraconsistent and semi-classical solutions which provide explanations for the paradox by weakening classical logic.
- B: The underspill and overspill solutions which provide their own explanation of the paradox by suggesting that for some predicates F, F is true of some objects that aren't F or vice-versa.
- C: We did not change logic to hide the defects in other flawed theories such as Ptolemaic astronomy, so why should we change the logic simply to hide these paradoxes?
- D: There is no known way of saving these flawed theories such as Ptolemaic astronomy and even if there was, there is little benefit to doing so.

- *F*: We have worked out the details of the new logics and they allow us to conserve the theory of truth.

- *G*: Changing the logic implies changing the meaning.

- *H*: Change of meaning is bad.

- *I*: The change is mere.

- *J*: This is no 'mere' relabelling.

- *K*: Change of truth schema is a change of the meaning of 'true'.

- *L*: The paradox forces a change of meaning.

The framework is represented in Figure 5.7 and its flattening in Figure 5.8.

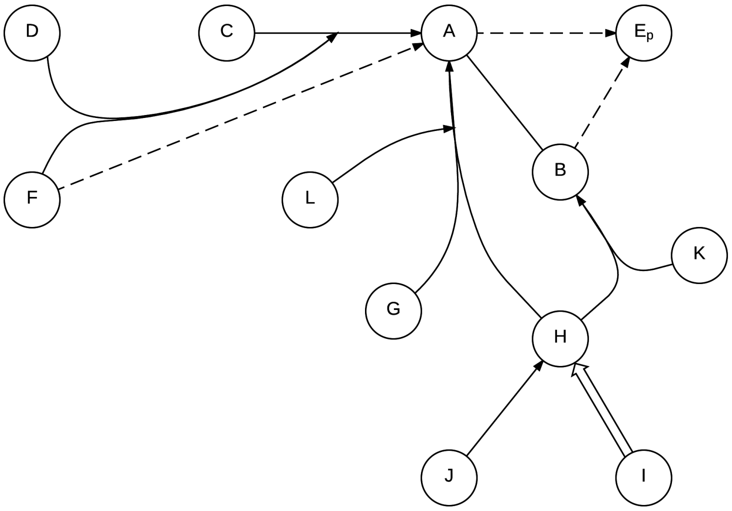

Figure 5.7: EEAF representing the reasoning behind Excerpt 6

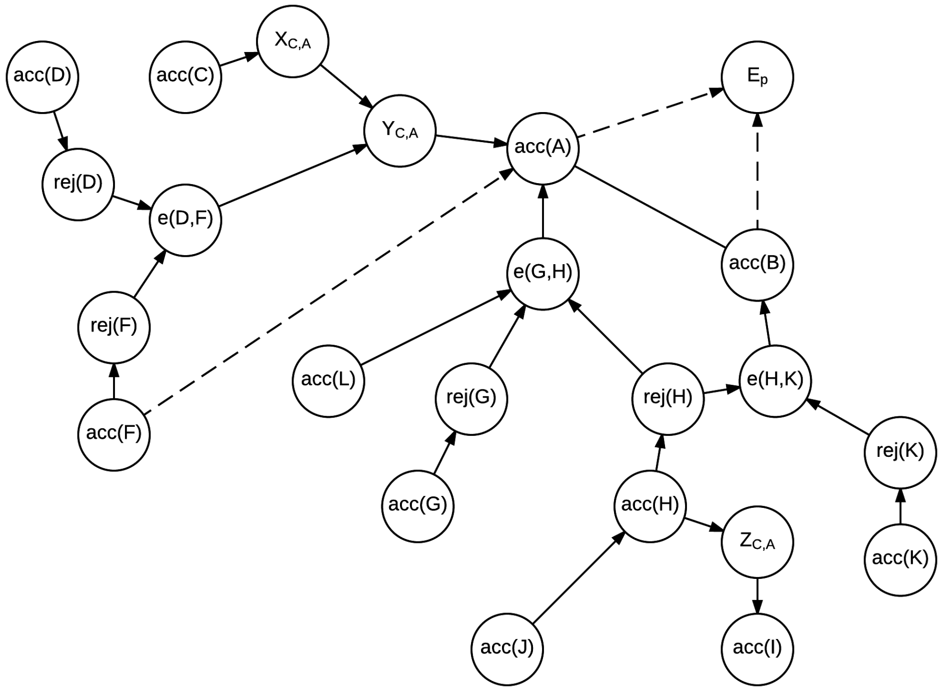

Figure 5.8: Flattened EEAF representing the reasoning behind Excerpt 6

Once again, we have omitted less-relevant auxiliary arguments for the sake of visibility.

By applying the selection procedures, we get that the argumentative cores are $\{A, C, D, F, L, G, J, K\}$ and $\{B, C, D, F, L, G, J, K\}$. We can distinguish here the two rivaling solutions which are both select. This is due to the fact that even though the author might have a preference for one or another, in the excerpt we have analyzed, he is merely defending the solutions represented in A from attacks and making no argument which attacks only the solutions represented in B.

The explanatory cores are $\{A, D, F, L\}$, $\{A, D, F, J\}$ and $\{B, J\}$. Notice that there are two different explanatory cores which contain A, as there are two arguments which individually defend A from the coalition attack of $\{G, H\}$. Each of these two sets include only of these two defending arguments for the sake of minimality.

Chapter 6

Future work and conclusions

6.1 Future work

Before concluding the thesis, we present some open problems related to the research of this thesis which it is worthwhile to research further.

One would be to investigate the equivalence between transpositions of all intuitively strict rules in $\mathcal{R}$ and allowing proofs by contradiction. We have noticed that under some circumstances, having proofs by contradiction is equivalent to having transpositions of all intuitively strict rules. However, it is not clear how to generalize this to all cases and thus this potential equivalence requires further research.

ASPIC-END is based on propositional logic, however philosophical literature about the liar paradox involves many use of quantifiers. Hence, it might be interesting to define a first-order version of ASPIC-END, in order to enrich the models and allow one to model this aspect of philosophical reasoning about the liar paradox.

Another point of interest would be to study the relationship between the ASPIC-END approach and the EEAF approach. The goal would be to get a better understanding of what one can do with each of these approaches. A formal correspondence could also be of interest, as well as informal comparisons of when which approach is best suited. It would also be interesting to combine both approaches and provide a structured argumentation system for building EEAFs using natural deduction.

In order to better study the viability of these approaches, one could also attempt to model more proposed solutions to the liar paradox, in particular paraconsistent ones, which have been left out of this thesis due to time limitations.

Note that as mentioned before, the models built in the course of this thesis will serve as material for the empirical studies of the theoretical work of the interdisciplinary project Cognitive Aspects of Formal Argumentation, which will start in November 2016 and will be led by Prof. Leon Van Der Torre and Prof. Christine Schiltz. In the course of this project, the cognitive plausibility of the models produced in this thesis, as well as of formal argumentation in general, will be studied.

While ASPIC-END was developed with the goal of modeling arguments about the solutions to the liar paradox in mind, it might be interesting to attempt to apply it to model arguments in a different context than logical paradoxes and allow us to to get a better idea of which are the features which fit well for explanatory debates in general and which are the features which are more unique to the modeling of logical paradox solutions.

6.2 Conclusions

In the course of this thesis, we have revisited some of the existing theories in formal argumentation, from the well-established abstract argumentation frameworks to the different extensions of it and also the structured argumentation approach of ASPIC+. We have then proposed a new formalism of structured argumentation, ASPIC-END, which includes explanatory features and natural deduction style proofs by contradiction, in order to provide us with more expressive power for the modeling of the debates around the solutions to the liar paradox. After this, we have used that formalism to model three excerpts describing three different solutions individually, and then have merged the models together in order to compare the three solutions. Additionally, we have presented a different approach called Extended Explanatory Argumentation Frameworks, where we tried to formalize the reasoning behind more complex excerpts in a manual way. We have finally reviewed a few open problems and possible future research related to the formalisms and models we have created in this thesis.

While the structured argumentation approach seemed more automated and well suited to model arguments in simple texts, attempts to use this formalism to model complex reasoning which uses several implicit assumptions and philosophical thinking have proven to be very difficult. In the manual formalization approach, while we were able to model more complex arguments which seemed to be closer to the text, multiple revisions had to be made in order to capture as best as possible the reasoning in the text, and even still then, different models might arise which also capture the same reasoning in a justifiable manner.

Bibliography

[1] Pietro Baroni, Federico Cerutti, Massimiliano Giacomin, and Giovanni Guida. Encompassing attacks to attacks in abstract argumentation frameworks. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, pages 83–94. Springer, 2009.

[2] Guido Boella, Dov M Gabbay, Leendert van der Torre, and Serena Villata. Meta-argumentation modelling i: Methodology and techniques. *Studia Logica*, 93(2-3):297–355, 2009.

[3] Guido Boella, Dov M Gabbay, Leendert WN van der Torre, and Serena Villata. Support in abstract argumentation. *COMMA*, 216:111–122, 2010.

[4] Claudette Cayrol and Marie-Christine Lagasquie-Schiex. Coalitions of arguments: A tool for handling bipolar argumentation frameworks. *International Journal of Intelligent Systems*, 25(1):83–109, 2010.

[5] 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.

[6] Hartry Field. *Saving truth from paradox.* Oxford University Press, 2008.

[7] Thomas Fowler. *The Elements of Deductive Logic.* At the Clarendon Press, 1887.

[8] Dov M Gabbay. Fibring argumentation frames. *Studia Logica*, 93(2-3):231–295, 2009.

[9] Sanjay Modgil. An abstract theory of argumentation that accommodates defeasible reasoning about preferences. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, pages 648–659. Springer, 2007.

[10] Sanjay Modgil and Henry Prakken. The aspic+ framework for structured argumentation: a tutorial. *Argument & Computation*, 5(1):31–62, 2014.

[11] Arthur N Prior et al. On a family of paradoxes. *Notre Dame Journal of Formal Logic*, 2(1):16–32, 1961.

[12] Raymond Reiter. A logic for default reasoning. *Artificial intelligence*, 13(1):81–132, 1980.

[13] Dunja Šešelja and Christian Straßer. Abstract argumentation and explanation applied to scientific debates. *Synthese*, 190(12):2195–2217, 2013.

Appendix A

Excerpts modeled in Section 5.3

A.1 Excerpt 4, taken from Saving Truth From Paradox [6], pages 4 and 5

We've seen that one possible reaction to Russell's paradox for (conceptual) properties is

Non-existence Solution: There is no such thing as "the Russell property", i.e. "the (conceptual) property of not instantiating itself". (Put loosely: There is no (conceptual) property corresponding to the predicate 'doesn't instantiate itself'. I take this as just an informal way of putting what's in the first sentence. Note that talk of "correspondence" doesn't appear in the official formulation.)

The idea is to decompose the schema (INST) into two components: a correct component

(INST$_w$) if there is a property of begin F, then it is instantiated by all and only the things that are F;

and a component that needs restriction

(COMP) There is a property of being F.

According to the Non-existence Solution, (INST$_w$) holds for all intelligible predicates F, but (COMP) fails for some of the, e.g. 'doesn't instantiate itself'. There will need to be may other failures of (COMP) too, if other paradoxes like Russell's are to be handled along similar lines. A serious theory that incorporates the Non-existence Solution will have to tell us for which F's (COMP) holds and for which ones it fails.

I've suggested, following Gödel, that the Non-existence solution isn't very attractive (for anyone who doesn't just reject conceptual properties out of hand). Conceptual properties aren't like sets, and it would violate their *raison d'etre* to suppose that for an intelligible predicate like 'doesn't instantiate itself', there is no corresponding property of *not instantiating itself.*

This is doubtless too quick, for one possibility would be to grant that every *intelligible* predicate has a corresponding property, but to deny that 'doesn't instantiate itself' is intelligible. On the most obvious version of this, we maintain the original (INST) (and hence, (COMP)),for any intelligible predicate F. But there's something very odd about holding that 'doesn't instantiate itself' isn't intelligible. It can be brought out by asking

“What part of ‘doesn’t instantiate itself’ don’t you understand?” It seems that if you accept (INST) (restricted to intelligible predicates), then you must regard the predicate ‘instantiates’ as intelligible, since it is used in the formulation of (INST)! Presumably ‘not’ and ‘itself’ are intelligible too, so how can ‘doesn’t instantiate itself’ fail to be intelligible? ‘Intelligible’ would have to be given a very special reading for this to make any sense. Of course we can always give it the special reading ‘expresses a property’, but (COMP) restricted to predicates that are “intelligible” in *this* sense becomes totally vacuous. We would in effect just be granting that (COMP) and the original (INST) fail for some predicates that are intelligible by ordinary standards. The pretense that we’ve *explained* the failures of (COMP) as due to the unintelligibility of the predicates is exposed as a fraud, and we will still need an account of which predicates it is for which (COMP) fails.

A.2 Excerpt 5, taken from Saving Truth From Paradox [6], pages 8 to 10

Step Four follows from Step Three, if we assume that $B \vee \neg B$ is a logical truth. That assumption is famously called the law of excluded middle (LEM). It is not only famous, it is famously controversial. There are some (e.g. mathematical intuitionists like the Dutch mathematicians Brouwer and Heyting) who deny its general applicability even within mathematics; indeed, some intuitionists (e.g. Michael Dummett) would deny its applicability to any statement that is not “in principle verifiable”. But on needn’t fall victim to “Dutch LEM disease” to suspect that there might be something suspicious about Step Four: questioning Step Four needn’t involve questioning it in connection with certain applications of “circular” predicates like ‘instantiates’.

Without Step Four, it isn’t immediately obvious why $B \leftrightarrow \neg B$ should be regarded as contradictory. This suggests the possibility of a fourth solution route:

Paracomplete Solutions: Excluded middle is not generally valid for sentences involving ‘instantiates’. In particular, the assumption that the Russell property R *either instantiates itself or doesn’t* should be rejected. Indeed the reasoning of the “paradox” shows that that assumption leads to contradiction.

For this to be an interesting option it must preclude the need for restrictions on (INST) and (COMP); that is, an interesting Paracomplete Theory must be one in which Naive Property Theory is consistent (where Naive Property Theory is the theory that for every predicate, there is a corresponding property that satisfies the “Instatiation Schema” (INST)). It is far from evident that an interesting Paracomplete Theory meeting this requirement is possible.

Indeed, a precondition of its possibility is that the logic vitiate *all* arguments from $B \leftrightarrow \neg B$ to contradiction, not just the Central Argument; and some logics without excluded middle, such as intuitionist logic, invalidate the Central Argument for the contradictoriness of $B \leftrightarrow \neg B$ while leaving other arguments intact. The most obvious route from $B \leftrightarrow \neg B$ to $B \wedge \neg B$ within intuitionism comes from the intuitionist *reductio* rule, which says that if $\neg B$ follows from Γ and B together, than it follows from Γ alone. Although intuitionists have reasons of their own for accepting this rule, the most obvious arguments for accepting it assumes excluded middle. For instance: $\neg B$ certainly follows from Γ and $\neg B$ together, so if it also follows Γ and B together then it must follow from Γ and $B \vee \neg B$ together, and hence *assuming excluded middle* it follows from Γ alone.

There is a class of logics without excluded middle ("deMorgan logics") that don't contain the *reductio* rule and that are in many respects for more natural than intuitionist logic: for instance, intuitionist logic restricts the inference from $\neg\neg A$ to A, and also restricts one of the deMorgan law (viz., the inference from $\neg(A \wedge B)$ to $\neg A \vee \neg B$), whereas deMorgan logics maintain all the deMorgan laws plus the equivalence of $\neg\neg A$ to A. And in deMorgan logics without excluded middle, $B \leftrightarrow \neg B$ is not contradictory.

A logic in which $B \leftrightarrow \neg B$ is not contradictory is *necessary* for a paracomplete solution, but far from *sufficient*: it remains to be shown that it is possible to consistently maintain the Naive Property Theory ((COMP) and (INST)) in such a logic. Indeed, we really want Naive Property Theory to include a bit more than (COMP) together with the Instatiation Schema (INST); we want it to include an intersubstitutivity claim, according to which

(i) o instantiates the property of being F

is fully equivalent to

(ii) Fo

in the sense that the claims (i) and (ii) can be intersubstituted even in embedded contexts (so long as these contexts are "transparent", that is, don't involve quotation marks, intentional operators, or the like). For instance, we want a principle that guarantees such a thing as

[It is not the case that o instantiates the property of being F] if and only if [it is not the case that Fo];

[If o instantiates the property of being F then B] if and only if [if Fo then B];

and so forth. Don't these follows from (INST)? They do in classical logic, but we're considering weakening classical logic. Still, I don't myself think a paracomplete solution would be very satisfactory if it required such a weak logic that these no longer followed.

So a good paracomplete solution would involve showing that (COMP) and (INST) can be maintained in a *reasonably strong* paracomplete logic that allows the derivation of the intersubstitutivity of (i) with (ii). Such a solution is in fact possible, as I will demonstrate in due course. In fact, it is paracomplete solution within a reasonably strong deMorgan logic that I will eventually be advocating in this book.

A.3 Excerpt 6, taken from Saving Truth From Paradox [6], pages 15 to 17

Weakening classical logic (whether by restricting excluded middle or in some other way) is not something to be done lightly. There are some obvious advantages to keeping to classical logic even for "circular" predicates: advantages of simplicity, familiarity, and so on. Choosing to forgo these advantages has its costs. But I will argue (primarily in Part II) that the *disadvantages* of keeping classical logic for "circular" predicates are also very great, so that the undoubted cost of weakening the logic is worth bearing.

Perhaps there are some who think that this cost-benefit analysis is inappropriate, that the very idea of tinkering with classical logic is irrational on its face since classical logic is obviously superior. The word ‘logicism’ would be a natural name for this attitude – in analogy to ‘sexism’, ‘racism’, ‘species-ism’ and so forth. Unfortunately it’s already taken, so let’s call the view ‘Logical Dogmatism’.

One possible defense of such Dogmatism is that if logic is not held fixed then anything goes. As an anonymous referee put it to me, “We didn’t weaken the logic as a way of hiding defects in Ptolemaic astronomy or old quantum theory; why should we modify the logic to hide the blemishes in the naive theory of truth?” The answer to this, I think, is that there is no known way (and little prospect of finding a way) to save either Ptolemaic astronomy or the old quantum theory by a change of logic, and little benefit to so doing since we have far more satisfactory alternatives. The proposal that we save the naive theory of truth by a change of logic is not the cheap non-solution that the objection envisages: it is something that must be earned by working out the details of the logic and of the theory based on it. Once we’ve worked out such a theory, we must weigh it against competing theories that keep classical logic but restrict the naive principles of truth, using the usual (somewhat vague and subjective) criteria for theory choice. With Ptolemaic astronomy or the old quantum theory, there is no serious prospect for such a theory being worked out that survives such a competition. The reader may think there is little prospect in the case of theory of truth either, but I invite him to withhold judgement until he has read the book.

A second common defense of Logical Dogmatism is based on the idea that “change of logic requires change of meaning”. To this I say, first, that the paradoxes force a change in the basic laws *either* of logic in a narrow sense *or* of the logic of truth, instantiation, etc.; or if you like, it forces a change in opinion about those laws. If change of (opinion about) the basic laws of ‘$\neg$’ and ‘$\rightarrow$’ counts as change of meaning, why doesn’t change of (opinion about) the basic laws of truth and the basic laws of instantiation? And as we’ll see, adhering to the principles of classical logic requires a *huge* change in standard principles about truth and instantiation. The upshot of this is that there is at least as good a case that the classical truth theorist is “changing the meaning ‘true’” as that the defender of the Intersubstitutivity Principle who restricts excluded middle is “changing the meaning of ‘not’” (or of ‘or’).

But second, why make a fetish about whether these things involve a change of meaning? As Putnam 1968 taught us, there is a distinction to be made between change of meaning and *mere* change of meaning. The switch from Euclidean geometry to generalized (variable curvature) Riemannian geometry involved revision of enough basic principles about straight lines that it may be somewhat natural to say that ‘straight line’ took on a different meaning. But if so, an abandonment of the old meaning and the invention of a new one was required to get a decent physical theory that is observationally adequate: for no reasonably simple observationally adequate theory allows for the existence of “straight lines in the Euclidean sense”. We needn’t of course have carried over the old term ‘straight line’ from Euclidean geometry to Riemannian, but there is enough continuity of doctrine to make it natural to do so. This is certainly not a *mere* change of meaning, i.e. a relabelling of terms without alteration of basic theory. The situation with truth is similar: here the “old theory”, involving both classical logic and the naive theory of truth, is simply inconsistent. Indeed, it’s trivial: it implies everything, e.g. that the Earth is flat. If you don’t want to be committed to the view that the Earth is flat you need a theory that differs from the naive theory in basic principles, either principles about truth or principles about logical matters more narrowly conceived. If giving up those basic principles involves a “change of meaning”, so be it: for then the “old meanings” aren’t really coherent, and they *need* changing. This is certainly no *mere* change of meaning, i.e. no mere relabelling.

Any resolution of the paradoxes will involve giving up (or at least restricting) some very firmly held principles: either principles of a narrowly logical sort, or principles central to the ordinary use of truth and instantiation predicates, or both. The principles to be given up are ones to which the average person simply can't conceive of alternatives. That's why the paradoxes are *paradoxes.* In this situation, I think we should be skeptical that asking whether the attempted resolution of the paradoxes "changes the meaning" of the terms involved is a clear question (whether these be ordinary logical terms like 'not' and 'if...then' or terms like 'true' and 'instantiates'). And I'm even more skeptical that it's a useful question.