Meta-Logical Argumentation with ASPIC-END – Can a Discussion about Logic be Modeled using Logic?

Dominic Deckert
Diplom Course in Computer Sciences
Technical University Dresden
dominic.deckert@mailbox.tu-dresden.de

Abstract

Structured argumentation theory is a versatile approach to knowledge representation and reasoning, where discourse between multiple different proponents can be represented simultaneously using formalism akin to the predicate language of first-order logic. By constructing attacks between these arguments, internally consistent and well-defended positions can be determined.

Among others, the ASPIC-style formalisms have been introduced to realize structured argumentation, including the well-known ASPIC$^+$. The ASPIC-END formalism is one of the latest approaches to ASPIC-style structured argumentation, specially designed to represent semi-formal discourse in formal sciences such as mathematics or philosophy ([CD20]). While it has been shown that ASPIC-END can represent discussions about the semantic paradoxes, that representation did not concern itself with the wider representation of the philosophy of logic. This field of philosophy is interesting for structured argumentation theory for a multitude of reasons: it is (relative to other fields) structured and formalized by representing inferences and statements in logical formulae, allowing for a mostly uncontroversial representation of the underlying knowledge as rules. At the same time, philosophy of logic allows for several conflicting positions to be held simultaneously (as opposed by mathematics, where the axiomatic representation allows for only one "correct" position). For these reasons, we believe that the philosophy of logic would pose an insightful test case for ASPIC-style structured argumentation in ASPIC-END. This thesis aims to survey the challenges modeling of this field of research will face.

We have chosen to use an article on philosophy of logic as a guidepost for our modeling efforts. The article "Truth and the Unprovability of Consistency" by Hartry Field ([Fie06]) provides a broad overview over the different *theories of truth* as well as an argumentative framework for how these theories relate to a general argument for theory consistency.

In this thesis, we propose several techniques to better model philosophy of logic and its different theories of truth. We judge these techniques on their general usefulness and discuss issues that merit further research.

Contents

Chapter 1

Introduction

Structured argumentation theory is a formalism designed to represent formal discourse between multiple proponents, modeling both the internal logic of their arguments and their ability to reject one-another's arguments through the use of counterarguments. This paradigm has been used to represent medical knowledge, legal discourse and other areas. In this thesis, we would like to apply structured argumentation theory to the field of philosophy, a field we believe has several unique challenges that could better inform the capabilities and limits of structured argumentation theory.

In particular, the field of *Computational Metaphysics* uses computational methods to analyze the logical underpinning of philosophical reasoning, putting it in association with Knowledge Representation and Reasoning. One of the major proponents of Computational Metaphysics, Christoph Benzmüller, famously used reasoning tools based on higher-order logic to analyze established philosophical proofs for the existence of a God. He found that one ontological God-proof was inconsistent with the assumptions it is based on, while a related proof by Gödel is internally consistent ([BP14]). This settled philosophical debates that had, until that point, not been able to conclusively show consistency or inconsistency for either of these God-proofs.

In philosophy, many questions fundamental to the human experience are discussed. Questions of how to characterize "correct" reasoning and "truth" are the focus of the philosophy of logic, a field within philosophy. Understanding the correct way of reasoning is of central importance for the foundations of mathematics, sciences and – one could even argue – everyday life. Logicians in this field have proposed different ways to perform logical inference and, consequently, different ways to express truth as a predicate within logical *theories of truth.*

Notably, an essential issue within philosophy of logic is that reasoning should be able to refer to its own sentences and assert that some of them are true. This ability to refer to its own sentences is the root of several semantic paradoxes in the philosophy of logic, such as the *Liar sentence* or *Curry's paradox.*[1] Semantic paradoxes allow any logic that fully accepts them to infer anything, leaving the logic *trivial.* Logicians have proposed a wide array of theories of truth that attempt to deal with the paradoxes or the triviality

[1] We will explore these paradoxes in more detail in 2.2.2.

they would entail. Ascertaining which theories of truth are internally consistent, do not "explode" (i.e. entail everything) and capture the *intuitions* behind correct reasoning is a central task within philosophy of logic.

Unfortunately, with the movement towards formalized mathematics in the early 20th century, it was proven that classical logic[2] could not characterize truth and be consistent. Gödel's Incompleteness theorems and Tarski's Undefinability theorem[3] showed disappointing limitations of classical logic when it came to truth.

We believe that philosophy of logic provides a rich area of research for the application of ASPIC-style argumentation theories. Research in philosophy is often done in the form of argumentative discourse, which argumentation theories lends itself as a medium for. Logicians may underline their central points with formalized and (internally) logically sound arguments, for which a reproduction in an ASPIC-style argumentation theory would be interesting. Furthermore, philosophy of logic is complex enough that different view points on logic can be held by different logicians, allowing for comparison between the theories of truth and discourse about which theory is "the correct one." This distinguishes philosophy of logic both from mathematics and other axiomatized sciences such as theoretical computer sciences. The mathematical view on logic can be considered somewhat monolithic in that it only accepts one logic; whereas computer science accepts a wide array of different logics, each fit to be used in different circumstances, without attempting to elevate one logic over the others as the "correct" logic.

We believe that there is merit in applying the methods of computational reasoning, and in particular Knowledge Representation and Reasoning, to the philosophy of logic. The aforementioned blurring of object and meta-language requires new approaches to be realized, and philosophy of logic poses a plethora of other atypical challenges to overcome to represent discourse in this field, reasoning in non-classical theories among them.

1.1 Aim of this Thesis

This thesis is intended to be a first step in opening the field of philosophy of logic to computational reasoning, in particular ASPIC-style argumentation theory. We wish to survey the area of research and both present and attempt to solve some of the challenges that reasoning within the philosophy of logic poses. For this purpose, we will try to construct an argumentation theory in the ASPIC-END formalism that can express multiple theories of truth simultaneously (although separately). This requires that we formalize relevant concepts such as entailment and truth, explore the different relevant theories of truth and assess how these theories deal with both the semantic paradoxes and Gödel's incompleteness theorem.

The ASPIC-END formalism has some distinct advantages over other ASPIC-style

[2] The modern notion of "classical logic is based on the work of Frege and Russell in the later 19th and early 20th century. They presented an abstract and formalized setting for the reasoning patterns common in mathematics. In particular, propositional logic and first-order predicate logic are both considered classical.

[3] We will discuss these theorems in greater detail in 2.1.1.

formalisms in that it allows even strict rules (called “intuitively strict” in ASPIC-END) to be attacked. This is crucial in a field of research where even the most well-established rule of logic may be discarded by some of the theories of truth.

In order to make surveying philosophy of logic easier, we have decided to focus in this thesis on one paper, the article “Truth and the Unprovability of Consistency” by professor Hartry Field ([Fie06]). In this article, he provides an overview of important theories of truth and how they relate to Gödel’s incompleteness theorem (i.e. he shows why these theories cannot show their own consistency). This article is well-suited for our purposes for a multitude of reasons: it narrows our focus to a few select theories of truth, it relates these theories to the mathematical underpinning of philosophy of logic, i.e. Gödel’s incompleteness theorem and the proof-theoretical nature of many logics, and the structure of the article is itself based on argumentation – Field first provides a general argument that any logic might construct to show its own consistency and then he constructs counter-arguments for each of the theories.

As mentioned above, this thesis is intended as a first step for analyzing philosophy of logic with argumentation theory. We believe that providing a complete modeling of the different theories of truth may bring new insights into both the discourse of the philosophy of logic as well as the capabilities and limits of argumentation theory. Argumentation theory also offers the possibility of representing the different theories simultaneously: each theory can be represented as a separate extension of the argumentation framework, attacks between extensions indicating that they assert contradictory statements or disagree on the allowed inferences.[4]

Furthermore, such an modeling may be capable of checking the consistency of arguments with the assumptions of an underlying theory of truth, check philosophical arguments for correctness or be used for a wide array of other tasks. It should be noted, however, that this ideal modeling (and in fact, even the argumentation theory provided in this thesis) are infinite in size. While this is not a result that can be avoided, we believe that it is necessary to fully represent the complexities of philosophy of logic. Unfortunately, this infiniteness poses a significant obstacle to reasoning and analysis tasks on the ontology, and not one that we believe can be fully solved. For the purposes of this thesis, we will not focus on the feasibility of reasoning tasks; instead our focus will be on the solution of more qualitative modeling problems.

1.2 Structure of this Thesis

This thesis is divided into seven chapters (including this one and the conclusion). Chapter 2 is intended to be a basic introduction into philosophy of logic and provide the necessary knowledge for Field’s article. Chapter 3 will define both argumentation theory in general, ASPIC-END in particular and the formal provisions we have found useful for formalizing philosophy of logic. Chapters 4 and 5 will present the argumentation theory we propose for philosophy of logic. The fourth chapter will be concerned with the

[4] We will illustrate this idea of extensions as theories in more detail in 4.3.1.

general principles both of logic as well as philosophy, while the fifth chapter will attempt the replicate Field's arguments for the different theories of truth. In Chapter 6, we will assess and discuss our modeling process and the resulting argumentation theory as well as note where future work may be most useful.

Chapter 2

Logical Theories in Philosophy

2.1 Philosophy of Logic

As mentioned in 1.1, the aim of this paper is to use the framework of argumentation theory to model the philosophical discourse on logic and truth; and in particular, an article by professor Hartry Field: "Truth and the Unprovability of Consistency" ([Fie06]). In this chapter, we will introduce the body of philosophical knowledge underlying this thesis and try to highlight where expectations may differ for anyone with a background only in computer science and mathematics.

In computer science – and especially in the field of knowledge representation and reasoning – logics can be seen as tools. They provide the basis with which to model knowledge about a system of interest, describe what inferences can be validly made from this knowledge and can be used to fulfill a wide variety of purposes. Logical formalisms such as Answer Set Programming or Description Logics have been developed concurrently for different purposes and without being held in open contradiction to one another, leading to a *plurality* of simultaneously-accepted logics. Furthermore, it is common to adapt existing logics to better fit a certain use case.[1]

In philosophy – and to a certain extent also in mathematics – the picture is different. In philosophy, logic is seen less as a tool or language of modeling, and more as a way to describe reasoning – the process of arriving at rational inferences based on prior knowledge. Here, reasoning is not restricted to one area of application: instead, philosophers ask how to generally describe correct reasoning, independent of the psychology of the thinker or the the "purpose" of the reasoning. Consequently, a plurality of contradictory logics is usually not accepted at the same time.

The *philosophy of logic*[2] is the main field of philosophy concerned with questions of

[1] For example, in 3.3, we will show how we adapt – or rather define a specific instance for – the ASPIC-END logical formalism to better suit discussions within the philosophy of logic.

[2] It is a topic of debate whether this area can also be called "philosophical logic." While some philosophers use both terms synonymously, there are others who see philosophical logic as a subfield of philosophy of logic, or note that the first term might give the false impression that philosophical logic is wholly divorced from mathematical logic. For the sake of simplicity we will only use the term "philosophy of logic."

logic. In the introduction to her book on philosophy of logic, Susan Haack presents a few of these central questions:

> What does it mean to say that an argument is valid? that one statement follows from another? that a statement is logically true? Is validity to be explained as relative to some formal system? Or is there an extra-systematic idea that formal systems aim to represent? What has being valid got to do with being a good argument? [...] Which formal systems count as logics, and why? ([Haa78, p. 1])

Notably, Haack does not require that logics are only expressed through formal systems. In fact, beyond the formalized logics (such as mathematical first-order logic) many informal theories are also considered in the philosophy of logic. This is in part due to the fact that the focus on "correct reasoning" extends beyond formal languages and into reasoning done in natural language. Since rational thought and arguments can be expressed in natural language, then so should logical theories.[3]

Allowing for expression and reasoning in natural language is one of the ways in which philosophy of logic goes beyond what is possible in computer science. However, this also introduces some problematic consequences for the philosophy of logic. One of them is the rise of a number of semantic paradoxes. These paradoxes, based on a semantical notion such as "truth" or reference to other sentences, can be expressed in natural language very easily.[4] Take for example the *Liar* sentence, which states that "this sentence is false." This sentence cannot hold, since then by its own definition it would entail that it is false (and does not hold), while if it did not hold, then it would be false and should hold (and thus cause a contradiction either way). Since it is important to the study of philosophy of logic, we will discuss the Liar sentence in greater detail in 2.2.2.

In fact, truth is a central interest of the philosophy of logic. While formal logics in mathematics and computer science tends to stay clear of such conceptual questions, it is quite important to the philosophy of logic. In natural language, where it is easy to reason about the truth of other sentences, defining truth as a conceptual notion within the logic itself would be desirable. However, philosophers such as Tarski have already shown that it is impossible to define truth within any sufficiently complex logic, using the paradoxicality of the Liar's sentence as a counter-example.[5] In fact, this not the only disappointing meta-result when it comes to logics: it is even impossible for any (internally consistent) logic to prove that it is consistent.[6]

[3] The classical logics of mathematics and computer science are commonly referred to as formal logics. Other logics, such as those that use natural language or non-classical inferences are usually referred to as informal.

[4] It can be argued that the problems of logic in natural language are due to the blurred lines between the object language (i.e. the medium of reasoning) and the meta-language (i.e. the medium of discussion about the reasoning). Concepts such as truth, consistency or the reference to sentences are ostensibly only possible within the meta-language. However, natural language acts as both the object and the meta-language for such logics, allowing these problematic constructions.

[5] This result, also called Tarski's Undefinability Theorem, will be discussed in greater detail in 2.1.1.

[6] We will also discuss Gödel's Incompleteness Theorems in greater detail in 2.1.1.

While it is certainly interesting to discuss the ramifications of reasoning in natural language, it can be hard to pin down the specific meaning and inferences done in natural language (especially in comparison to the more formalized mathematical logics). As such, Field (as other philosophers before him) restricts his view in [Fie06] to logics that can be represented with a proof-theoretic framework. *Proof theory*, initially introduced as a formalized underpinning of mathematics by Hilbert, is the approach of representing reasoning as proofs from a set of assumed propositions (the *axioms*), making repeated steps using predescribed *rules of inference* and arriving at new conclusions (also called *theorems*). Notationally, we will indicate axioms in proof theory as $\vdash \psi$ and rules of inferences as $\varphi_1, ..., \varphi_n \vdash \psi$ for any sentences φ_i and ψ. (We will introduce the specific rules and axioms used in different theories of truth in 2.2.)

2.1.1 Truth in Proof-Theoretic Logics

As has been discussed above, the idea of reasoning in natural language comes both with advantages and with its own problems. The ability to reference its own sentences within the logic itself can be seen as both. To provide a formal underpinning for the act of referencing sentences within the logic itself, the concept of *Gödel codes* is widely used in both mathematics and philosophy of logic. Gödel originally proposed these codes as a way to represent elements of the object language of a logic (i.e. sentences) by natural numbers.[7]

Definition 1. *In the following, let* $\langle\cdot\rangle : \mathcal{L} \to \mathcal{N}$ *be the injective Gödel code function. We will denote the Gödel code of a sentence* φ *by* $\langle\varphi\rangle$*.*

Consequently, any logic that can represent arithmetics on the natural numbers can arguably refer to their own sentences through the associated Gödel codes.

While this construction is easy to achieve and formalize within any logic that can characterize arithmetic, it is however a problem for the logic itself: as mentioned above, Gödel has already shown two (disappointing) results for these logics:[8]

Definition 2 (Gödel's Incompleteness Theorems)**.** *Let T be a consistent theory that can axiomatize arithmetics. Then the following two results hold:*

- *There are sentences within T that can neither be proven nor disproven*
- *It cannot be shown within T that T is consistent*[9]

While both theorems have interesting consequences in the philosophy of logic, for this paper the second Incompleteness Theorem will be more relevant, as Field uses it in his article (more on this in 5.1).

[7] While Gödel proposed an encoding similar to a reversed prime number factorization, there is a simpler intuition for how this can be achieved: in a computer any string of letters is encoded as a binary string. This binary string can then be interpreted as a natural number, allowing for an mapping from any letter string to the natural number (it is easy to enforce that this mapping is injective by adding a leading 1 to the binary string).

[8] We will introduce and express Gödel's Incompleteness Theorems here only very roughly. A more complete explanation of the theorems can be found at [Raa22].

[9] A theory is consistent if it does not accept a sentence and its negation at the same time.

Furthermore, the construction of Gödel codes allows the use of "meta-properties" that refer to properties such as truth or entailment that in a usual logic would only be possible in the meta-language. One such property is the *truth* predicate (i.e. the primitive predicate that contains exactly the Gödel codes of sentences that are held as true in the theory). Commonly in philosophy of logic, the requirements that should hold for a truth predicate are described using the T-scheme, first formulated by Tarski:[10]

Definition 3 (Tarski's T-scheme)**.** *For a characterization of truth in a logic, the following should hold for all sentences φ:*
$\langle\varphi\rangle$ is true iff φ

In proof-theory, this property can be expressed with two axioms called T-IN and T-OUT. (Here, → denotes implication.)

$$\begin{array}{lll} \text{T-IN:} & \vdash & \varphi \quad \to True(\langle\varphi\rangle) \\ \text{T-OUT:} & \vdash & True(\langle\varphi\rangle) \quad \to \varphi \end{array}$$

Notably, these axioms can be used with the inference rule *modus ponens* (more on this rule in 2.2.1). Some proof-theoretical logics that distinguish between implications and inferences through rules[11] may also wish to express the intuition behind the axioms T-IN and T-OUT as rules of inference. This allows the logics to accept the "weaker" rules of inference while rejecting T-IN and T-OUT:

$$\begin{array}{lll} \text{T-Introduction:} & \varphi & \vdash True(\langle\varphi\rangle) \\ \text{T-Elimination:} & True(\langle\varphi\rangle) & \vdash \varphi \end{array}$$

While the construction of a truth predicate might seem innocuous at first glance, logicians have shown that several problems arise from the introduction of a truth predicate. On the one hand, according to Tarski's Undefinability Theorem ([Tar36]), no classical logic can actually fulfill the full T-scheme. Instead, there must be instances of T-IN or T-OUT that the logic does not accept.[12]

On the other hand, several prominent paradoxes central to philosophy of logic can be formulated with a truth predicate (we will discuss some of these paradoxes in 2.2.2).

However, the idea of a truth predicate is an enticing one: the notion of truth is intuitively understood, and the usage in meta-language very common. Accordingly, there are even attempts at formalizing such predicates and their semantical properties in philosophy of logic, for example the Kripke-Feferman theory of truth (also called KF). It aims to construct the truth predicate in a classical logic as the least fixed point of a consequence operator[13] for sentences containing the truth predicate. A deeper

[10] While Tarski first published the T-scheme in 1935, its wider dissemination in the philosophy of logic did not start until he published another article in 1944 ([Tar44]).

[11] For example, Field's paracomplete logic does not always accept the conditional $\varphi \to \psi$ when it accepts the rule of inference $\varphi \vdash \psi$.

[12] We will discuss Tarski's Undefinability Theorem in greater detail in 5.3.1.

[13] This construction method is also used to define truth in other theories, some of them non-classical.

exploration of the semantic construction (and axiomatization) of KF would however be beyond the focus of this work.[14] Instead, we will focus on how ASPIC-END may be able to realize the truth predicate (and other predicates relying on self-reference of the logic) in 3.3.1.

2.2 Types of Logics

To understand Field's article, the focus of this paper, we will here present an overview of different logics in philosophy of logic. This is by no means an extensive account. Instead this account is concentrated on the most important logics Field studies in his article.

To do this, we will first introduce the common proof-theoretic "building blocks" (i.e. axioms and rules of inference) of these theories. Notably, while some of these building blocks are shared between all theories, others are essential to distinguishing between the different theories.

Then, we will introduce two essential semantic paradoxes that the theories have to contend with. The Liar's paradox and Curry's paradox can lead to inconsistency using some of the building blocks introduced below, and therefore motivate what combinations of building blocks cannot be accepted simultaneously.

Lastly, we will roughly sketch both classical and non-classical theories of truth important for Field's article.

2.2.1 Proof-Theoretic Building Blocks

As noted above, the proof-theoretical logics can usually be characterized by what rules of inference and axioms they permit. Here, we will present an overview of the different inference rules relevant to the article by Field.

Some of these "building blocks" are generally accepted by every theory of truth (under consideration). On the one hand, all the logics generally agree on the handling of the logical connectives $\wedge$, $\vee$, $\forall$ and $\exists$. The intended meaning of conjunction and disjunction can be achieved with a set of simple and uncontroversial rules:

$$\begin{array}{lrcl}
\wedge\text{-Introduction:} & \varphi_1, \varphi_2 & \vdash & \varphi_1 \wedge \varphi_2 \\
\wedge\text{-Elimination:} & \varphi_1 \wedge \varphi_2 & \vdash & \varphi_1 \\
 & \varphi_1 \wedge \varphi_2 & \vdash & \varphi_2 \\
\vee\text{-Introduction:} & \varphi_1 & \vdash & \varphi_1 \vee \varphi_2 \\
 & \varphi_2 & \vdash & \varphi_1 \vee \varphi_2
\end{array}$$

We will discuss the rules for the introduction and elimination of quantifiers in 4.2.3. Suffice it to say that they can be easily and uncontroversially formulated and are relevant only for more formalized systems that use both the the quantifiers and variables as symbols.

[14] For a more thorough understanding of this area of philosophy of logic, see [Kri76] for the original truth theory, or [HH06] for an attempt at an explicit axiomatization.

Another uncontroversial rule is the modus ponens, also called implication elimination. This rule can be used to characterize the implication symbol →: If A implies B, then whenever A holds, B must also hold.

$$\text{modus ponens:}\quad \varphi_1, \varphi_1 \rightarrow \varphi_2 \quad \vdash \quad \varphi_2$$

Beyond these simple and uncontroversial rules, there are some rules that are accepted within some logics and rejected within others.

On the one hand, there are several purely syntactical rules which are only accepted by some theories: *Double Negation Elimination* (DNE), the *Principle of Explosion* (PE),[15] the *Law of the Excluded Middle* (LEM)[16] and the *deMorgan Laws.* They are all accepted within classical theories, but may be rejected in non-classical cases.

$$\begin{array}{lrcl}
\text{Double Negation Elimination:} & \neg\neg\varphi & \vdash & \varphi \\
\text{Principle of Explosion:} & \varphi \wedge \neg\varphi & \vdash & \bot \\
\text{Law of the Excluded Middle:} & & \vdash & \varphi \vee \neg\varphi \\
\text{deMorgan Laws:} & \neg[\varphi_1 \vee \varphi_2] & \vdash & \neg\varphi_1 \wedge \neg\varphi_2 \\
 & \neg\varphi_1 \wedge \neg\varphi_2 & \vdash & \neg[\varphi_1 \vee \varphi_2] \\
 & \neg[\varphi_1 \wedge \varphi_2] & \vdash & \neg\varphi_1 \vee \neg\varphi_2 \\
 & \neg\varphi_1 \vee \neg\varphi_2 & \vdash & \neg[\varphi_1 \wedge \varphi_2]
\end{array}$$

Beyond these rules of inference, some logics employ (or forbid) the *meta-rules* of the calculus of natural deduction. They consist of *proof by contradiction*,[17] *implication introduction*[18] and *reasoning by cases*[19] and are commonly used in philosophy of logic. ASPIC-END uses these meta-rules of inference, so another application illustrating these meta-inferences can be found in 3.2.1.

Aside from these regular building blocks, theories in the philosophy of logic also allow or disallow inferences regarding the truth predicate. As mentioned above, the T-scheme (represented by T-IN and T-OUT) expresses the properties an ideal truth predicate should have, although several theories of truth reject parts of the T-scheme.

Furthermore, Field introduces the concept of *Intersubstitutivity* as an alternative way of characterizing the T-scheme: a logic satisfies this property when the "alike"

[15] The principle of explosion states that from a contradiction, anything follows. This is usually represented by the symbol ⊥, which stands for triviality and entails everything. The non-classical theories where contradiction does not explode are called paraconsistent.

[16] The law of the excluded middle is an axiom that restricts the possible entailment status of sentences: when it is satisfied, then for every sentence, it or its negation must be a theorem. Logics that do not accept the Law of the Excluded Middle are called paracomplete.

[17] Proof by Contradiction states that if φ entails inconsistency, then $\neg\varphi$ must hold instead. As a note: the LEM can be proven using proof by contradiction and DNE.

[18] With implication introduction, the material implication $A \rightarrow B$ can be shown by assuming A and showing that B must follow.

[19] With reasoning by cases, if a conclusion can be shown for the individual cases A_1 and A_2 and these cases encompass all possibilities, then the conclusion must also hold generally.
(This rule is in fact general: if there are more than two cases, then by nesting binary disjunctions this can be written as a disjunction of two "cases.")

formulae φ and $True(\langle\varphi\rangle)$ can be seen as equivalent – even within larger contexts.[20] This property can easily be realized using syntactical substitution and expresses the properties of the primitive truth predicate: With it, the truth predicate can be "introduced" or "eliminated" in all contexts (through the equivalence of the involved sentences).

Notably, the rule of Intersubstitutivity can be used to induce the full T-scheme: any logic that accepts it and the trivial implication $\varphi \to \varphi$ can derive T-IN and T-OUT by substituting part of the implication (see [Fie06][p. 582] for a more thorough explanation). Similarly, some combinations of building blocks are capable of reproducing others: for example, reasoning by cases can use the Law of the Excluded Middle and the Principle of Explosion to show that Double Negation Elimination must hold.

2.2.2 Paradoxes in Logic

One of the focuses in the study of philosophy of logic is the way in which paradoxical statements can be constructed and whether logical theories are able to handle or refute them. In this section, we will present two relevant paradoxes that find use in Field's article as well.

The Liar's Paradox

A significant paradox in philosophy of logic is the so-called Liar's paradox.[21] It revolves around the truth predicate and a self-referential sentence $Liar$[22] which (informally) states that "this sentence is false" (or $Liar$ being equivalent to $\neg True(\langle Liar\rangle)$).
It can easily be seen that this sentence leads to inconsistency in classical logic when T-IN and T-OUT are accepted. Reasoning by cases can easily show that a classical logic with the T-scheme leads to inconsistency for the Liar sentence:

[20] Field expresses the Principle of Intersubstitutivity for truth as follows: "If A and B are alike except that [...] one has 'p' where the other has '$\langle p\rangle$ is true' then one can legitimately infer B from A and A from B." ([Fie06, p. 583])

[21] For a more in-depth discussion, see [BGR20]

[22] In his article, Field used the symbol Q for his Liar sentence. We have not been able to find a motivation in the literature and have opted for a more expressive symbol instead.

A_1	By LEM, there are two cases:		$\Rightarrow$	$True(\langle Liar \rangle) \vee \neg True(\langle Liar \rangle)$
	Reasoning by cases			
A_2	*Case 1*			$True(\langle Liar \rangle)$
A_3	(T-OUT)		$\Rightarrow$	$True(\langle Liar \rangle) \rightarrow Liar$
A_4	(modus ponens)	A_2, A_3	$\Rightarrow$	$Liar$
A_5	By definition of *Liar*:	A_4	$\Rightarrow$	$\neg True(\langle Liar \rangle)$
A_6	This case is *inconsistent.*	A_2, A_5	$\Rightarrow$	$\bot$
A_7	*Case 2*			$\neg True(\langle Liar \rangle)$
A_8	By definition of *Liar*:	A_7	$\Rightarrow$	$Liar$
A_9	(T-IN)		$\Rightarrow$	$Liar \rightarrow True(\langle Liar \rangle)$
A_{10}	(modus ponens)	A_8, A_9	$\Rightarrow$	$True(\langle Liar \rangle)$
A_{11}	This case is *inconsistent.*	A_7, A_{10}	$\Rightarrow$	$\bot$
A_{12}	By reasoning by cases, this logic is generally inconsistent .			$\bot$

Table 2.1: Outline of the Liar's paradox

This proof relies on the law of the excluded middle (LEM), reasoning by cases as well as the T-scheme to establish general inconsistency. Since a classical logic accepts both LEM and the reasoning by cases, it must reject some part of the T-scheme. Non-classical logics may alternatively reject LEM or reasoning by cases.

The Curry Paradox

Another significant paradox related to the primitive truth predicate and self-reference is Curry's paradox (see [SB21] for a more involved discussion).

The Curry paradox centers around a statement S of the form "*If this sentence S is true, then C follows*" (this could "formally" be written as $S = True(\langle S \rangle) \rightarrow C$) where C can be any statement, however absurd.[23] With the inference rules of classical logic, the meta-inference of implication introduction and the T-scheme, one can show that this statement must be true (and that thus an arbitrary C must follow). The proof is done in two steps: first show that the implication $True(\langle S \rangle) \rightarrow C$ must hold, using *implication introduction*, and then conclude C by using *modus ponens* again (since the result of the implication introduction both is equivalent to and has as antecedent $True(\langle S \rangle)$).

[23] While traditionally, the Curry sentence concludes that "the moon is made of cheese," one can just as easily construct a Curry sentence for any conclusion. Accordingly, the construction of arbitrary Curry sentences allows a logic to infer any arbitrary sentence, leading to inconsistency (or more precisely, triviality).

	Implication Introduction			
A_1	Assume that S is true:			$True(\langle S\rangle)$
A_2	(T-OUT):		$\Rightarrow$	$True(\langle S\rangle) \rightarrow S$
A_3	By modus ponens:	A_1, A_2	$\Rightarrow$	S
A_4	By definition of S:	A_3	$\Rightarrow$	$True(\langle S\rangle) \rightarrow C$
A_5	By modus ponens:	A_1, A_4	$\Rightarrow$	C
A_6	Implication Introduction is successful:			$True(\langle S\rangle) \rightarrow C$
A_7	By definition of S:	A_6	$\Rightarrow$	S
A_8	(T-IN):		$\Rightarrow$	$S \rightarrow True(\langle S\rangle)$
A_9	By modus ponens:	A_7, A_8	$\Rightarrow$	$True(\langle S\rangle)$
A_{10}	Again by modus ponens:	A_9, A_6	$\Rightarrow$	C

Table 2.2: Outline of the standard Curry argument

Any logic that allows the construction of Curry sentences and this argument for arbitrary C therefore entails every sentence and is called *trivial*. Triviality is a more problematic form of inconsistency, since any trivial logic is useless in the sense that it cannot formally distinguish between valid conclusions and inane derivations. While (as noted in 2.2.3) philosophers have introduced paraconsistency as a way to prevent inconsistency from immediately trivializing a logic, this only exacerbates the problematic nature of Curry's paradox. Any logic where the above argument can be constructed for arbitrary Curry sentences is automatically trivial (and thus *more* than just inconsistent).

2.2.3 Logics

Here, we would like to give a brief overview of the most important groups of logics discussed by Field.[24] Unfortunately, the limited scope of this thesis did not allow modeling of all theories introduced by Field. Below, we will introduce the most important theories Field considered and give an explanation for those theories we decided to exclude.

Field approaches most of the logics from a proof-theoretical perspective, though he does not provide fully explicit definitions for any specific logic. In fact, most of the logics are introduced in relation to one another, i.e. what additional "building blocks" they admit or reject.

Classical Logics

The field of classical logics encompasses what, at least to a mathematician, would be *the* standard logics. All standard rules of inference and the meta-rules of natural deduction are valid in the classical logic.[25] However, as discussed above, the (full) T-scheme and

[24] Beyond the logics introduced here, Field briefly mentions intuitionist logics. While not without its proponents, intuitionist logic is not integral to Field's argument. We have therefore decided to exclude it from the logics introduced below.

[25] While the interest of the philosophy of logic is mostly on first-order logic as the classical logic, propositional logic would also fall under this umbrella.

Intersubstitutivity are incompatible with classical logics. In fact, any classical logic that allows a truth predicate has to provide a solution for the Liar's paradox in order to avoid inconsistency (and triviality).[26] Field characterizes classical logics in relation to first-order logic: they are the logics where "all arguments that are valid classically are taken as legitimate." ([Fie06], p. 571). Notably, since classically valid arguments do not involve the truth predicate, different classical theories of truth may disagree on what formulae they regard as true.

As mentioned above, the only recourse classical logic has when dealing with the Liar sentence is to restrict the T-scheme. Field notes two general ways classical theories of truth have done this: logics that restrict T-IN consequently do not include all entailed sentences in the truth predicate, whereas logics restricting T-OUT accept sentences as true that are not entailed. Both cases suppress the inconsistency argument for the Liar Sentence.

As mentioned above, a prominent representative of classical logic with a truth predicate in philosophy of logic is *KF*.[27] KF seeks to construct a truth predicate as a least fixed point construction: Using a consequence operator, the truth values of ever-more "remote"[28] sentences can be decided. While this logic can be more fully axiomatized (see [HH06]) and has been studied extensively, it does not evade the problems of the undefinability theorem and has to reject instances of the axiom T-IN.

Another group of interesting (though less widely discussed) classical theories are the classical dialetheic theories. These logics reject the axiom T-OUT and are otherwise classical in nature. Consequently, while they do not allow contradictions, they nevertheless allow cases where both a sentences and its negation are accepted as true (so-called dialetheia).[29] However, this does not inherently lead to inconsistency, since without T-OUT one cannot infer inconsistent sentences from these incoherent truth assertions.

Hyperdialetheism is a special case of classical dialetheism that Field briefly considers as a thought experiment: it is a theory of truth that regards all sentences as true. Field notes that this is ultimately a very uninteresting theory since it robs the truth predicate of any meaning. It is however an interesting thought experiment where representation in ASPIC-END might be insightful.

"Weakly" Classical Logics

In his article, Field considers an interesting, albeit unusual approach to classical logic that he calls "weakly classical" logics (see [Fie06], ch. 6).[30] He motivates this approach by the following observation: accepting either part of the T-scheme leaves one with a problematic situation in classical logic. For any Liar sentence $Liar$ (see 2.2.2), the logic would have to either accept $Liar$ and $\neg True(\langle Liar \rangle)$ (when accepting T-OUT)

[26] As mentioned in 2.2.2, philosophy of logic distinguishes between entailing an inconsistency and entailing everything. In classical logic, these two notions coincide.

[27] KF is based on a truth theory described by Kripke in [Kri76] and refined by Feferman.

[28] i.e. containing truth assertions that refer to other truth-containing sentences.

[29] For example, such a logic could accept both $True(\langle Liar \rangle)$ and $True(\langle \neg Liar \rangle)$ simultaneously.

[30] Field ascribes strong revision theories and strong supervaluational theories to this approach.

or $\neg Liar$ and $True(\langle Liar \rangle)$ (when accepting T-IN). However, to Field accepting either option would be absurd. Nevertheless, in order to preserve as much of the T-scheme as possible, it is necessary to accept *one or the other.* With this in mind, Field proposed to consider an alternative – rejecting reasoning by cases (and implication introduction[31]) in sentences where the truth predicate is involved. This makes it possible to accept the disjunction of two problematic options without having to arrive at either absurd situation (since without reasoning by cases, one cannot consider the individual cases that make up the disjunction). Field notes that discarding reasoning by cases is not very desirable since it makes disjunction (essentially) meaningless and we have to agree. For this reason we have decided to exclude weakly classical logics from the scope of theories of truth we will model in 5.3.2.

Paraconsistent Logics

One approach that has been suggested for the solution of the logical paradoxes in philosophy of logic are *paraconsistent* logics. For any Liar sentence *Liar* it can be shown that under classical logic both *Liar* and $\neg Liar$ must be entailed at the same time, which is the hallmark of an inconsistent logic. Under the Principle of Explosion, this even means that anything can be derived using the inconsistent Liar sentence, damning the logic to triviality.

Paraconsistent logics seek to mend this problem in a straightforward manner: they restrict the Principle of Explosion for "pathological" sentences,[32] and by doing so these logics can accept specific inconsistencies without slipping into triviality.

While the group of paraconsistent logic is mostly defined by their rejection of the Principle of Explosion, dialetheic logics go even a step further: they accept specific dialetheia, i.e. sentences where both it and its negation are accepted. Most dialetheic logics subscribe to a paraconsistent approach for the allowed rules of inference.[33]

Field focuses his exploration of paraconsistent theories on those that contain dialetheia. However, proponents of dialetheism have not agreed on one fully defined theory of truth. Instead, Field presents two main proponents with differing logics: Graham Priest and Jeffrey Beall. While Priest's version of paraconsistency – the older of the two – accepts the truth scheme, it does not accept the principle of Intersubstitutivity for formulae with negation (as well as modus ponens and the axiom scheme of induction), whereas Beall accepts Intersubstitutivity fully and instead seeks to reject the Law of the Excluded Middle. Field discusses both versions of paraconsistency and notes some inadequacies that they possess in his opinion.

[31] Field notes that a "classical logic still cannot maintain both T-IN and T-OUT as axioms without inconsistency." However, Field notes that one could accept the rules of inference – T-Elim and T-Introd – by rejecting the →-Introduction that would otherwise infer the axioms.

[32] Most paraconsistent logics restrict the Principle of Explosion for any sentences containing the truth predicate – such as the Liar sentence or any version of Curry's sentence.

[33] As mentioned above, Field discusses 2 types of "dialetheic" logics in philosophy of logic: 1) dialetheism as spearheaded by Priest and Beall (and introduced here), and 2) a variant of classical logic that does not universally accept T-OUT.

Unfortunately, we have encountered problems while trying to model a paraconsistent theories of truth in ASPIC-END (we will discuss these in 6.2.4). For this reason, we have decided to exclude paraconsistency from the theories of truth we modeled.

Paracomplete Logics

While other logical theories have solved the Liar's paradox by rejecting parts of the (rather intuitive) T-scheme or by accepting inconsistency without triviality, there are logicians (including Field) that lean towards a different solution. They propose rejecting the Law of the Excluded Middle (or instances containing pathological sentences). By rejecting the assertion that the Liar sentence must be true or false, one can avoid the inconsistency that any of those options would lead to:[34] With the rejection of LEM, it is possible to treat the Liar (or any other pathological) sentence as neither true nor not true. These "truth value gaps" are the paracomplete solution to the Liar paradox. Other logics that contain such truth-value gaps, such as Kleene's three-valued logic K_3, can be counted under the umbrella of paracomplete logics.

As an aside: *intuitionistic logic*, a non-classical approach to mathematics and logic, rejects LEM and can be seen as close to the paracomplete theories of truth. Notably though, intuitionism differs from them in a few key aspects: on the one hand, it is not inherently a theory of truth. It does not usually contain a truth predicate and thus need not make a statement about the T-scheme or Intersubstitutivity. On the other hand, intuitionism rejects many of the classical building blocks that paracomplete theories attempt to preserve (such as the deMorgan rules or double negation elimination). Since Field did not focus on intuitionism, we have also decided not to further assess the possible arguments for this theory.[35]

A variety paracomplete theories have been proposed[36] that disagree on how to handle the different kinds of implication. Consequently, these different paracomplete theories of truth accept different building blocks as well. In this thesis, we will concentrate on the paracomplete theory Field prefers (and outlines in his article).

Field's preferred type of paracomplete theory accepts many of the most intuitively appealing building blocks of philosophy of logic. It accepts the T-scheme fully (as well as Intersubstitutivity), most classical rules such as modus ponens, the principle of explosion, double negation elimination and the deMorgan laws as well as the meta-inference of reasoning by cases.[37] Disappointingly, Field has to restrict some other building blocks that he argues can be seen as relying on the intuition LEM represents: proof by con-

[34] As mentioned in 2.2.2, the Liar paradox hinges on LEM, the T-scheme and reasoning by cases. Rejecting one of them disables the standard argument for inconsistency.

[35] For those interested in a deeper exploration of intuitionism, see [Iem20].

[36] As a matter of fact: the categorization and name of "paracomplete logic" is a modern invention, first proposed by Field after a suggestion by Beall (see [Fie08, p. 12]).

[37] Intuitionist logic for example rejects double negation elimination as well as the deMorgan laws.

tradiction[38] and →-introduction.[39] However, Field rejects these meta-rules of inference only for pathological formulae.

2.3 Overview

Below is an overview of the rules of inference present in the different logics according to Field.

Two notes for this overview: (1) While weakly classical logics cannot accept the full T-scheme, they are an outlier in this overview since they do not specify which rule to restrict. Instead they accept the disjunction that one of the rules needs to be restricted, and reject reasoning by cases. To note this indeterminateness as to which rule is restricted, we will use a special $\sim$ symbol.
(2) As mentioned above, intuitionism is a theory of mathematics more than it is a theory of truth, and is therefore usually not formulated with the T-scheme or Intersubstitutivity in mind. For this reason, we have left these entries for intuitionism empty.

Logics	T-OUT	T-IN	T-Elim	T-Introd	Intersubstitutivity	deMorgan rules	LEM	DNE	PE	→-Introduction	Reasoning by Cases	Proof by Contradiction
KF	✓	X	✓	X	X	✓	✓	✓	✓	✓	✓	✓
Cl. Hyperdialetheism	X	✓	X	✓	X	✓	✓	✓	✓	✓	✓	✓
Cl. “Dialetheic”	X	✓	X	✓	X	✓	✓	✓	✓	✓	✓	✓
Weakly Cl. Logics	$\sim$	$\sim$	✓	✓	X	✓	✓	✓	✓	X	X	✓
Paracomplete Logics	✓	✓	✓	✓	✓	✓	X	✓	✓	X	✓	X
Intuitionism						X	X	X	✓	✓	✓	✓
Paraconsistency (Priest)	✓	✓	✓	✓	X	✓	✓	✓	X	✓	✓	X
Paraconsistency (Beall)	✓	✓	✓	✓	✓	✓	X	✓	X	✓	✓	X

Table 2.3: Rules of Inference in the different logics

[38] For proof by contradiction, also called the *reductio* rule $(\Gamma \wedge A \rightarrow \neg A) \vdash (\Gamma \rightarrow \neg A)$ in intuitionist logic, Field provides the following argument: In a context Γ, if A (and $\neg A$ trivially) lead to $\neg A$, then Γ alone must entail $\neg A$ since A and $\neg A$ are by LEM the only possible options, one of which must hold (see [Fie06]).

[39] We will discuss the problem with implication introduction in greater detail in 5.3.3.

Chapter 3

Argumentation Theory as a Modeling Approach

3.1 Argumentation Frameworks

In Field's paper, he presents a central argumentation as to why any logic should be able to prove its own consistency. He then presents different logics and how they contradict this central argumentation in different ways. we have chosen to represent and model this back-and-forth between different positions in ASPIC-END, an ASPIC-style structured argumentation theory.

3.1.1 History of Formal Argumentation

The notion of *argumentation frameworks* was first introduced by Dung as abstract argumentation frameworks ([Dun95]). Intuitively, an abstract argumentation framework can be used to express the *mechanics* of formal debate: debaters posit different arguments, some supporting and some in conflict with one another. Their positions (or rather, their sets of believed arguments) can then be judged on whether they are internally consistent and whether they can defend themselves against the "attacks" of the other debaters.

While versatile, Dung's abstract argumentation frameworks are just that: abstract. They do not express the contents of the represented arguments, only their attack relations to one another. Several approaches have been suggested in order to provide more concrete details on the arguments involved in a debate.[1] One of the most well-developed approaches is *structured argumentation theory*, based on proof theory and Pollock's work in formalizing argumentation and defeasible reasoning ([Pol87]). Structured argumentation theory aims to model arguments as trees of inferences based on axioms and assumptions.[2]

ASPIC-END, the modeling framework chosen in this report is a structured argumentation theory based on ASPIC$^+$, which itself was based on the ASPIC formalism. ASPIC

[1] See [BGGVdT18, section A] for a more detailed account.

[2] An overview of the different structured argumentation frameworks can be found in [BGH$^+$14].

was created as part of the ASPIC research project in 2006, although a well-rounded introduction can also be found in [Pra10]. ASPIC offered a structured argumentation approach based on any logical language, creating arguments based on inference rules of this language. It incorporated the idea of *defeasible conditionals*; inferences that usually, but not always, hold.

Soon after its official introduction, Caminada and Amgoud found some inadequacies of the ASPIC formalism: they showed that ASPIC does not keep to their Rationality Postulates, a list of rather uncontroversial conditions for a reasoning formalism ([CA07]). In correcting these inadequacies, they proposed changes to the ASPIC formalism that, together with other changes, would eventually lead to the conception of ASPIC$^+$. A thorough introduction to ASPIC$^+$ can be found in [MP14].

In order to refine the reasoning capabilities of argumentation frameworks, the notion of an argument *explaining* certain observations was proposed by Šešelja and Strasser in 2013 (see [ŠS13]). Their explanatory argumentation frameworks extend Dung's ideas by an explanation relation between arguments. With this relation, sets of arguments can additionally be judged on how many explananda (i.e. "goal" statements for which explanations are sought) they can explain. The recently introduced ASPIC-END incorporates explanatory argumentation into its design.

Among these ASPIC-style argumentation theories, ASPIC-END is the most suited framework for representing discourse in the philosophy of logic. We will elaborate on the advantages of ASPIC-END in 3.2, but it can be noted here that ASPIC-END was designed specifically with discourse in philosophy and other formalized sciences in mind.

3.1.2 Argumentation Frameworks

While Dung's abstract argumentation frameworks do not represent the inner structure of arguments, they do realize an attack relation between arguments, inducing semantics that judge *extensions* (i.e. sets of arguments) on how well-defended they are. Multiple formalisms have been proposed for constructing the inner structure of arguments and how an attack relation can arise naturally from this. One of the most well-known (and well-researched) is the ASPIC$^+$ framework, on which the ASPIC-END formalism is based (see [MP14] for an introduction). While we will introduce ASPIC-END in section 3.2, we will define the underlying Dung-style argumentation framework here.

Definition 4. *An* argumentation framework *is a tuple* $(\mathcal{A}, \longrightarrow)$ *where*

- $\mathcal{A}$ *is a set of arguments*
- $\longrightarrow \subseteq \mathcal{A} \times \mathcal{A}$ *is the* attack *relation*

In this formalism we will focus mostly on extensions that are particularly justified. These extensions are called admissible:

Definition 5. *An extension* $\mathcal{T} \subseteq \mathcal{A}$ *is called* admissible *if*

- *every argument* $o \in \mathcal{T}$ *is* defended, *i.e. for every* n *with* $n \longrightarrow o$ *there is a* $m \in \mathcal{T}$ *such that* $m \longrightarrow n$, *and*

- $\mathcal{T}$ *is* conflict-free*, i.e. there are no* $m, n \in \mathcal{T}$ *such that* $m \longrightarrow n$

The admissible extensions are the basis for any semantics in abstract argumentation frameworks. We only consider positions that cannot be refuted by other arguments and are internally consistent. Furthermore, Dung proposed *semantics* as a way of refining the selection of extensions that are accepted. Here we will only introduce the preferred semantics, but other semantics with their own advantages and disadvantages exist.

Definition 6. *An admissible extension is accepted under the* preferred *semantics iff it is set-maximal, i.e. any superset is inadmissible.*

The preferred semantics reduces the wide area of all admissible extensions only to those that "decide" on as many arguments as they can while preserving internal consistency. There is always a stable extension for an abstract argumentation framework, though it may in some cases only be the empty set.

3.2 ASPIC-END

While abstract argumentation frameworks can model the external relations between different arguments and positions in a discourse, they cannot represent the internal structure of any argument. This falls to ASPIC-END.[3] While similar to the earlier ASPIC$^+$-formalism, it has been designed with meta-logical reasoning in mind. In ASPIC-END, the very rules that are used to express logic, i.e. *intuitively strict*, can be rejected (see 3.2.1 for the formal specifics). Furthermore, ASPIC-END has additional desirable qualities: it is well-behaved with regards to the rationality postulates of Caminada and Amgoud, and it offers additionally reasoning methodsin the meta-inferences of the calculus of *natural deduction*. These inferences include proof by *contradiction*, proof by *case distinction* and *implication introduction*.[4]

As mentioned in 3.1.1, ASPIC-END realize *explanatory argumentation frameworks* (EAF). Unfortunately, we have found the usage of EAFs beyond the scope of this thesis and will not go into greater detail on their definition. As stated in 1.1, the focus of this thesis is on how to apply the methods of structured argumentation theory to philosophy of logic. While the additional infrastructure afforded by EAFs could offer more specific reasoning goals or allow further selection between the admissible extensions of the argumentation theory, both these applications have proven too specific for the broad nature of this thesis. In 6.2.1, we will however discuss possible ways in which explanatory argumentation could be useful in future works.

3.2.1 Formal Definitions

In ASPIC-END, the internal structure of arguments is represented by a sequence of rules that infer and support the argument. These rules are presented in a "logical"

[3] We heavily base the theoretical foundations of ASPIC-END on the foundational paper by Cramer and Dauphin ([CD20]). All definitions can be read up on there.

[4] The meta-rules of natural deduction have been introduced in 2.2.1.

language (i.e. using the formal definitions of predicate logic) as part of the *argumentation theory*, with attacks between arguments defined according to the rules that the arguments employ. We will define both argumentation frameworks and the attack relation below.

Argumentation Theories

Central to ASPIC-END are the *argumentation theories* that can be defined in it. An argumentation theory represents the knowledge underlying a discourse as inference rules. ASPIC-END, similar to ASPIC$^+$, distinguishes two types of inference rules: *intuitively strict* rules and *defeasible* rules. Intuitively strict rules try to capture the notion of "unbendable laws" in the subject matter.[5] *Defeasible* rules on the other hand express an inference that *usually* holds, but exceptions may exist.[6]

Definition 7. *An* argumentation theory *is a tuple* $(\mathcal{L}, \mathcal{R}, n, <)$ *where:*

- *$\mathcal{L}$ is a logical language[7] of free variables $\mathcal{L}_v$, function and relation symbols and closed under the logical connectives $\vee$, $\neg$ and $\exists$. Furthermore, $\mathcal{L}$ contains three special unary predicates $\mathsf{Assumable}_\neg, \mathsf{Assumable}_\vee, \mathsf{Assumable}_\rightarrow$ and the nullary predicate $\bot$ (which represents triviality).*

- *$\mathcal{R} = \mathcal{R}_{is} \cup \mathcal{R}_{def}$ is the set of rules, consisting of a set of intuitively strict rules $\mathcal{R}_{is}$ of the form $\phi_1, ..., \phi_n \rightsquigarrow \phi$ and a set of defeasible rules $\mathcal{R}_{def}$ of the form $\phi_1, ..., \phi_n \Rightarrow \phi$ (in both cases, $n \geq 0$ and all $\phi_i, \phi \in \mathcal{L}$).*
  To represent triviality, we require that the rule set $R_\bot = \{\bot \rightsquigarrow \phi \mid \phi \in \mathcal{L}\} \subset \mathcal{R}_{is}$

- *$n : \mathcal{R} \rightarrow \mathcal{L}$ is a (partial) naming function (n is undefined for all $r_\phi \in R_\bot$)*

- *$<$ is an ordering relation over $\mathcal{R}_{def}$ that expresses preference between defeasible rules*

The three special predicates $\mathsf{Assumable}_\neg, \mathsf{Assumable}_\vee, \mathsf{Assumable}_\rightarrow$ provide ways to allow (or rather, reject) arguments that rely on the meta-inferences of natural deduction.

Furthermore, to facilitate attacks on the validity of rules, we use the naming function n to represent rules in the arguments. Of note is that the rules in $R_\bot$ define the meaning of triviality and should not be attackable. We achieve this by not defining a name for these rules.

An argument can then be constructed inductively by applying one of the rules or meta-inferences on smaller sub-arguments. It should be noted, however, that all rules in $\mathcal{R}$ are treated as proof-theoretical rules (instead syntactically) instead of semantically. For example, it is not always the case that $\neg\neg\phi$ is equivalent to ϕ.

[5] The laws of mathematics and logics would be easy-to-grasp examples: if we believe that c is the sum of a and b, then it must **always** be the case that c is also the sum of b and a, by the commutative property of addition.

[6] A simple example is the usual rule that "all birds fly." While there are exceptions to this rule, it is still true in most situations.

[7] We will specify the logical language used in this project in 3.3.1.

In this report, we will use the customary notation for different properties of arguments: the conclusion of an argument A is denoted as $\mathsf{Conc}(A)$, the assumptions made for the special proofs are denoted as $\mathsf{As}_{\neg}(A), \mathsf{As}_{\vee}(A), \mathsf{As}_{\to}(A)$, or $\mathsf{As}(A)$ in totality. We call arguments with $\mathsf{As}(A) \neq \emptyset$ as *hypotheticals*. $\mathsf{Sub}(A)$ describes the sub-arguments of A, while $\mathsf{TopRule}(A)$ specifies the last applied rule (if the last step of the argument is indeed a rule application) and $\mathsf{DefRules}(A)$ collects all the defeasible rules in A.

With this, we can formally define arguments in a bottom-up manner in ASPIC-END:[8]

Definition 8. *An* argument A *on the basis of an argumentation theory* $\Sigma = (\mathcal{L}, \mathcal{R}, n, <)$ *has one of the following forms:*

1. $A_1, \ldots, A_n \rightsquigarrow \psi$, *where* $A_1, \ldots, A_n$ *are arguments such that there exists an intuitively strict rule* $\mathsf{Conc}(A_1), \ldots, \mathsf{Conc}(A_n) \rightsquigarrow \psi$ *in* $\mathcal{R}_{is}$.
   $\mathsf{Conc}(A) := \psi$, $\qquad \mathsf{As}_{\neg}(A) := \mathsf{As}_{\neg}(A_1) \cup \cdots \cup \mathsf{As}_{\neg}(A_n)$,
   $\mathsf{As}_{\vee}(A) := \mathsf{As}_{\vee}(A_1) \cup \cdots \cup \mathsf{As}_{\vee}(A_n)$, $\qquad \mathsf{As}_{\to}(A) := \mathsf{As}_{\to}(A_1) \cup \cdots \cup \mathsf{As}_{\to}(A_n)$,
   $\mathsf{Sub}(A) := \mathsf{Sub}(A_1) \cup \cdots \cup \mathsf{Sub}(A_n) \cup \{A\}$,
   $\mathsf{DefRules}(A) := \mathsf{DefRules}(A_1) \cup \cdots \cup \mathsf{DefRules}(A_n)$,
   $\mathsf{TopRule}(A) := \mathsf{Conc}(A_1), \ldots, \mathsf{Conc}(A_n) \rightsquigarrow \psi$.

2. $A_1, \ldots, A_n \Rightarrow \psi$, *where* $A_1, \ldots, A_n$ *are arguments s.t.* $\mathsf{As}(A_1) \cup \ldots \cup \mathsf{As}(A_n) = \emptyset$ *and there exists a defeasible rule* $\mathsf{Conc}(A_1), \ldots, \mathsf{Conc}(A_n) \Rightarrow \psi$ *in* $\mathcal{R}_d$.
   $\mathsf{Conc}(A) := \psi$, $\qquad \mathsf{As}_{\neg}(A) := \emptyset$,
   $\mathsf{As}_{\vee}(A) := \emptyset$, $\qquad \mathsf{As}_{\to}(A) := \emptyset$,
   $\mathsf{Sub}(A) := \mathsf{Sub}(A_1) \cup \cdots \cup \mathsf{Sub}(A_n) \cup \{A\}$,
   $\mathsf{DefRules}(A) := \mathsf{DefRules}(A_1) \cup \cdots \cup \mathsf{DefRules}(A_n) \cup$
   $\{\mathsf{Conc}(A_1), \ldots, \mathsf{Conc}(A_n) \Rightarrow \psi\}$,
   $\mathsf{TopRule}(A) := \mathsf{Conc}(A_1), \ldots, \mathsf{Conc}(A_n) \Rightarrow \psi$.

3. $\mathsf{Assume}_{\neg}(\varphi)$, *where* $\varphi \in \mathcal{L}$.
   $\mathsf{Conc}(A) := \varphi$, $\qquad \mathsf{As}_{\neg}(A) := \{\varphi\}$,
   $\mathsf{As}_{\vee}(A) := \emptyset$, $\qquad \mathsf{As}_{\to}(A) := \emptyset$,
   $\mathsf{Sub}(A) := \{\mathsf{Assume}_{\neg}(\varphi)\}$,
   $\mathsf{DefRules}(A) := \emptyset$, $\qquad \mathsf{TopRule}(A)$ *is undefined.*

4. $\mathsf{Assume}_{\vee}(\varphi)$, *where* $\varphi \in \mathcal{L}$.
   $\mathsf{Conc}(A) := \varphi$, $\qquad \mathsf{As}_{\neg}(A) := \emptyset$,
   $\mathsf{As}_{\vee}(A) := \{\varphi\}$, $\qquad \mathsf{As}_{\to}(A) := \emptyset$,
   $\mathsf{Sub}(A) := \{\mathsf{Assume}_{\vee}(\varphi)\}$,
   $\mathsf{DefRules}(A) := \emptyset$, $\qquad \mathsf{TopRule}(A)$ *is undefined.*

5. $\mathsf{Assume}_{\to}(\varphi)$, *where* $\varphi \in \mathcal{L}$.
   $\mathsf{Conc}(A) := \varphi$, $\qquad \mathsf{As}_{\neg}(A) := \emptyset$,
   $\mathsf{As}_{\vee}(A) := \emptyset$, $\qquad \mathsf{As}_{\to}(A) := \{\varphi\}$,

[8] The following definition has been taken from [CD20] and was provided courtesy of Dr. Marcos Cramer. The change from $\supset$ to $\to$ as the symbol for implication is due to this being the symbol Field uses for implication.

$\mathsf{Sub}(A) := \{\mathsf{Assume}_{\rightarrow}(\varphi)\}$,
$\mathsf{DefRules}(A) := \emptyset$, $\mathsf{TopRule}(A)$ *is undefined.*

6. $\mathsf{ProofByContrad}(\neg\varphi, A')$, *where* A' *is an argument such that* $\varphi \in \mathsf{As}_{\neg}(A')$ *and* $\mathsf{Conc}(A') = \bot$.
   $\mathsf{Conc}(A) := \neg\varphi$, $\mathsf{As}_{\neg}(A) := \mathsf{As}_{\neg}(A') \setminus \{\varphi\}$,
   $\mathsf{As}_{\vee}(A) := \mathsf{As}_{\vee}(A')$, $\mathsf{As}_{\rightarrow}(A) := \mathsf{As}_{\rightarrow}(A')$,
   $\mathsf{Sub}(A) := \mathsf{Sub}(A') \cup \{\mathsf{ProofByContrad}(\neg\varphi, A')\}$,
   $\mathsf{DefRules}(A) := \mathsf{DefRules}(A')$, $\mathsf{TopRule}(A)$ *is undefined.*

7. $\mathsf{ReasonByCases}(\psi, A_1, A_2, A_3)$, *where:*
   A_1 *is an argument such that* $\varphi \in \mathsf{As}_{\rightarrow}(A_1)$ *and* $\mathsf{Conc}(A_1) = \psi$,
   A_2 *is an argument such that* $\varphi' \in \mathsf{As}_{\rightarrow}(A_2)$ *and* $\mathsf{Conc}(A_2) = \psi$,
   A_3 *is an argument such that* $\mathsf{Conc}(A_3) = \varphi \vee \varphi'$.
   $\mathsf{Conc}(A) := \psi$,
   $\mathsf{As}_{\neg}(A) := \mathsf{As}_{\neg}(A_1) \cup \mathsf{As}_{\neg}(A_2) \cup \mathsf{As}_{\neg}(A_3)$,
   $\mathsf{As}_{\vee}(A) := (\mathsf{As}_{\vee}(A_1) \setminus \{\varphi\}) \cup (\mathsf{As}_{\vee}(A_2) \setminus \{\varphi'\}) \cup \mathsf{As}_{\vee}(A_3)$,
   $\mathsf{As}_{\rightarrow}(A) := \mathsf{As}_{\rightarrow}(A_1) \cup \mathsf{As}_{\rightarrow}(A_2) \cup \mathsf{As}_{\rightarrow}(A_3)$,
   $\mathsf{Sub}(A) := \mathsf{Sub}(A_1) \cup \mathsf{Sub}(A_2) \cup \mathsf{Sub}(A_3) \cup \{\mathsf{ReasonByCases}(\psi, A_1, A_2, A_3)\}$,
   $\mathsf{DefRules}(A) := \mathsf{DefRules}(A_1) \cup \mathsf{DefRules}(A_2) \cup \mathsf{DefRules}(A_3)$,
   $\mathsf{TopRule}(A)$ *is undefined.*

8. $\rightarrow$-$\mathsf{intro}(\varphi \rightarrow \psi, A')$, *where* A' *is an argument such that* $\varphi \in \mathsf{As}(A')$ *and* $\mathsf{Conc}(A') = \psi$.
   $\mathsf{Conc}(A) := \varphi \rightarrow \psi$, $\mathsf{As}_{\neg}(A) := \mathsf{As}_{\neg}(A')$,
   $\mathsf{As}_{\vee}(A) := \mathsf{As}_{\vee}(A')$, $\mathsf{As}_{\rightarrow}(A) := \mathsf{As}_{\rightarrow}(A') \setminus \{\varphi\}$,
   $\mathsf{Sub}(A) := \mathsf{Sub}(A') \cup \{\rightarrow\text{-}\mathsf{intro}(\varphi \rightarrow \psi, A')\}$,
   $\mathsf{DefRules}(A) := \mathsf{DefRules}(A')$, $\mathsf{TopRule}(A)$ *is undefined.*

9. $\forall$-$\mathsf{intro}(\forall x.\varphi(x), A')$, *where* A' *is an argument and* $x \in \mathcal{L}_v$ *is a variable such that there is no* $\psi \in \mathsf{As}(A')$ *such that* x *is free in* ψ, *and* $\mathsf{Conc}(A') = \varphi(x)$.
   $\mathsf{Conc}(A) := \forall x.\varphi(x)$, $\mathsf{As}_{\neg}(A) := \mathsf{As}_{\neg}(A')$,
   $\mathsf{As}_{\vee}(A) := \mathsf{As}_{\vee}(A')$, $\mathsf{As}_{\rightarrow}(A) := \mathsf{As}_{\rightarrow}(A')$,
   $\mathsf{Sub}(A) := \mathsf{Sub}(A') \cup \{\forall\text{-}\mathsf{intro}(\forall x.\varphi(x), A')\}$,
   $\mathsf{DefRules}(A) := \mathsf{DefRules}(A')$, $\mathsf{TopRule}(A)$ *is undefined.*

Of note is the construction of hypotheticals, an improvement over the ASPIC$^+$ formalism. All three of the meta-inferences of natural deduction rely on and work with assumptions: a proof by contradiction for example makes an assumption φ that leads to $\bot$ (i.e. triviality). Accordingly, the assumption cannot hold and $\neg\varphi$ must hold instead.[9]

In ASPIC-END, it is however possible to raise doubt about the applicability of these meta-inferences. In logics with a non-standard treatment of inconsistency or without the law of non-contradiction proof by contradiction is suspect: without LNC, showing

[9] ASPIC-END restricts hypotheticals to strict rule applications. As noted in the seminal ASPIC-END paper ([CD20]), proof by contradiction can be used to achieve rule contraposition which is not usually accepted for defeasible rules.

that φ cannot hold does not entail that the opposite must hold instead. In ASPIC-END, the non-applicability of a proof-method $m \in \{\neg, \vee, \rightarrow\}$ can be indicated by $\neg\mathsf{Assumable}_m$(this is one of the ways in which one argument might attack another).

ASPIC-END distinguishes different types of attacks between arguments, all of which take the internal structure of the arguments into account.

Definition 9. *For an argumentation theory $\Sigma = (\mathcal{L}, \mathcal{R}, n, <)$ we say that an argument A attacks an argument B (on some $B' \in \mathsf{Sub}(B)$) if:*

- *$\mathsf{Conc}(A) = \neg\varphi$ or $\neg\mathsf{Conc}(A) = \varphi$, B' is of the form $B'_1, ..., B'_n \Rightarrow \varphi$ and $As(A) = \emptyset$. This attack is called a* rebuttal*.*
- *$\mathsf{Conc}(A) = \neg n(r)$ or $\neg\mathsf{Conc}(A) = n(r)$, $\mathsf{TopRule}(B) = r$, there is no $\varphi \in \mathsf{As}(B')$ such that $\neg\varphi = \mathsf{Conc}(A')$ for $A' \in \mathsf{Sub}(A)$ and there are arguments $B_1, ..., B_n$ such that $B_1 = B'$, $B_n = B$, $B_i \in \mathsf{Sub}(B_{i+1})$ and $\mathsf{As}(A) \subseteq \mathsf{As}(B_1) \cup ... \cup \mathsf{As}(B_n)$. This attack is called an* undercut*.*
- *$As(A) = \emptyset$ and for a proof method $m \in \{\neg, \vee, \rightarrow\}$: $B' = \mathsf{Assume}_m(\varphi)$ and $\mathsf{Conc}(A) = \neg\mathsf{Assumable}_m(\varphi)$. This is called an* assumption-attack*.*

A rebuttal is an attack on some (possibly intermediary) conclusion of the other argument. It is one way in which ASPIC-END distinguishes between intuitively strict and defeasible rules: if an intuitively strict rule is accepted, then its conclusions cannot be refuted (since that is the intended meaning of intuitively strict rules). The results of a defeasible rule on the other hand can be rejected more readily. An undercut disputes the validity of a rule itself (even an intuitively strict rule), whereas an assumption attack rejects some assumption that was used for one of the meta-rules of natural deduction.

Hypothetical arguments should not be able to attack any arguments with less or different assumptions. In the case of rebuttals, since B' must end with a defeasible rule, which cannot have any assumptions, A must also have no assumptions. Similarly, assumption-attacks target the start of a hypothetical, which cannot have any assumptions.

We have to note, however, that a rebuttal from A on B' can easily cause an opposed rebuttal from B' to A. To decide this stalemate, ASPIC-END includes the preference relation $<$: an attack on B' in ASPIC-END requires that the least preferred (defeasible) rule of the arguments (according to $<$) occurs in B. This rule is also called the *weakest link* of B and the resulting semantics the *weakest-link* semantics.

Definition 10. *Let $\Sigma = (\mathcal{L}, \mathcal{R}, n, <)$ be an argumentation theory and A, B be two arguments from Σ. Then we define the lifting $\prec$ of $<$ to arguments to be the ordering where for any two arguments A, B from Σ, we have $A \prec B$ iff there is $r_b \in \mathsf{DefRules}(B)$ such that $r_b < r_a$ for all rules $r_a \in \mathsf{DefRules}(A)$. Then, we say that A* successfully rebuts *B if A rebuts B on some sub-argument B' and $A \not\prec B$.*

With the weakest-link condition for attacks, the induced *attack* relation between arguments can be defined.

Definition 11. *Let $\Sigma = (\mathcal{L}, \mathcal{R}, n, <)$ be an argumentation theory and A, B be two arguments from Σ. Then, A successfully attacks B ($A \longrightarrow B$) iff A undercuts, assumption-attacks or successfully rebuts B.*

With this, any argumentation theory $(\mathcal{L}, \mathcal{R}, n, <)$ induces an abstract argumentation framework and the arguments of the argumentation theory can be judged on how "reasonable" they are.

3.3 The Meta-Logical Instance of ASPIC-END

Among argumentation theories, the ASPIC-END formalism was (in part) designed to model meta-logical argumentation (i.e. reasoning about logic). For example, ASPIC-END allows arguments to undercut even strict rules. These rules are often used to reflect the seemingly "untouchable rules" of logic (such as double negation elimination). In meta-logical discussion, these rules may only hold in parts of the argument space – only some extensions (i.e. theories of truth in the philosophical sense) accept DNE or any other of the seemingly strict rules of logic.

However, so far there have not been attempts to fully define the primitive truth predicate in ASPIC-END. As part of this report, we would like to define a meta-logical instance of ASPIC-END[10] by realizing Gödel codes as *meta-terms* in the logical language. Furthermore, we will introduce (non-instantiated) *rule schemes* as a simplification, since the rules of logic (e.g. DNE) can best be represented by a collection of similarly-structured instantiated rules that treat all possible cases.[11]

The meta-logical instance introduced here has some features not commonly found in other ASPIC-style literature. The formalisms explored below present conceptually interesting approaches to ASPIC-END, but they unfortunately also result in an significant increase to the rule set. Under these formalisms, the resulting rule sets are countably infinite.[12] We will discuss this consequence in greater detail in 6.1.2.

3.3.1 Meta-Logical Terms

In his article, Field refers to several different meta-properties of other formulae. Apart from directly using the truth predicate, Field also singled out those formulae that indirectly referred to the truth predicate or displayed any other important syntactical properties. In particular, Field used the underlying Gödel codes in a variety of contexts. To this end, instead of allowing specific meta-predicates, we will expand an existing first-order logical language to allow for the inclusion of formulae as meta-terms (i.e. include Gödel codes in the terms that can be used by the theory). For this, we assume certain syntactical elements (in particular, predicate symbols) and will require that the underlying logic is a *first-order language.*

[10] In fact, what we propose here does not breach the confines of what is possible in ASPIC-END, since ASPIC-END broadly accepts any *logical* language.

[11] The Peano rule scheme for induction is a simple example for such a rule scheme.

[12] Barring uninteresting cases with only nullary function and predicate symbols.

Let $\mathcal{L}_v$ be a set of free variable symbols, $\mathcal{L}_f$ be a set of function symbols and $\mathcal{L}_P$ be a set of predicate symbols, each element of which has a fixed arity. For an n-ary function symbol f, we will use $f^{(n)}$ as a shorthand (and similarly $P^{(n)}$).

The logical languages $\mathcal{L}$ under scrutiny here consist of a set of terms $\mathcal{T}(\mathcal{L})$ and a set of formulae $\mathcal{F}(\mathcal{L})$.

Definition 12. *The set of all terms of a* first-order *logical language $\mathcal{L}$ is the smallest set $\mathcal{T}(\mathcal{L})$ such that:*

- *for all* 0*-ary functions symbols $c^{(0)}$ (called* constants*), $c \in \mathcal{T}(\mathcal{L})$*
- *for all other function symbols $f^{(n)}$ and $t_1, ..., t_n \in \mathcal{T}(\mathcal{L})$, we have $f(t_1, ..., t_n) \in \mathcal{T}(\mathcal{L})$*

Definition 13. *Given a set of terms $\mathcal{T}(\mathcal{L})$, the set of all "standard" formulae is the smallest set $\mathcal{F}(\mathcal{L})$ such that:*

- *for all* 0*-ary predicate symbols $P^{(0)}$, we have $P \in \mathcal{F}(\mathcal{L})$*
- *for all other predicate symbols $P^{(n)}$ and terms $t_1, ..., t_n \in \mathcal{T}(\mathcal{L})$, we have $P(t_1, ..., t_n) \in \mathcal{F}(\mathcal{L})$*
- *for all formulae $\varphi_1, \varphi_2 \in \mathcal{F}(\mathcal{L})$, we have that $\neg\varphi_1$, $[\varphi_1 \vee \varphi_2]$, $[\varphi_1 \wedge \varphi_2]$, $[\varphi_1 \rightarrow \varphi_2]$ is in $\mathcal{F}(\mathcal{L})$*
- *for all formulae $\varphi \in \mathcal{F}(\mathcal{L})$ and variables $x \in \mathcal{L}_v$, we have that $\forall x.\varphi$, $\exists x.\varphi$ is in $\mathcal{F}(\mathcal{L})$*

As seen in the definition above, any composite formula is delineated using square brackets. While necessary to avoid ambiguity, these brackets can hamper readability of the formula. It is easy to see that although we cannot completely discard them, some of the brackets may be superfluous. We will aim to display these brackets only when necessary to clarify the intended syntax.[13] Similarly without an impact on the syntax of the logical language we will change the size of corresponding pairs of brackets when multiple pairs of nested brackets appear.[14]

As mentioned above, Field's arguments contained meta-predicates that take other formulae (or other syntactical objects) as arguments. To allow this in ASPIC-END, we define a special self-containing language based on an underlying first-order language.

Definition 14. *Let $\mathcal{L}$ be a* first-order *logical language according to definitions 12 and 13. Then the associated* meta-logical language *$\mathcal{L}^{\langle\rangle}$ consists of a set of terms $\mathcal{T}(\mathcal{L}^{\langle\rangle})$ and a set of formulae $\mathcal{F}(\mathcal{L}^{\langle\rangle})$.*
$\mathcal{T}(\mathcal{L}^{\langle\rangle})$ is the smallest set such that:

- *$\mathcal{T}(\mathcal{L}^{\langle\rangle})$ is closed under the application of function symbols from $\mathcal{L}_f$*
- *for all function symbols $f \in \mathcal{L}_f$, the string $\langle f\rangle$ is in $\mathcal{T}(\mathcal{L}^{\langle\rangle})$*

[13] There is obviously a need to display some brackets for clarity. For example, the formula $\neg A \vee B$ could mean either $[\neg A \vee B]$ or $\neg[A \vee B]$. These formulae would be simplified as $\neg A \vee B$ or $\neg[A \vee B]$.

[14] For example, the formula $P(f(g(x)), f(f(g(x))))$ can be better displayed as $P\Big(f\big(g(x)\big), f\Big(f\big(g(x)\big)\Big)\Big)$.

- *for all predicate symbols $p \in \mathcal{L}_P$, the string $\langle p \rangle$ is in $\mathcal{T}(\mathcal{L}^{\langle\rangle})$*
- *for all terms $t \in \mathcal{T}(\mathcal{L}^{\langle\rangle})$, the string $\langle t \rangle$ is in $\mathcal{T}(\mathcal{L}^{\langle\rangle})$*
- *for all formulae $\varphi \in \mathcal{F}(\mathcal{L}^{\langle\rangle})$, the string $\langle \varphi \rangle$ is in $\mathcal{T}(\mathcal{L}^{\langle\rangle})$*

and $\mathcal{F}(\mathcal{L}^{\langle\rangle})$ is the set of all "standard" formulae and defined on the basis of $\mathcal{T}(\mathcal{L}^{\langle\rangle})$ in the same way $\mathcal{F}(\mathcal{L})$ is based on $\mathcal{T}(\mathcal{L})$ in definition 13.

Notably, this definition for meta-logical terms encapsulates a recursive definition: the set of formulae can be constructed from the given set of terms, and this in turn produces new formulae, extending the set of terms. It might be prudent here to show that the above definition is in fact well-formed. Since the proof for this is rather straightforward, we will only give a sketch of it.

One can easily imagine a fixed-point approach where $\mathcal{T}(\mathcal{L}^{\langle\rangle})$ can be obtained as the least fixed-point of an infinite chain of $\mathcal{T}_i$, with $\mathcal{T}_0 = \mathcal{T}_\mathcal{L}$ and $\mathcal{T}_{i+1}$ is the result of applying $\langle\rangle$ to all relevant elements of $\mathcal{T}_i$ (and the associated formulae). This step from $\mathcal{T}_i$ to $\mathcal{T}_{i+1}$ can be accomplished as a monotonously increasing function on the sets of terms. Under these conditions, a smallest fixed point must exist[15] and, it must be a logical language according to the construction method of the chain. (Notably, each $\mathcal{T}_i$ induces an associated set of formulae, meaning it suffices to keep an eye on the terms.)

3.3.2 Rule Schemes

In ASPIC-END, the rule set $\mathcal{R}$ is a set of *instantiated* rules, i.e. all occurring terms and variables are part of our logical language $\mathcal{L}$. In fact, most ASPIC-style argumentation theories present their rules explicitly, resulting in an overall finite set of rules. While definition 7 does not explicitly require that $\mathcal{R}$ is finite, this would still provide some desirable qualities: for example, a finite set of rules might feasibly be analyzed by an automated solver, yielding all possible extensions. Furthermore, it might be easier for any human modeler to read and understand a finite set of rules.

However, in the course of representing Field's argument, we have found in multiple instances that the modeling would benefit from a larger degree of *generality.* Field states the rules of inference (see 2.2.1) as abstract schemes that require instantiation to be used.

A similar example can be found in mathematics with the axiom scheme of *induction.* For any property that can be expressed through a formula φ,[16] induction states that if φ holds for all base cases (i.e. $\varphi(0)$ for natural numbers), and that φ is preserved when making minimal steps (i.e. $\varphi(n) \to \varphi(n+1)$ for natural numbers), then P must hold for all elements of the structure. This can be expressed in a single instantiated rule. However, general induction for all possible properties φ cannot be expressed as one rule in first-order logic, since the individual rule instances have disparate predicate symbols, which cannot be quantified or substituted in first-order logic.

In the example above, the symbol φ can be seen as a predicate acting as a placeholder for the property over which induction is proven. With rule schemes, we will introduce

[15] This conclusion can be based on the monotonic nature of the chain and the Knaster-Tarski fixed point theorem.

[16] For the sake of simplicity, we will provide an explanation using a unary φ

similar placeholder symbols for formulae, variables and terms. and define how they can be instantiated.

We would like to argue that there is merit in allowing rule schemes as a tool for simplification: with rule schemes, the respective rule set can be specified finitely and with increased readability.

To establish how this rule scheme produces rules when replacing terms or even formulae, we must first define substitutions to facilitate the replacing.

Definition 15. *Let $\mathcal{L}_1$ and $\mathcal{L}_2$ be logical languages. Then, for a formula $\varphi \in \mathcal{L}_1$ the* substitution *$\varphi[t_1/t_2]$ can be obtained by replacing all occurrences of the term $t_1 \in \mathcal{T}(\mathcal{L}_1)$ with the term $t_2 \in \mathcal{T}(\mathcal{L}_2)$. We will use $\varphi[t_1/t_2, s_1/s_2]$ as a shorthand for $(\varphi[t_1/t_2])[s_1/s_2]$. For any rule r, either of the form $\varphi_1, ..., \varphi_n \rightsquigarrow \psi$ or $\varphi_1, ..., \varphi_n \Rightarrow \psi$ (called* intuitively strict *and* defeasible*, respectively), the substitution $r[t_1/t_2]$ can be obtained by replacing all the subformulae φ_i and ψ by the result of applying the substitution $\varphi_i[t_1/t_2]$ and $\psi[t_1/t_2]$ respectively.*

Notably, a substitution here transforms formulae from one logical language into another. This is necessary below since rule schemes might contain placeholder symbols that are replaced during the instantiation process, and thus belong to a different logical language. We will use additional variable and predicates symbols to indicate placeholder symbols.

Definition 16. *Let $\mathcal{L}$ be a first-order logical language, $\widetilde{\mathcal{L}_v}$ and $\widetilde{\mathcal{L}_x}$ be disjoint sets of variables and $\widehat{\mathcal{P}}$ be a set of predicate symbols, where no symbols from $\widetilde{\mathcal{L}_v}$, $\widetilde{\mathcal{L}_x}$ and $\widehat{\mathcal{P}}$ occur in $\mathcal{L}$.*
Then, let the mixed schematic language *$\widetilde{\mathcal{L}}$ be the first-order logical language (in accordance with definitions 12 and 13) that results from including $\widetilde{\mathcal{L}_v} \cup \widetilde{\mathcal{L}_x}$ in the free variables of $\mathcal{L}$, where only variables from $\mathcal{L}_v \cup \widetilde{\mathcal{L}_x}$ can be used as quantified variables, and including $\widehat{P}$ in the predicate symbols of $\mathcal{L}$.*
Furthermore, let the schematic formula language $\widehat{\mathcal{L}}$ be the first-order logical language that results from only including $\widehat{P}$ in the predicate symbols of $\mathcal{L}$. It is easy to see that $\mathcal{L} \subseteq \widehat{\mathcal{L}} \subseteq \widetilde{\mathcal{L}}$.

To illustrate this definition (as well as the ones coming below), we will present two simple examples. The first will show the instantiation process for a single formula in lockstep with the definitions. We will present the instantiation process for a rule in the second example below.

Example 1. *The formula $\psi = \widehat{p}(g, f(\widetilde{t})) \wedge R(\widetilde{t})$ is a schematic formula. The placeholder symbols are the term placeholder $\widetilde{t} \in \widetilde{\mathcal{L}_v}$ and the predicate placeholder $\widehat{p} \in \widehat{P}$. Accordingly: $\psi \in \widetilde{\mathcal{L}}$.*

Below, we will introduce substitution methods for all variable, term and formula placeholders. In order to fully instantiate any mixed rule scheme, we will first apply the variable and *term* instantiation $\sigma_T : \widetilde{\mathcal{L}} \to \mathcal{P}(\widehat{\mathcal{L}})$ to all mixed rule schemes and then apply *formula* instantiation $\sigma_F : \widehat{\mathcal{L}} \to \mathcal{P}(\mathcal{L})$ to all the resulting rules.

Now, we can define the easier term substitution σ_F where quantified variable placeholders are replaced by any variable symbol and term placeholders can be replaced by any arbitrary terms.

Definition 17. *A* mixed rule scheme *is any (intuitively strict or defeasible) rule r from $\widetilde{\mathcal{L}}$ that contains variables $\widetilde{x}_1, ..., \widetilde{x}_m \in \widetilde{\mathcal{L}_x}$ and $\widetilde{w}_1, ..., \widetilde{w}_n \in \widetilde{\mathcal{L}_v}$ (for $m + n \geq 1$).*
The term instantiation of a rule can be obtained in two steps:

- *First, for every variable placeholder $\widetilde{x}_i$, generate separate substitutions $r[\widetilde{x}_i/v]$ for any variable symbol $v \in \mathcal{L}_v$ such that v does not occur in r. The set of all these substitution results for r shall be called $\sigma_X(r)$.*
- *Second, we will produce the set of all possible* term instances *for a given term: $\sigma_T(r) = \{r'[\widetilde{w}_1/t, ..., \widetilde{w}_n/t_n] \mid t_1, ..., t_n \in T(\widehat{\mathcal{L}}), r' \in \sigma_X(r)\}$.*

It is easy to apply term instantiation to this simple term:[17]

Example 2. *As mentioned above, $\psi = \widehat{p}(g, f(\widetilde{t})) \wedge R(\widetilde{t})$ is a schematic formula containing the term placeholder $\widetilde{t}$. This term can be instantiated with any term of $\mathcal{T}(\widehat{\mathcal{L}})$. As examples, we will use the terms g and $f(f(g))$:*

- $\psi_1 = \widehat{p}(g, f(\widetilde{t})) \wedge R(\widetilde{t})[\widetilde{t}/g] = \widehat{p}(g, f(g)) \wedge R(g)$
- $\psi_2 = \widehat{p}(g, f(\widetilde{t})) \wedge R(\widetilde{t})[\widetilde{t}/f(g)] = \widehat{p}(g, f(f(g))) \wedge R(f(g))$

It is important to note that the substitution transforms the mixed rule schemes into rules of the schematic formula language $\widehat{\mathcal{L}}$, from which it can be transformed into a rule of $\mathcal{L}$ in a second step. Notably, if no predicates from $\widehat{P}$ remain after the first step, then the resulting rules are already in $\mathcal{L}$.

While the direct substitution approach works well for terms, it is slightly different for formulae: When substituting a formula for a n-ary predicate symbol, the arguments that the predicate symbol applied to should be preserved. In the above induction example,[18] the axiom scheme should preserve n when substituting φ with an arbitrary formula (with one term "placeholder"). To facilitate this, we will first define the structure that such rule schemes with replaceable formulae (represented by simple predicates) can take: In order to actually achieve a formula substitution, we will require a support structure that represents the formula that can be inserted during substitution while preserving the terms that were arguments of the predicate. We will achieve this by constructing new "schematic" formulae with numbered variables $\widehat{x}_i$ (where $\widehat{x}_i$ is intended as a stand-in for the i-th argument).

Definition 18. *Let $\mathcal{L}_v^*$ be a countable set of variables that do not occur in $\mathcal{L}$ and are linearly ordered (we will here represent these variables by $\widehat{x}_1, \widehat{x}_2, \widehat{x}_3, ...$). Then let $\mathcal{L}^*$ be the first order logical language that results from including $\mathcal{L}_v^*$ in the free variables of $\mathcal{L}$. An n-ary* schematic formula *$\varphi \in \mathcal{L}^*$ is a formula where the largest variable from $\mathcal{L}_v^*$ that occurs in φ has an index smaller or equal to n.*

[17] We will present the term and formula instantiation in example 5.

[18] Recall that the induction step requires $\varphi(n) \to \varphi(n + 1)$ to hold.

Example 3. *In the running example, the schematic formula $\psi_1 = \widehat{p}(g, f(g)) \wedge R(g)$ contains only the binary predicate placeholder $\widehat{p}$. Accordingly, the schematic formula we aim to replace it with should have at most two variables from $\mathcal{L}_v^*$ (i.e. be 2-ary). Two examples for such variables are:*

- $Q(\widehat{x}_1, \widehat{x}_2)$
- $R(\widehat{x}_1) \vee R(\widehat{x}_2)$

To replace the predicate placeholders $\widehat{p}$ by any arbitrary formula (and be consistent across a substitution), we will require a mapping between the representative predicates and corresponding schematic formulae. With this, we can finally define how to obtain formula instances from the rule schemes.

Definition 19. *We shall call a mapping function $f : \widehat{P} \to \mathcal{L}^*$* arity-preserving *iff for all n-ary $\widehat{p} \in \widehat{P}$, $f(\widehat{p})$ is an n-ary schematic formula. Then, for any arity-preserving f, any n-ary predicate $\widehat{p} \in \widehat{P}$ and any formula $\widehat{p}(t_1, ..., t_n) \in \widehat{\mathcal{L}}$, the instantiation $\widehat{p}(t_1, ..., t_n)[f]$ is the substitution $f(\widehat{p})[\widehat{x}_1/t_1, ..., \widehat{x}_n/t_n]$.*

With this, we can fully instantiate the running example:

Example 4. *In the running example, the schematic formula $\psi_1 = \widehat{p}(g, f(g)) \wedge R(g)$ contains only the binary predicate placeholder $\widehat{p}$. Supposing this is the only variable in $\widehat{P}$, the function $f : \widehat{p} \mapsto R(\widehat{x}_1) \vee R(\widehat{x}_2)$ is an arity-preserving mapping function. Accordingly, the instantiation can be obtained as:*

$$\begin{aligned} \widehat{p}(g, f(g))[f] &= f(\widehat{p})[\widehat{x}_1/g, \widehat{x}_2/f(g)] \\ &= R(\widehat{x}_1) \vee R(\widehat{x}_2)[\widehat{x}_1/g, \widehat{x}_2/f(g)] \\ &= R(g) \vee R(f(g)) \end{aligned}$$

Definition 20. *For any rule r and arity-preserving f, a complete instantiation $r[f]$ of the rule r can be achieved componentwise : let $\varphi_1, ...\varphi_n$ and ψ be the subformulae of r,*[19] *then $r[f]$ can be obtained by applying for every $\widehat{p}(t_1, ..., t_n)$ occurring in any subformula the substitution $\big[\widehat{p}(t_1, ..., t_n)/f(p)[\widehat{x}_1/t_1, ..., \widehat{x}_n/t_n]\big]$ to the subformulae.*
Then, the formula instantiation of a rule is $\sigma_F(r) = \{r[f] \mid f \text{ is arity-preserving}\}$.

Applying σ_T and σ_F, every rule scheme can be transformed into a set of rules that satisfies definition 6 of [CD20].

Definition 21. *Given any mixed rule scheme r, the set of all instances of r can be obtained as $\bigcup_{r1 \in \sigma_T(r)} \sigma_F(r_1)$.*

To illustrate the way in which the term and formula instantiation are applied, we will present a second example for a single rule:

Example 5. *Let $r = \forall\widetilde{x}.\widehat{p}(\widetilde{x}) \rightsquigarrow \widehat{p}(\widetilde{\tau})$*[20] *be a rule of $\widetilde{\mathcal{L}}$. This rule contains the variable placeholder $\widetilde{x}$, the term placeholder $\widetilde{\tau}$ and the predicate placeholder $\widehat{p}$.*
The instantiation process for this rule is done in three steps:

[19] That is, r is of the form $\varphi_1, ...\varphi_n \rightsquigarrow \psi$ or $\varphi_1, ...\varphi_n \Rightarrow \psi$.

[20] This is the rule for ∀-introduction as it will be introduced in 4.2.3.

- Instantiating variable placeholders*: since r does not contain any other variables, $\widetilde{x}$ can be substituted with any variable. Thus, for any $v \in \mathcal{L}_v$, the rule $r_1 = \forall v.\widehat{p}(v) \rightsquigarrow \widehat{p}(\widetilde{\tau})$ is in $\sigma_X(r)$.*
- Instantiating term placeholders*: for any term $t \in \mathcal{T}(\mathcal{L})$ and variable $v \in \mathcal{L}_v$, the rule $\forall v.\widehat{p}(v) \rightsquigarrow \widehat{p}(t)$ is in $\sigma_T(r)$.*
  As an example, the rule $r_2 = \forall v.\widehat{p}(v) \rightsquigarrow \widehat{p}(\langle True(x)\rangle)$ is a result of term instantiation.
- Instantiating predicate placeholders*: the predicate placeholder $\widehat{p}$ is a unary predicate, and accordingly, any arity-preserving mapping function will map it to an 1-ary schematic formula. This formula is then inserted at the place of $\widehat{p}$.*
  For example, $f : \widehat{p} \mapsto True(\widehat{x}_1)$ is an an arity-preserving mapping function for r, and for the intermediary rule r_2, this instantiation results in
  $r_2[f] = \forall v.True(v) \rightsquigarrow True(\langle True(x)\rangle)$.

Unfortunately, when $\mathcal{L}$ has an infinite number of terms or formulae, then there are infinitely many rule instances.[21] While the ASPIC-END formalism can reason with such infinite rule sets, these rule sets might be rather inefficient to reason with.

[21] Any non-nullary function symbol results in an infinite logical language, making this condition very common. Likewise, any meta-logical language fulfills this condition.

Chapter 4

General Inference Rules for the Philosophy of Logic

In this chapter, we will present a representation for a selection of the inference rules used in philosophy of logic and, in particular, Field's argumentation.

First, we will present and motivate a collection of modeling principles that we have found useful for this project. With these principles in mind, we will then present new rules and rule schemes for defining several logical operators in our knowledge base: disjunction, equality and the universal and existential quantifiers. On this basis, we can then define structural induction, a central proof method of Field's arguments.

Furthermore, we will introduce ASPIC-END rules for the different rules of inference presented in 2.2.1. To facilitate comparison between the different logical theories that Field considers, we will present a set of rules that restricts the applicability of these rules to only those logical theories where they actually hold.

4.1 Methodology

4.1.1 Modeling Principles

Over the course of this project, we have considered different approaches to modeling the arguments Field provides in his paper. We have found a few guiding principles that we will follow in the rest of this thesis.

- *Minimality*: The rule corpus that we will produce as part of the modeling is quite expansive, and therefore unwieldy. This poses a problem to tractability: without automatic solving, any rule that we add to the argumentation theory adds an additional chance that unintended (or unrecognized) problems arise. It might however be tempting to solve some modeling problems through the addition of individual rules: Reasoning patterns such as contraposition or variable renaming, while possible with the meta-inferences of natural deduction, could be represented more elegantly using explicit rules. But for the sake of tractability, we have decided to keep the argumentation theory as small as possible, avoiding redundant rules wherever possible.

- *Authenticity*: Wherever possible, we will try to keep our formalization close to Field's arguments. In some instances, this might come at the detriment of minimality. In such an instance, we will explain how we have weighed between the two principles.
- *Explicitness*: We will try to express all formal arguments that are presented by Field as explicitly as possible. In some instances, we have found that Field's arguments, while rather formal, do not make explicit all the rules of inference and other logical steps they employ. Only in such instances will we "fill in the gaps" by presenting a formal argument that we believe mirrors Field's argument. However, wherever it is feasible we will keep to Field's arguments step-by-step.
- *Closeness to Natural Deduction*: The design of ASPIC-END has in part been based on the inference rules of Natural deduction. In particular, reasoning by cases, $\rightarrow$-Introduction and proof by contradiction all allow hypothetical reasoning, which is an improvement over other formal argumentation formalisms. However, even beyond these meta-inferences, Natural Deduction provides formal characterizations for the operators it uses (such as $\vee$). In keeping with the spirit of ASPIC-END, we will base our formal characterizations on Natural deduction wherever applicable.

Explicitness differs from authenticity insofar as explicitness is concerned with the grade of detail that is given to the formulated arguments, whereas authenticity is concerned with how close these arguments mirror Field's.

4.2 General Rules for the Modeling Language

Having established these guiding principles, we will now present the body of general ASPIC-END rules that we will use to reason within philosophy of logic.

4.2.1 Conjunction and Disjunction

Two general logical notions used by Field in his arguments are conjunction and disjunction. As discussed in 2.2.1, their intended meaning can be ensured with the following rules:[1]

$$\begin{array}{rcl} \widehat{p_1} \wedge \widehat{p_2} & \rightsquigarrow & \widehat{p_1} \\ \widehat{p_1} \wedge \widehat{p_2} & \rightsquigarrow & \widehat{p_2} \\ \widehat{p_1}, \widehat{p_2} & \rightsquigarrow & \widehat{p_1} \wedge \widehat{p_2} \\ \widehat{p_1} & \rightsquigarrow & \widehat{p_1} \vee \widehat{p_2} \\ \widehat{p_2} & \rightsquigarrow & \widehat{p_1} \vee \widehat{p_2} \end{array}$$

[1] While $\vee$-elimination, i.e. *reasoning by cases*, is a rule in natural deduction, in ASPIC-END this is realized with hypotheticals.

This selection of rules satisfies the principle of minimality from 4.1.1. While it would (for example) be easy to introduce a rule for commutativity of $\land$, this can also be concluded from the given rules and would therefore be superfluous: with the rules, one can single out $\widehat{p_2}$ and $\widehat{p_1}$, then combining these results to $\widehat{p_2} \land \widehat{p_1}$.[2]

4.2.2 The Equality Symbol "="

Furthermore, while not all logics traditionally require the existence of a unique equality predicate, several steps of Field's argumentation utilize it. The equality predicate "="[3] in ASPIC-END can be characterized by requiring that it must fulfill the AC0 properties: = must be *reflexive*, *symmetrical* and *transitive.* Furthermore, it must fulfill the *substitution* property (see below).

$$\begin{array}{rcl} & \rightsquigarrow & \widehat{p_1} = \widehat{p_1} \\ \widehat{p_1} = \widehat{p_2} & \rightsquigarrow & \widehat{p_2} = \widehat{p_1} \\ \widehat{p_1} = \widehat{p_2}, \widehat{p_2} = \widehat{p_3} & \rightsquigarrow & \widehat{p_1} = \widehat{p_3} \end{array}$$

Usually, the substitution property is expressed for a function application $f(t_1, ...t_n)$, where any t_i can be substituted by an equivalent term s. However, in this thesis the function application and the "context" terms t_j $(j \neq i)$ can be provided by rule schemes, leading to a simpler version of the substitution property:

$$\widehat{p}(\widetilde{t_1}), \widetilde{t_1} = \widetilde{t_2} \quad \rightsquigarrow \quad \widehat{p}(\widetilde{t_2})$$

Even beyond specific objects, it might be useful to state the equivalence of sentences (using the meta-logical terms introduced in 3.3.1). In fact, Field utilizes the equivalence of sentences in several of his argument to replace one sentence with an equivalent sentence within the truth context. In accordance with the principle of authenticity, we will try to replicate this approach.[4]

$$\begin{array}{rcl} \widehat{p} \to \widehat{q}, \widehat{q} \to \widehat{p}, \mathit{True}(\langle\widehat{p}\rangle) & \Rightarrow & True(\langle\widehat{q}\rangle) \\ \widehat{p} \to \widehat{q}, \widehat{q} \to \widehat{p}, \neg \mathit{True}(\langle\widehat{p}\rangle) & \Rightarrow & \neg True(\langle\widehat{q}\rangle) \end{array}$$

4.2.3 The Quantifiers $\forall$ and $\exists$

Many of the arguments used by Field operate on an abstract level, generalizing from specific formula. In formal language, this is easiest represented with the universal and existential quantifiers. However, in ASPIC-END these quantifiers are not defined aside

[2] In fact, this characterization of $\land$ and $\lor$ is based on their characterizations in Natural Deduction and can be completed using their meta-inferences.

[3] As is custom, we will denote equality in infix notation, i.e. $a = b$ for the relation formally denoted by $= (a, b)$.

[4] This type of substitutivity is not a universal property. In fact, there are logics that cannot preserve it. Accordingly, it is presented here as a defeasible rule.

from the introduction of the universal quantifier. We will characterize the quantifiers fully with the following approach:[5] typical mathematical arguments about an “arbitrary” element of a set (i.e. ”Let x be an arbitrary ...”) can usually be modeled as arguing with a free variable and vice versa. We will introduce equivalent substitutions between terms with free variables and quantified terms.

While the ∀-intro in ASPIC-END has to ascertain that we do not generalize hypothetical free variables, the other quantifier operations can be achieved as rules (similar to the approach in [Vel06]). In order to ensure uniqueness for the introduced variable when eliminating quantifiers, we will use the special unary function symbol g in existential quantifier elimination. In order to be named apart from any other function application, the specific function application will refer to the formula that the existential quantifier was applied to.

We will use the following rule schemes to handle quantifiers:[6]

$$\begin{aligned}\forall\widetilde{x}.\widehat{p}(\widetilde{x}) &\quad\rightsquigarrow\quad \widehat{p}(\widetilde{\tau})\\ \widehat{p}(\widetilde{\tau}) &\quad\rightsquigarrow\quad \exists\widetilde{x}.\widehat{p}(\widetilde{x})\\ \exists\widetilde{x}.\widehat{p}(\widetilde{x}) &\quad\rightsquigarrow\quad \widehat{p}\Big(g\big(\langle\widehat{p}(\widetilde{x})\rangle\big)\Big)\end{aligned}$$

4.3 Domain-Specific Rules

In this section, we will introduce the ASPIC-END rules for modeling the philosophy of logic. We will begin with the overarching structure used to represent the different logical theories and how they can be compared. Then, we will present the implementations for the rules of inference presented in 2.2.1. Lastly, we will propose a way to use structural induction on terms in ASPIC-END.

4.3.1 The Structure for Different Logical Theories

One important point of Field’s explanations is that specific theories (which allow or reject some of the building block) vary in what specific results they can reach. As an example, let us consider KF. Field argues that this theory, while it accepts all rules of classical logic, rejects instances of the axiom T-IN[7] and thus not all axioms of the logic are true (we will present a more detailed representation of KF in 5.3.2).

Ultimately, our goal is to realize a framework where several theories can be described simultaneously and where their strengths and weaknesses (especially in terms of their impact on paradoxes) can be compared.

However, this framework must also ensure that no extension adopts contradictory theories at the same time. Instead, every extension should represent at least one, but only mutually consistent theories. This choice of theory, represented by the unary predicate

[5] This approach has also been outlined by Velleman in [Vel06].

[6] As a reminder: there will be separate rule instances for any pair of variables and terms of the language to replace $\widetilde{x}$ and $\widetilde{\tau}$ respectively.

[7] See 2.2.1 for a more complete introduction of T-IN and the other buidling blocks.

AdoptTheo, determines the acceptability of the building blocks, i.e. the rule schemes and axioms of philosophy of logic. We will use the binary meta-predicates *Rule*[8] and *Axiom* to denote which building blocks are accepted in a theory (they will be discussed in greater detail below).

It is important to remember that any proof-theoretical approach, including the one used in the philosophy of logic, distinguishes between axioms and rules. Even though it is a common convention in ASPIC-style structured argumentation frameworks to refer to any rule without antecedents as an *axiom*, this is not the case in the philosophy of logic. Here, we consider as an axiom (of a logical theory) any formula that holds unconditionally. Similarly, the *Axiom* predicate takes as its first argument a formula, while the *Rule* predicate uses the name of an instantiated rule.

Any building block that is not accepted by the adopted theory should not be usable within the corresponding extension. To ensure this, we require on the one hand that the rule naming function $n(\rho)$ for any inference rule of the philosophy of logic returns the string $Accept_r(\rho)$ (and thus, any occurrence of $\neg Accept_r(\rho)$ allows for undercut attacks against any use of the rule within that theory).

$$\begin{array}{rcl} AdoptTheo(\widetilde{t}), Rule(\widetilde{r}, \widetilde{t}) & \rightsquigarrow & Accept_r(\widetilde{r}) \\ AdoptTheo(\widetilde{t}), \neg Rule(\widetilde{r}, \widetilde{t}) & \rightsquigarrow & \neg Accept_r(\widetilde{r}) \end{array}$$

On the other hand, for axioms we require that the rule naming $n(r_{ax(x)})$ for any rule $r_{ax(x)}$ representing an axiom x returns $Accept_a(r_{ax(x)})$.[9]

$$\begin{array}{rcl} AdoptTheo(\widetilde{t}), Axiom(\langle\widehat{x}\rangle, \widetilde{t}) & \rightsquigarrow & Accept_a(r_{ax(\widehat{x})}) \\ AdoptTheo(\widetilde{t}), \neg Axiom(\langle\widehat{x}\rangle, \widetilde{t}) & \rightsquigarrow & \neg Accept_a(r_{ax(\widehat{x})}) \end{array}$$

It is important to note that Field assumes in his paper that all logical reasoning must be based in one of the logics under consideration.Consequently, the described ASPIC-END system requires a mechanism to adopt one of the theories. While intuitively, this approach might resemble hypothetical reasoning (i.e. "If we chose to adopt this logic, then X and Y would follow..."),[10] there are however some problems with this approach. Most importantly, representing the different theories of truth by hypotheticals would not be authentic to the philosophy of logic, where these theories are treated as fully existent. Furthermore, this approach would introduce issues on the technical level: a hypothetical argument cannot attack other arguments, meaning that no actual comparison between

[8] Notably, there is an inevitable mismatch between rules as defined by Field and the rules we refer to in ASPIC-END: "Rules" in Field's sense of the word refer to rule schemata, which can be instantiated into individual rule instances, and do not include axiom schemata. All rules in ASPIC-END are fully instantiated and axioms are not treated as an inherently separate concept from rules.

[9] This presents a certain symmetry between rules and axioms between ASPIC-END: they are both represented by binary predicates and can be rejected via the rules that represent them, using representatives for the naming function n.

[10] An argument is hypothetical if it is based on one or more non-substantiated assumptions, i.e. $As(A) \neq \emptyset$ for an argument A in ASPIC-END.

the different theories would be possible. Furthermore, a hypothetical argument cannot contain defeasible inferences and would limit the theories to intuitively strict rules.

Instead, we will provide separate defeasible rules of the form $\Rightarrow \mathit{AdoptTheo}(t)$ for all theories t under consideration. These rules can be seen as a "justification" for choosing the theory. While it might seem as though this mechanism would allow multiple theories to be chosen simultaneously, this is not so: Contradictions between theories will result in rebuttals between their arguments, ensuring that incongruent theories cannot be adopted at the same time.[11] We will illustrate this rebuttal in more detail in 6.1.3.

There is however a problem with the proposed approach: if logical theories are represented by extensions, then typical Dung-style reasoning would only accept those theories that are admissible. And admissibility would require consistency, disregarding those theories of philosophy of logic that are paraconsistent. As noted in 2.2.3, we have decided to exclude paraconsistent logics from the scope of this thesis. Luckily, the logics we have chosen for modeling are all classically consistent. We will discuss a possible solution for reasoning with paraconsistency in 6.2.4.

Important Concepts

In order to fully implement the different logics as both object of argumentation and the basis on which to discuss, we require definitions for the most basic concepts of logical theories: *rules*, *theorems* and *negation*.

Here, we will use the binary predicate *Rule* (see 4.3.1) to denote all rules and which theories they hold in. A binary *Ante* predicate connects these rules to all their corresponding antecedents and a unary *conc* function denotes the unique conclusion of a rule.[12] To illustrate this, we will here provide an example for an instance of the (generally-accepted) modus ponens: $A, A \to B \vdash B$.

$$\begin{aligned} & \rightsquigarrow && \mathit{Rule}\big(r_{MP}(\langle A\rangle, \langle B\rangle), t\big) \\ \mathit{Ante}\big(a, r_{MP}(\langle A\rangle, \langle B\rangle)\big) & \rightsquigarrow && a = \langle A\rangle \lor a = \langle A \to B\rangle \\ & \rightsquigarrow && \mathit{conc}\big(r_{MP}(\langle A\rangle, \langle B\rangle)\big) = \langle B\rangle \end{aligned}$$

With this, we can provide rules defining theorems within these theories:[13]

$$\begin{aligned} \mathit{Axiom}(\widetilde{s}, \widetilde{t}) & \rightsquigarrow && \mathit{Theorem}(\widetilde{s}, \widetilde{t}) \\ \forall \widetilde{a}.[\mathit{Ante}(\widetilde{a}, \widetilde{r}) \to \mathit{Theorem}(\widetilde{a}, \widetilde{t})] & \rightsquigarrow && \mathit{Theorem}\big(\mathit{conc}(\widetilde{r}), \widetilde{t}\big) \end{aligned}$$

[11] While this setup makes it impossible to adopt incongruent theories simultaneously, it might however be possible (and even preferable) to adopt multiple congruent theories at the same time. After all, theories can only be congruent if they do not contradict one another on what inference rules and axioms to in- or exclude. Accordingly, this setup can only adopt multiple theories when they are only more specific on some of the rules. This might be desirable when one of the theories is an "instantiation" of the other (broader) theory.

[12] The conclusion is in fact unique since every rule we consider in the theories is assumed to be fully instantiated.

[13] A *theorem* refers to an entailed sentence of a theory. This entailment can be shown by providing a proof from axioms.

Furthermore, in the following modeling, contradictory sentences are presented with a separate *neg* function symbol instead of as $\langle \neg \hat{p} \rangle$. This is done so that properties for complementary sentences can be discussed on variables, i.e. arbitrary elements, instead of via meta-variables that are instantiated into an infinite number of separate elements.

$$\rightsquigarrow \quad neg(\langle \hat{p} \rangle) = \langle \neg \hat{p} \rangle$$

A common syntactical approach in languages that allow self-reference is to specify that a property holds for every *formula*. In order to replicate this in ASPIC-END, we will define a unary predicate *Form* that holds for all formulae.

It can be characterized very straightforwardly.

$$\rightsquigarrow \quad \mathit{Form}(\langle \widehat{p} \rangle)$$

The $True$-Predicate

The aim of this thesis is to represent different theories of truth as used in the philosophy of logic. To our mind, each theory of truth aims to express theory-specific properties of *the* truth meta-predicate. They express proof-theoretic principles for reasoning with sentences (esp. those involving the truth-predicate), thereby allowing justifications for the truth of sentences within the theory. Furthermore, they usually also explain how truth-based paradoxes (e.g. the Liar and Curry's paradox, see 2.2.2) can be avoided or their influence contained.

In accordance with this, we have decided to represent the truth predicate as applying generally, i.e. without specifying the theory a sentence is regarded as true in. This unary $True$ predicate takes as arguments "true" sentences of the modeling language itself (and will use the meta-logical terms introduced in 3.3.1 for this). This approach has the advantage of emphasizing contradictory truth results between different theories (and their corresponding extensions): since they are using the same truth predicate, they will attack one another on their truth findings (and thus ensure separation of different theories of truth).

This approach is specific to the $True$ predicate. Other properties that are specific to one theory of truth[14] will be realized with predicates that specifies the theory asa separate argument.

Fundamental Assumptions of Theories of Truth

As Field states in his argument, is is assumed for any theory of truth that all inference rules of that theory are *truth-preserving.*[15] Here, truth preservation (denoted as the unary predicate *TPres*) can be expressed with the following rule:

$$\rightsquigarrow \quad \mathit{DefinedAs}\Big(\langle \mathit{TPres}(\widetilde{r}) \rangle, \big\langle \big[\forall a.[\mathit{Ante}(\widetilde{r}, a) \rightarrow \mathit{True}(a)]\big] \rightarrow \mathit{True}(\mathit{conc}(\widetilde{r})) \big\rangle\Big)$$

[14] For example, what rules they accept and what sentences they entail as theorems.

[15] A rule is called truth-preserving if its conclusion is true when all its antecedents of the rule are true.

We will use the binary predicate *DefinedAs* to denote any bidirectional definitions in this thesis.[16] We will express this bidirectionality with the following rules:

$$\begin{array}{rcl} \mathit{DefinedAs}(\langle\widehat{p}\rangle, \langle\widehat{q}\rangle), \widehat{p} & \rightsquigarrow & \widehat{q} \\ \mathit{DefinedAs}(\langle\widehat{p}\rangle, \langle\widehat{q}\rangle), \widehat{q} & \rightsquigarrow & \widehat{p} \end{array}$$

As mentioned above, it is generally assumed in the theory of logic that all inference rules contained within a theory are truth-preserving:

$$r_{AssumeTPres} \quad \Rightarrow \quad \forall\widetilde{r}.\forall\widetilde{t}.\big[[\mathit{Rule}(\widetilde{r},\widetilde{t}) \wedge \mathit{AdoptTheo}(\widetilde{t})] \rightarrow \mathit{TPres}(\widetilde{r})\big]$$

Another important assumption to reasoning in the philosophy of logic is that a theory regards all its axioms as true. This can easily be represented in ASPIC-END as well:

$$r_{AssumeAxiom} \quad \Rightarrow \quad \forall\widetilde{x}.\forall\widetilde{t}.\big[[\mathit{Axiom}(\widetilde{x},\widetilde{t}) \wedge \mathit{AdoptTheo}(\widetilde{t})] \rightarrow \mathit{True}(\widetilde{x})\big]$$

As mentioned in 2.1.1, the usage of Gödel codes allows a logical theory to refer to its own syntactical objects. This construction is essential to the truth predicate and several central results of the theory of logic.[17] Furthermore, Gödel codes are one of the ways to achieve self-reference, central to some of the most important paradoxes of logic such as the Liar's paradox or the Curry paradox. In 3.3.1, we introduced *meta-logical terms* as a simplified version of Gödel codes for the logic language used in our ASPIC-END argumentation theory. Notably, any theory introduced in this thesis may (by the definition of our language) use meta-logical terms. This was by design: In his paper, Field analyzes only logics that use the *True* predicate (and therefore presupposes Gödel codes). In accordance with his presupposition, we will generally assume that all theories under consideration are capable of self-reference.

$$r_{AssumeSRef} \quad \Rightarrow \quad \forall\widetilde{t}.\mathit{PermitSelfref}(\widetilde{t})$$

4.3.2 Realizing the "Building Blocks" of Proof-Theoretic Logics

As noted in 2.2.1, logics in the philosophy of logic can be distinguished by what "building blocks," i.e. rules of inference and foundational axioms, they admit or reject (as discussed in 2.2.1).

Since they are realized differently in ASPIC-END, we will handle rules of inference and axioms separately. First, we will introduce the formalization for the rules of inference.

[16] The *DefinedAs* predicate allows an easier identification of pathological sentences, as described in 5.3.3.

[17] Such as Gödel's Incompleteness theorems or Tarski's related Undefinability theorem.

Rules of Inference

As a matter of fact, all the rules of inference must be realized as rule schemes in ASPIC-END. In order to be able to refer to individual instances of these rules, we will name them in a consistent manner: When referring to the complete set of all instances of a rule scheme, we will use r_{ab} (we will use the abbreviation introduced in 2.2.1, denoted by ab, as a reference to the rules), while instances will specify the placeholder terms in order of appearance in the rule. For example, the rule scheme for Double Negation Elimination can be referred to as r_{DNE}, while the specific instantiation $\neg\neg \mathit{Liar} \rightsquigarrow \mathit{Liar}$ can be denoted as $r_{DNE}(\langle \mathit{Liar} \rangle)$.

Then, these rules of inference can be represented in ASPIC-END as:

$$\begin{array}{lrcl} r_{MP}\colon & \widehat{p}_1, \widehat{p}_1 \rightarrow \widehat{p}_2 & \rightsquigarrow & \widehat{p}_2 \\ r_{DNE}\colon & \neg\neg\widehat{p} & \rightsquigarrow & \widehat{p} \\ r_{PE}\colon & \widehat{p} \wedge \neg\widehat{p} & \rightsquigarrow & \bot \end{array}$$

Furthermore, all of the logics under consideration accept several less-controversial rules of inference as well: the deMorgan laws[18] and conditional contraposition.[19] Notably, the deMorgan laws encompass more than one rule, and are here represented using the constant symbol dml.

$$\begin{array}{lrcl} & \mathit{Rule}(dml, \widetilde{t}) & \rightsquigarrow & \mathit{Rule}(r_{DML\_1}(\langle\widehat{p}\rangle, \langle\widehat{q}\rangle), \widetilde{t}) \wedge \mathit{Rule}(r_{DML\_2}(\langle\widehat{p}\rangle, \langle\widehat{q}\rangle), \widetilde{t}) \wedge \\ & & & \mathit{Rule}(r_{DML\_3}(\langle\widehat{p}\rangle, \langle\widehat{q}\rangle), \widetilde{t}) \wedge \mathit{Rule}(r_{DML\_4}(\langle\widehat{p}\rangle, \langle\widehat{q}\rangle), \widetilde{t}) \\ r_{DML\_1}\colon & \neg\widetilde{p} \vee \neg\widetilde{q} & \rightsquigarrow & \neg[\widetilde{p} \wedge \widetilde{q}] \\ r_{DML\_2}\colon & \neg[\widetilde{p} \vee \widetilde{q}] & \rightsquigarrow & \neg\widetilde{p} \wedge \neg\widetilde{q} \\ r_{DML\_3}\colon & \neg\widetilde{p} \wedge \neg\widetilde{q} & \rightsquigarrow & \neg[\widetilde{p} \vee \widetilde{q}] \\ r_{DML\_4}\colon & \neg[\widetilde{p} \wedge \widetilde{q}] & \rightsquigarrow & \neg\widetilde{p} \vee \neg\widetilde{q} \\ r_{cCP}\colon & \widetilde{p} \rightarrow \widetilde{q} & \rightsquigarrow & \neg\widetilde{q} \rightarrow \neg\widetilde{p} \end{array}$$

4.3.3 Proof-Theoretic Axioms

As noted in 4.3.1, the proof-theoretic approach of many theories of the philosophy of logic distinguishes between rules and axioms. In particular, while rules refer to inference mechanisms consisting of multiple formulae, an axiom is a single formula and holds without any inferences. Accordingly, axioms in ASPIC-END are represented by the binary *Axiom* predicate, referring to the formula and a theory that this formula is an axiom of. Similar to the *Rule* predicate, an axiom that is rejected by the adopted theory can be undercut.

In 2.2.1, we have presented some of the most important axioms for different theories of truth. Here, we can represent them in ASPIC-END in the following manner:

[18] The deMorgan laws consist of 4 laws expressing the interplay between conjunction, disjunction and negation in classical logic, e.g. $\neg[A \wedge B] \vdash \neg A \vee \neg B$.

[19] Contraposition is the rule in classical logic that an implication $A \rightarrow B$ is equivalent to the "reverse" $\neg B \rightarrow \neg A$.

$$\begin{array}{lll} \rightsquigarrow & \neg[\widehat{p} \wedge \neg\widehat{p}] & \text{Law of the Excluded Middle} \\ \rightsquigarrow & \widehat{p} \rightarrow True\big(\langle\widehat{p}\rangle\big) & \text{T-IN} \\ \rightsquigarrow & True\big(\langle\widehat{p}\rangle\big) \rightarrow \widehat{p} & \text{T-OUT} \end{array}$$

In order to increase readability, we will introduce *function symbols* (as "names") for the any instantiations of the proof-theoretic axioms.

$$\begin{array}{ll} \rightsquigarrow & ax_{LEM}\big(\langle\widehat{p}\rangle\big) = \langle\widehat{p} \vee \neg\widehat{p}\rangle \\ \rightsquigarrow & ax_{T\_IN}\big(\langle\widehat{p}\rangle\big) = \big\langle\widehat{p} \rightarrow True\big(\langle\widehat{p}\rangle\big)\big\rangle \\ \rightsquigarrow & ax_{T\_OUT}\big(\langle\widehat{p}\rangle\big) = \big\langle True\big(\langle\widehat{p}\rangle\big) \rightarrow \widehat{p}\big\rangle \end{array}$$

4.3.4 The Intersubstitutivity Principle

As mentioned in 2.2.1, Field has proposed an alternative way of characterizing the T-scheme: Intersubstitutivity. Field defined this property as follows:

> **Intersubstitutivity Principle:** If A and B are alike except that (in some transparent context) one has 'p' while the other has '$\langle p\rangle$ is true', then one can legitimately infer B from A and A from B. ([Fie06, p. 583])

While Field does not present a formalized definition of "transparent" contexts (he only notes that they should be "non-quotational, non-intentional, etc." ([Fie06])) we propose the following inductive definition for two sentences being "alike:"

$$\begin{array}{rcl} & \rightsquigarrow & Alike_T\big(\langle\widehat{p}\rangle, \langle True(\langle\widehat{p}\rangle)\rangle\big) \\ Alike_T\big(\langle\widehat{p}\rangle, \langle\widehat{q}\rangle\big) & \rightsquigarrow & Alike_T\Big(\big\langle True(\langle\widehat{p}\rangle)\big\rangle, \big\langle True(\langle\widehat{q}\rangle)\big\rangle\Big) \\ Alike_T\big(\langle\widehat{p}\rangle, \langle\widehat{q}\rangle\big) & \rightsquigarrow & Alike_T\big(\langle\widehat{p} \wedge \widetilde{r}\rangle, \langle\widehat{q} \wedge \widetilde{r}\rangle\big) \\ Alike_T\big(\langle\widehat{p}\rangle, \langle\widehat{q}\rangle\big) & \rightsquigarrow & Alike_T\big(\langle\widetilde{r} \wedge \widehat{p}\rangle, \langle\widetilde{r} \wedge \widehat{q}\rangle\big) \end{array}$$

The last two rules should be done analogously for the connectives $\vee$, $\rightarrow$, $\neg$, $\exists\widetilde{x}.$ and $\forall\widetilde{x}.$ as well. Then, Intersubstitutivity can be characterized with the following rule scheme:

$$\begin{array}{llcl} r_{Intersub\_1}(\langle\widehat{p}\rangle, \langle\widehat{q}\rangle): & Alike_T(\langle\widehat{p}\rangle, \langle\widehat{q}\rangle), \widehat{p} & \rightsquigarrow & \widehat{q} \\ r_{Intersub\_2}(\langle\widehat{p}\rangle, \langle\widehat{q}\rangle): & Alike_T(\langle\widehat{p}\rangle, \langle\widehat{q}\rangle), \widehat{q} & \rightsquigarrow & \widehat{p} \end{array}$$

With these rules, it is possible to legitimately *infer* from one alike situation to another. However, some logics reject the Intersubstitutivity principle. Similarly to the rules of inference, this rejection can be realized using the *Rule* predicate, enabling an undercut attack on any application of Intersubstitutivity, should it not be accepted.

4.3.5 Inductive Reasoning

In his standard argumentation, Field relies on the inference rule of *induction.* While it is not explicitly stated whether Field utilizes induction over natural numbers or any structure, Field's argumentation is closer to the practice of structural induction. In keeping with the principle of authenticity, we have decided to formalize his arguments with structural induction.

In structural induction, a property $\widehat{p}$ can be proven on a recursively defined structure[20] by showing that

1. $\widehat{p}$ holds for all base cases, **and**

2. every step in the recursive definition of the structure preserves the property (i.e. if it holds for all sub-elements, then it also holds for the combined element).

In particular, Field used induction on the structure of theorems to prove properties for them. We will present the induction rule scheme for this case, since it is the most pertinent to this report. Other structural induction schemes can be constructed in a similar manner. For arguments, the base cases consist of all axioms of the theory, while composite arguments can be defined by their last rule application on sub-arguments. In ASPIC-END, the structural induction for theorems can be expressed in the following manner:

$$\begin{array}{l} \quad\quad\;\; \forall\widetilde{x}.[\mathit{Axiom}(\widetilde{x},\widetilde{t}) \to \widehat{p}(\widetilde{x})], \\ \forall\widetilde{r}.\Big[\mathit{Rule}(\widetilde{r},\widetilde{t}) \to \big[\forall\widetilde{a}.[\mathit{Ante}(\widetilde{a},\widetilde{r}) \to \widehat{p}(\widetilde{a})] \\ \quad\quad\quad\quad\quad \to \widehat{p}\big(\mathit{conc}(\widetilde{r})\big)\big]\Big] \end{array} \quad \rightsquigarrow \quad \forall\widetilde{x}.[\mathit{Theorem}(\widetilde{x},\widetilde{t}) \to \widehat{p}(\widetilde{x})]$$

Importantly, while mathematical induction is usually presented as a rule scheme of axioms (and the same can be done here in ASPIC-END), we have chosen to present induction as a scheme of inference rules for multiple reasons. Firstly, this change increases readability, as the conditions for and the conclusion of induction need not appear in any argument conclusion together. Furthermore, this change is in keeping with Field's presentation of induction (see [Fie06, p. 587]). While this change is not significant for any classical logic (where → -Intro makes both forms equivalent), it might be relevant for the study of non-classical logics without → -Intro, such as the paracomplete theories discussed in 5.3.3. Luckily, in this thesis we have not encountered reasoning where it would be necessary to represent induction as a conditional instead of a rule.

4.4 Notable Paradoxes

As noted in 2.2.2, certain paradoxical sentences, mainly the well-known Liar sentence[21] and the Curry sentence[22] are a challenge to many of the logics presented by Field. Classical logic produces problematic results for both these sentences: assigning any truth

[20] Structural induction on $\mathcal{N}$ is the same as induction according to the Peano induction scheme, while structural induction on decision-trees or logical formulae has no obvious equivalent in this kind of induction.

[21] i.e. "This sentence is false"

[22] i.e. "If this sentence is true, then the moon is made of cheese."

value to the Liar sentence leads to inconsistency, and the Curry sentences can be used to infer any conclusion (such as the "moon being made of cheese"). While it has been shown in the philosophy of logic that these paradoxes can be constructed using very basic syntactical constructions, formalizing the syntactical construction process that these paradoxes rely upon in ASPIC-END would be beyond the scope of this thesis. Instead, we will represent the Liar and Curry sentence as nullary predicates *Liar* and *Curry* that are connected to their definition with the binary *DefinedAs* predicate. Then, the pertinent properties of *Liar* and *Curry* can be expressed in the following way:[23]

$$\begin{array}{lll} r_{DefLiar}\colon & \rightsquigarrow & \mathit{DefinedAs}\big(\langle \mathit{Liar}\rangle, \big\langle \neg True(\langle \mathit{Liar}\rangle)\big\rangle\big) \\ r_{DefCurry}\colon & \rightsquigarrow & \mathit{DefinedAs}\big(\langle \mathit{Curry}\rangle, \big\langle True(\langle \mathit{Curry}\rangle) \rightarrow CheeseMoon\big\rangle\big) \end{array}$$

As noted in 4.3.1, the usage of meta-logical terms is only possible in theories that allow self-reference. Since the logic we have presented generally allows the usage of meta-logical terms, we have (implicitly) assumed that all theories under consideration are self-referential. However, it might be desirable to consider logics without self-reference at some point. In such theories, the Liar's paradox as well as the Curry paradox cannot be defined. Accordingly, the defining rules for the paradoxes should be undercut.

$$\begin{array}{lll} \neg \mathit{PermitSelfref}(\widetilde{t}), \mathit{AdoptedTheo}(\widetilde{t}) & \rightsquigarrow & \neg n(r_{DefLiar}) \\ \neg \mathit{PermitSelfref}(\widetilde{t}), \mathit{AdoptedTheo}(\widetilde{t}) & \rightsquigarrow & \neg n(r_{DefCurry}) \end{array}$$

[23] This definition using the $DefinedAs$ predicate will be relevant for the definition of pathological sentences in 5.3.3.

Chapter 5

Representing Field's Arguments

5.1 Field's "Truth and the Unprovability of Consistency"

As mentioned in 1.1, the focus of this thesis is to analyze and represent the argumentations Field'a arguments from the article "Truth and the Unprovability" ([Fie06]) as a representative for the philosophy of logic. This article provides a well-rounded and clearly argued account of several logics important to the field and highlights some connections between philosophy of logic and logic in mathematics. Whenever useful, we will quote pertinent passages of Field's article to serve as an explanation for our modeling decisions.

In his article, Field provides a seemingly simple meta-logical argument, the "Consistency Argument," for how any logic should be able to prove its own consistency. He then notes how this would run counter to Gödel's Incompleteness theorem and concludes that this Consistency argument cannot be correct for any logical theory. Field then proceeds to introduce and examine different logics, discussing how they interfere with the Consistency Argument.

As mentioned in 3.3, the definition of meta-terms and rule schemes in the logical language resulted in an countably infinite argument space. In the sections below, we will explore the resulting argument-space without automated reasoning support and outline a select group of arguments that best represent Field's argumentation. We believe that this approach results in a conceptually insightful (even if not complete) exploration of the field of philosophy of logic. For the sake of readability, we will intersperse the ASPIC-END arguments with short explanations of the next steps.

5.2 The Consistency Argument

Here, we will present the Consistency Argument as used by Field in ASPIC-END. He outlines his Consistency argument in the following two steps:

> [W]e ought to be able to prove the consistency of a mathematical theory T within T by

> (i) inductively proving within T that all its theorems are true,
> and
> (ii) inferring from the truth of all theorems of T that T is consistent.
>
> ([Fie06, p. 567])

We will follow the same structure for the Consistency argument in ASPIC-END: in 5.2.1, we will present the structural induction that Field sketches to show that all theorems of T should be true, while in 5.2.2 we will model the two ways Field introduces how a theory can show its own consistency (when all its theorems are true). These two ways rely on different, but equally "simple," assumptions.

Notably, Field's argument is based on the idea that we use the logical inference rules of a theory of truth to show that that theory is consistent. Since the Consistency Argument is constructed generally (and should be considered in every theory), we will assume that an arbitrary theory is adopted (represented by a free variable t).

Let t be the underlying theory of our argument.
C_0: $\mathsf{Assume}_{\rightarrow}(AdoptTheo(t))$

5.2.1 Step (i): Showing Theorem Truth

As discussed in 4.3.5, we have found that Field's argument for step (i) is easiest represented by a proof via structural induction. It presents a more well-rounded argumentation theory and, in particular, allows us to disregard any musings on the length of proofs.[1]

In particular, Field shows separately that (1) all axioms are true and (2) that all rules of inference preserve truth.[2] Combined, these observations show that any theorem must be true.[3]

To model Field's Consistency argument properly, we will use an instantiation of the structural induction rule scheme introduced in 4.3.5:

[1] Representing Field's argument as a mathematical induction would require pinpointing a relevant, monotonously increasing, number that can be associated with the theorems, such as the length of theorem proofs. Using structural induction significantly shortens some of the resulting arguments in ASPIC-END.

[2] These two steps both use assumptions that we introduced as true for all proof-theoretic logics.

[3] Field sketches the induction as follows:

> [...]the reasoning for Step (i) is:
>
> 1. Each axiom of T is true
> 2. Each rule of inference preserves truth [...]
>
> Since a theorem of T is just a sentence that results from the axioms by successive application of rules of inference, a simple mathematical induction yields
>
> 3 All theorems of T are true
>
> ([Fie06, p. 568])

$$\begin{array}{l} \quad\forall x.[Axiom(x,t) \rightarrow True(x)], \\ \forall r.\Big[Rule(r,t) \rightarrow \big[\forall a.[Ante(a,r) \rightarrow True(a)] \\ \qquad\qquad \rightarrow True\big(conc(r)\big)\big]\Big] \end{array} \rightsquigarrow \forall x.[Theorem(x,t) \rightarrow True(x)]$$

Step 1: The Induction Start

The induction start can now be formalized:

`Let` r `be an arbitrary axiom.`

C_{1_1}: $\quad \mathsf{Assume}_{\rightarrow}\big(Axiom(x,t)\big)$

`We generally assume all axioms to be true.`

C_{1_2}: $\quad \Rightarrow \quad \forall x.\forall t.\big[[Axiom(x,t) \wedge AdoptTheo(t)] \rightarrow True(x)\big]$

C_{1_3}: $\quad \mathrm{C}_{1\_2} \rightsquigarrow \forall t.\big[[Axiom(x,t) \wedge AdoptTheo(t)] \rightarrow True(x)\big]$

C_{1_4}: $\quad \mathrm{C}_{1\_3} \rightsquigarrow [Axiom(x,t) \wedge AdoptTheo(t)] \rightarrow True(x)$

C_{1_5}: $\quad \mathrm{C}_{1\_1}, \mathrm{C}_0 \rightsquigarrow Axiom(x,t) \wedge AdoptTheo(t)$

C_{1_6}: $\quad \mathrm{C}_{1\_5}, \mathrm{C}_{1\_4} \rightsquigarrow True(x)$

`Then, the induction start is proven.`

C_{1_7}: $\quad \rightarrow\text{-}\mathsf{intro}\big(Axiom(x,t) \rightarrow True(x), \mathrm{C}_{1\_6}\big)$

C_{1_8}: $\quad \forall\text{-}\mathsf{intro}\big(\forall x.[Axiom(x,t) \rightarrow True(x)], \mathrm{C}_{1\_7}\big)$

Step 2: The Induction Step

The induction step can then be formalized in the following way:

`Consider an arbitrary theorem with a composite proof.`

`This proof has a last rule application.`

C_{2_1}: $\quad \mathsf{Assume}_{\rightarrow}\big(Rule(r,t)\big)$

`Furthermore, we require that all rules are truth-preserving.`

C_{2_2}: $\quad \Rightarrow \quad \forall r.\forall t.\big[[Rule(r,t) \wedge AdoptTheo(t)] \rightarrow TPres(r)\big]$

C_{2_3}: $\quad \mathrm{C}_{2\_2} \rightsquigarrow \forall t.\big[[Rule(r,t) \wedge AdoptTheo(t)] \rightarrow TPres(r)\big]$

C_{2_4}: $\quad \mathrm{C}_{2\_3} \rightsquigarrow [Rule(r,t) \wedge AdoptTheo(t)] \rightarrow TPres(r)$

C_{2_5}: $\quad \mathrm{C}_{2\_1}, \mathrm{C}_0 \rightsquigarrow Rule(r,t) \wedge AdoptTheo(t)$

C_{2_6}: $\quad \mathrm{C}_{2\_5}, \mathrm{C}_{2\_4} \rightsquigarrow TPres(r)$

`Accordingly, if the sub-proofs are true, then so is the theorem.`

C_{2_7}: $\quad \rightsquigarrow DefinedAs\Big(\langle TPres(r)\rangle, \big\langle\big[\forall a.[Ante(r,a) \rightarrow True(a)]\big] \rightarrow True\big(conc(r)\big)\big\rangle\Big)$

C_{2_8}: $\quad \mathrm{C}_{2\_6}, \mathrm{C}_{2\_7} \rightsquigarrow \big[\forall a.[Ante(a,r) \rightarrow True(a)]\big] \rightarrow True\big(conc(r)\big)$

`Then, the induction step can be made.`

C_{2_9}: $\to$ -intro$\big(Rule(r,t) \to \big[\forall a.[Ante(a,r) \to True(a)]\big] \to True\big(conc(r)\big), C_{2_8}\big)$

C_{2_10}: $\forall$-intro$\Big(\forall r.\Big[Rule(r,t) \to \big[\forall a.[Ante(a,r) \to True(a)]\big] \to True\big(conc(r)\big)\Big], C_{2_9}\Big)$

```
By structural induction, every theorem of t must be true.
```

C_{2_11}: $C_{1\_8}, C_{2\_10} \rightsquigarrow \forall x.[Theorem(x,t) \to True(x)]$

5.2.2 Step (ii): Concluding Consistency

In step (ii), Field argues that if the structural induction is correct and it can therefore be shown within the theory t that all theorems of t were true, then t must be *consistent* (meaning that t does not entail any contradictions, i.e. formulae φ and $\neg\varphi$). He presents two different ways to reach this conclusion, both relying on a simple, but not universal, assumption. Here, we will present the first approach that Field takes for this argument: he argues that no sentence and its negation should both be true (Field calls this assumption an *elementary property* of truth), and since all theorems of t are true (according to (i)), it then follows that no sentence and its negation could be theorems of t.[4] Since according to Gödel's Incompleteness theorem no logic can show its own consistency, there must be a blocking element either in this step or the structural induction of step (i).

We will use the following additional rules to characterize this "elementary property of truth" and consistent logics:

$$\begin{aligned} &\Rightarrow \forall x.\neg\big[True(x) \land True\big(neg(x)\big)\big] \\ Consistent(\widetilde{t}) &\rightsquigarrow \forall x.[Theorem(x,\widetilde{t}) \to \\ &\qquad \neg Theorem\big(neg(x),\widetilde{t}\big)] \\ \forall x.[Theorem(x,\widetilde{t}) \to \neg Theorem\big(neg(x),\widetilde{t}\big)] &\rightsquigarrow Consistent(\widetilde{t}) \end{aligned}$$

Then the resulting argument for part (ii) can be constructed in the following manner in ASPIC-END:

```
Let x be a theorem of t.
```

C_{3_1}: Assume$_{\to}\big(Theorem(x,t)\big)$

```
All theorems of T are true.
```

C_{3_2}: $C_{2\_11} \rightsquigarrow Theorem(x,t) \to True(x)$

[4] Field presents this reasoning as follows:

> The reasoning of Step (ii) is that by elementary property of truth, no sentence and its negation can both be true. So from (3) [the structural induction] it follows that no sentence and its negation can both be theorems of T, which is to say that T is consistent. ([Fie06, p.568])

Then x is true.

$$\begin{array}{rrcl}
\mathrm{C}_{3\_3}: & \mathrm{C}_{3\_1}, \mathrm{C}_{3\_2} & \leadsto & True(x)
\end{array}$$

By elementary properties of truth, x cannot be true at the same time as its negation.

$$\begin{array}{rrcl}
\mathrm{C}_{3\_4}: & & \Rightarrow & \forall x.\neg\big[True(x) \wedge True\big(neg(x)\big)\big] \\
\mathrm{C}_{3\_5}: & \mathrm{C}_{3\_4} & \leadsto & \neg[True(x) \wedge True\big(neg(x)\big)]
\end{array}$$

Then, the negation of x cannot be a theorem.

$$\begin{array}{rrcl}
\mathrm{C}_{3\_6}: & & & \mathsf{Assume}_{\neg}\Big(Theorem\big(neg(x), t\big)\Big) \\
\mathrm{C}_{3\_7}: & \mathrm{C}_{2\_11} & \leadsto & Theorem\big(neg(x), t\big) \rightarrow True\big(neg(x)\big) \\
\mathrm{C}_{3\_8}: & \mathrm{C}_{3\_6}, \mathrm{C}_{3\_7} & \leadsto & True\big(neg(x)\big) \\
\mathrm{C}_{3\_9}: & \mathrm{C}_{3\_3}, \mathrm{C}_{3\_8} & \leadsto & True(x) \wedge True\big(neg(x)\big) \\
\mathrm{C}_{3\_10}: & \mathrm{C}_{3\_9}, \mathrm{C}_{3\_5} & \leadsto & [True(x) \wedge True\big(neg(x)\big)] \wedge \neg[True(x) \wedge True\big(neg(x)\big)] \\
\mathrm{C}_{3\_11}: & \mathrm{C}_{3\_10} & \leadsto & \bot \\
\mathrm{C}_{3\_12}: & & & \mathsf{ProofByContrad}\big(\neg Theorem\big(neg(x), t\big), \mathrm{C}_{3\_11}\big)
\end{array}$$

Then T is consistent.

$$\begin{array}{rrcl}
\mathrm{C}_{3\_13}: & & & \rightarrow\text{-}\mathsf{intro}\Big(Theorem(x, t) \rightarrow \neg Theorem\big(neg(x), t\big), \mathrm{C}_{3\_12}\Big) \\
\mathrm{C}_{3\_14}: & \mathrm{C}_{3\_13} & \leadsto & \forall x.[Theorem(x, t) \rightarrow \neg Theorem\big(neg(x), t\big)] \\
\mathrm{C}_{3\_15}: & \mathrm{C}_{3\_14} & \leadsto & Consistent(t) \\
\mathrm{C}_{3\_15}: & & & \rightarrow\text{-}\mathsf{intro}\big(AdoptTheo(t) \rightarrow Consistent(t), \mathrm{C}_{3\_15}\big)
\end{array}$$

Variant Argument for Consistency

Furthermore, Field proposed a variant argument for consistency. Instead of relying on the introduced “elementary property” of truth, Field uses another assumption that is usually true for logics to show consistency: the assumption being that there must be sentences in the logic that are regarded as not true and which are consequently not theorems the theory. Accordingly, the logic cannot be trivial[5] (and is usually consistent).[6]

We will use the following additional rules to characterize triviality and Field’s assumption:

$$\begin{array}{rcl}
& \Rightarrow & \exists x.\neg True(x) \\
Trivial(\widetilde{t}) & \leadsto & \forall x.Theorem(x, \widetilde{t}) \\
Trivial(\widetilde{t}) & \leadsto & \neg Consistent(\widetilde{t}) \\
\neg Consistent(\widetilde{t}) & \leadsto & Trivial(\widetilde{t})
\end{array}$$

[5] As a reminder, a logic is considered trivial if all sentences are theorems of the logic. This is usually seen as an even more damning form of inconsistency.

[6] In the article, Field argues that “some sentences are not true, so by (3) [the structural induction] some sentences are not theorems of T, and so T is consistent” ([Fie06, p. 568]).

Notably, this logic contains a inference from inconsistency to triviality. This is not an universal truth in the philosophy of logic since any paraconsistent logic does not permit this inference. Consequently, in this formalization paraconsistent logics will reject this inference rule.

This variant argument for step (ii) can be expressed in ASPIC-END in the following way:[7]

`There are sentences not considered true.`

$$\begin{array}{llcl}
\mathrm{C}_{4\_1}: & & \Rightarrow & \exists x.\neg True(x)
\end{array}$$

`These sentences cannot be theorems of the theory.`

$$\begin{array}{rrcl}
\mathrm{C}_{4\_2}: & \mathrm{C}_{4\_1} & \rightsquigarrow & \neg True\Big(g\big(\langle\neg True(x)\rangle\big)\Big) \\
\mathrm{C}_{4\_3}: & & & \mathsf{Assume}_{\neg}\Big(Theorem\Big(g\big(\langle\neg True(x)\rangle\big)\Big), t\Big) \\
\mathrm{C}_{4\_4}: & \mathrm{C}_{2\_11} & \rightsquigarrow & Theorem\Big(g\big(\langle\neg True(x)\rangle\big), t\Big) \to True\Big(g\big(\langle\neg True(x)\rangle\big)\Big) \\
\mathrm{C}_{4\_5}: & \mathrm{C}_{4\_3}, \mathrm{C}_{4\_4} & \rightsquigarrow & True\Big(g\big(\langle\neg True(x)\rangle\big)\Big) \\
\mathrm{C}_{4\_6}: & \mathrm{C}_{4\_5}, \mathrm{C}_{4\_2} & \rightsquigarrow & True\Big(g\big(\langle\neg True(x)\rangle\big)\Big) \wedge \neg True\Big(g\big(\langle\neg True(x)\rangle\big)\Big) \\
\mathrm{C}_{4\_7}: & \mathrm{C}_{4\_6} & \rightsquigarrow & \bot \\
\mathrm{C}_{4\_8}: & & & \mathsf{ProofByContrad}\Big(\neg Theorem\Big(g\big(\langle\neg True(x)\rangle\big), t\Big), \mathrm{C}_{4\_7}\Big)
\end{array}$$

`Then, the theory is not trivial and therefore consistent.`

$$\begin{array}{rrcl}
\mathrm{C}_{4\_9}: & & & \mathsf{Assume}_{\neg}\big(\neg Consistent(t)\big) \\
\mathrm{C}_{4\_10}: & \mathrm{C}_{4\_9} & \rightsquigarrow & Trivial(t) \\
\mathrm{C}_{4\_11}: & \mathrm{C}_{4\_10} & \rightsquigarrow & \forall x.Theorem(x, t) \\
\mathrm{C}_{4\_12}: & \mathrm{C}_{4\_11} & \rightsquigarrow & Theorem\Big(g\big(\langle\neg True(x)\rangle\big), t\Big) \\
\mathrm{C}_{4\_13}: & \mathrm{C}_{4\_12}, \mathrm{C}_{4\_8} & \rightsquigarrow & Theorem\Big(g\big(\langle\neg True(x)\rangle\big), t\Big) \wedge \neg Theorem\Big(g\big(\langle\neg True(x)\rangle\big), t\Big) \\
\mathrm{C}_{4\_14}: & \mathrm{C}_{4\_13} & \rightsquigarrow & \bot \\
\mathrm{C}_{4\_15}: & & & \mathsf{ProofByContrad}\big(\neg\neg Consistent(t), \mathrm{C}_{4\_14}\big) \\
\mathrm{C}_{4\_16}: & \mathrm{C}_{4\_15} & \rightsquigarrow & Consistent(t) \\
\mathrm{C}_{4\_17}: & & & \to\text{-}\mathsf{intro}\big(AdoptTheo(t) \to Consistent(t), \mathrm{C}_{4\_16}\big)
\end{array}$$

5.3 Logics under Consideration

After formulating his Consistency Argument in Chapter 1, Field begins his process of analyzing how different logics break the argument. Chapters 2 and 3 are spent establishing different categories the theories of truth fall into, while the later chapters introduce (and assess) individual theories.

[7] As a reminder: the function symbol g is introduced as the representative element when eliminating an universal quantifier. See 4.2.3.

In this section, we will first try to roughly outline Field's arguments in chapter 2 and 3, but our focus will be on some of the theories Field considers and attempt to express his arguments in ASPIC-END. To this end, we will paraphrase the chapters (and the main points of Field's arguments), introduce any new knowledge as ASPIC-END rules and subsequently provide ASPIC-END arguments for these chapters.

5.3.1 Field's Categorization of Approaches

As mentioned above, Field uses chapter 2 and 3 to introduce different categorizations for the theories under consideration. These categories are used in the rest of Field's paper, and even though they do not represent individual logics (the main focus of this thesis), it might still be insightful to attempt to model Field's thoughts on the different approaches here.

Chapter 2

In chapter 2, Field uses Tarski's well-known Undefinability Theorem (originally shown in [Tar36]) to conclude that any classical logic that does contain a general truth predicate cannot preserve the T-scheme completely.[8]

Using the Undefinability theorem, Field notes two ways that different logics have aimed to fix this gap: either as a classical logic that restricts the T-scheme in some manner *or* as a non-classical logic while preserving the T-scheme.[9] The Undefinability Theorem can be represented in ASPIC-END with the following rule:

$$\begin{aligned}\Rightarrow \quad \neg\exists t.\Big[&Classical(t) \wedge PermitSelfref(t) \wedge \forall x.\big[Form(x) \rightarrow \\ &[Axiom(ax_{T\_IN}(x), t) \wedge Axiom(ax_{T\_OUT}(x), t)]\big]\Big]\end{aligned}$$

This subdivision of theories of truth can now be derived in ASPIC-END rather straightforwardly (for the sake of simplicity, we will use $\varphi_{\neg\mathsf{TarUndef}}$ to abbreviate $\exists t.\Big[Classical(t) \wedge PermitSelfref(t) \wedge \forall x.\Big[Form(x) \rightarrow \big[Axiom\big(ax_{T\_IN}(x), t\big) \wedge Axiom\big(ax_{T\_OUT}(x), t\big)\big]\Big]\Big]$):

[8] The Undefinability Theorem asserts that any theory T that is sufficiently complex to express arithmetics and contains self-reference cannot define a sentence Tr (or predicate $True$) such that $\varphi \Leftrightarrow True(\langle\varphi\rangle)$ for every sentence φ. This equivalence, also called the T-scheme, is introduced in more detail in 2.1.1.

[9] Field chooses to frame Tarski's Undefinability theorem in the following manner in his paper:

> Of course, Tarski proved an important negative result about theories with unified truth predicates: he proved that no theory of truth (in a sufficiently rich metalanguage that permits self-reference) whose logic is classical can have a general truth predicate that obeys the truth schema [i.e. the T-scheme] ([Fie06, p. 569, 570])

We will use a case distinction to prove Field's subdivision.

D_{1_1}:		$\rightsquigarrow$	$Classical(t) \vee \neg Classical(t)$

First, for a classical t, the T-Scheme cannot be preserved.

D_{1_2}:			$\mathsf{Assume}_{\vee}\big(Classical(t)\big)$
D_{1_3}:		$\Rightarrow$	$\forall t. PermitSelfref(t)$
D_{1_4}:	D_{1_3}	$\rightsquigarrow$	$PermitSelfref(t)$
D_{1_5}:			$\mathsf{Assume}_{\neg}\Big(\forall x.\Big[Form(x) \to \big[Axiom\big(ax_{T\_IN}(x), t\big) \wedge Axiom\big(ax_{T\_OUT}(x), t\big)\big]\Big]\Big)$
D_{1_6}:	$\mathrm{D}_{1\_2}, \mathrm{D}_{1\_4}, \mathrm{D}_{1\_5}$	$\rightsquigarrow$	$Classical(t) \wedge PermitSelfref(t) \wedge \forall x.\Big[Form(x) \to \big[Axiom\big(ax_{T\_IN}(x), t\big) \wedge Axiom\big(ax_{T\_OUT}(x), t\big)\big]\Big]$
D_{1_7}:	D_{1_6}	$\rightsquigarrow$	$\varphi_{\neg\mathsf{TarUndef}}$
D_{1_8}:		$\Rightarrow$	$\neg\varphi_{\neg\mathsf{TarUndef}}$
D_{1_9}:	$\mathrm{D}_{1\_7}, \mathrm{D}_{1\_8}$	$\rightsquigarrow$	$\varphi_{TarUndef} \wedge \neg\varphi_{TarUndef}$
D_{1_10}:	D_{1_9}	$\rightsquigarrow$	$\bot$
D_{1_11}:			$\mathsf{ProofByContrad}\Big(\neg\forall x.\Big[Form(x) \to \big[Axiom\big(ax_{T\_IN}(x), t\big) \wedge Axiom\big(ax_{T\_OUT}(x), t\big)\big]\Big], \mathrm{D}_{1\_10}\Big)$
D_{1_12}:	D_{1_11}	$\rightsquigarrow$	$\neg Classical(t) \vee \neg\forall x.\Big[Form(x) \to \big[Axiom\big(ax_{T\_IN}(x), t\big) \wedge Axiom\big(ax_{T\_OUT}(x), t\big)\big]\Big]$

Otherwise, the theory might be non-classical.

D_{1_13}:			$\mathsf{Assume}_{\vee}\big(\neg Classical(t)\big)$
D_{1_14}:	D_{1_13}	$\rightsquigarrow$	$\neg Classical(t) \vee \neg\forall x.\Big[Form(x) \to \big[Axiom\big(ax_{T\_IN}(x), t\big) \wedge Axiom\big(ax_{T\_OUT}(x), t\big)\big]\Big]$

Then, a subdivision of approaches can be made.

D_{1_15}:			$\mathsf{ReasoningByCases}\Big(\neg Classical(t) \vee \neg\forall x.\Big[Form(x) \to \big[Axiom\big(ax_{T\_IN}(x), t\big) \wedge Axiom\big(ax_{T\_OUT}(x), t\big)\big]\Big], \mathrm{D}_{1\_12}, \mathrm{D}_{1\_14}, \mathrm{D}_{1\_1}\Big)$

Chapter 3

In chapter 3, Field rejects two technical "solutions" for the problematic Consistency Argument: on the one hand, he notes that an argument could be made against using the principle of induction on any statements containing the *True* predicate. He does note that this argument need not be entertained since neither are there real proponents for it nor has any reasonable logic resorted to this defense.

On the other hand, an argument could be made that any sufficiently powerful theory with an infinite number of axioms could be a problem: Field notes that in such a theory, the step from proving individually for each axiom that it is true to proving that all (infinitely many) axioms are true could be suspect. He argues that it is possible to construct a super-logic for any "infinite" logic such that the super-logic is based around a single, universally quantified rule that asserts truth for all infinitely many rules of the original logic. This is enough to reject the argument, since the Consistency argument cannot hold for this new logic either.

With these two technical "solutions" rejected, Field argues that the Consistency Argument must for *classical* logics break down on the induction for either a specific axiom or a specific rule. This is the only remaining point of attack since step (ii) of the Consistency Argument is classically valid and neither induction in general nor the universal quantification can be rejected.

It should be noted that the argumentation and the results of chapter 3 pose a more subdued role in the context of Field's paper. The two arguments he counteracts do not have any major proponents, and his conclusion is mostly relevant for the structure of the following chapters (i.e. finding specific instances of axioms or rules where the induction of the Consistency argument breaks down). For this reason, we have decided not to model chapter 3 in greater detail. Field's arguments have been explained here for the sake of completeness.

5.3.2 Classical Logics

As mentioned in 2.2.3, Field defines as "classical" all those theories where all arguments that are valid classically are taken as legitimate. Accordingly, classicality under Field's definition applies to a wide array of different logics that Field examines in the chapters 4 – 8 of the article.

In keeping with the subdivision Field introduced in chapter 2, Field will separately assess how the Consistency argument is blocked[10] for logics that reject only instances of T-IN (chapter 4) or T-OUT (chapter 5). In chapters 6 and 7, Field discusses weakly classical logics that do not specify which part of the T-scheme they reject. As explained in 2.2.3, weakly classical logics were beyond the scope of this thesis and are only mentioned here for the sake of completeness. Lastly, he discusses how the different classical theories treat the Intersubstitutivity Principle (chapter 8).

We express classicality in ASPIC-END by accepting all the standard building blocks – i.e. axioms and rules of inference – we have considered (see 4.3.2).

[10] As a reminder: by Gödel's Incompleteness theorem, it is clear that no consistent theory can prove its own consistency. Therefore, the Consistency argument Field constructed must be blocked for every consistent logical theory.

This can be achieved with the following rules in ASPIC-END:[11]

$$\begin{aligned}
Classical(\widetilde{t}) &\rightsquigarrow Rule\big(r_{DNE}(\langle\widehat{p}\rangle\rangle, \widetilde{t}\big)\\
Classical(\widetilde{t}) &\rightsquigarrow Rule\big(r_{MP}(\langle\widehat{p}_1\rangle, \langle\widehat{p}_2\rangle), \widetilde{t}\big)\\
Classical(\widetilde{t}) &\rightsquigarrow Rule\big(r_{PE}(\langle\widehat{p}\rangle), \widetilde{t}\big)\\
Classical(\widetilde{t}) &\rightsquigarrow Axiom\big(ax_{LEM}(\langle\widehat{p}\rangle), \widetilde{t}\big)
\end{aligned}$$

However, similar to the T-scheme, classical logics cannot accept the principle of Intersubstitutivity as introduced in 4.3.4. This rejection can be realized with the following rules:

$$\begin{aligned}
Classical(\widetilde{t}) &\rightsquigarrow \neg Rule(r_{Intersub\_1}(\langle\widehat{p}\rangle, \langle\widehat{q}\rangle), \widetilde{t})\\
Classical(\widetilde{t}) &\rightsquigarrow \neg Rule(r_{Intersub\_2}(\langle\widehat{p}\rangle, \langle\widehat{q}\rangle), \widetilde{t})
\end{aligned}$$

Chapter 4: Kripke-Feferman Logic

In chapter 4, Field assesses the classical approach of preserving all instances of the axiom T-OUT, while rejecting some instances of T-IN. As an example of this approach, he introduces the logic KF (named after two major contributors, Kripke and Feferman).

Field then shows that the instance of T-OUT for the Liar Sentence[12] is not regarded as true while being an axiom of the theory. This contradicts the induction start of the Consistency Argument.

The following rules can be used to characterize theories like KF (the axiom symbols used here are introduced in 4.3.3):

$$\begin{aligned}
&\Rightarrow AdoptTheo(KF)\\
&\rightsquigarrow Axiom\big(ax_{T\_OUT}(\langle\widehat{p}\rangle), KF\big)\\
r_{\neg T\_IN(\langle\widehat{p}\rangle)}: \quad &\rightsquigarrow \neg\big[Axiom\big(ax_{T\_IN}(\langle\widehat{p}\rangle), KF\big)\big]
\end{aligned}$$

In accordance with the Undefinability theorem (see 5.3.1), classical self-referential logics cannot accept the full T-scheme. Consequently, since KF accepts all instances of T-OUT, it cannot accept all instances of T-IN. However, since Field did not specify which instances of T-IN to reject, we will see all instances as suspect. Using $r_{\neg T\_IN(\langle\widehat{p}\rangle)}$, we can construct an undercut on any instance of the T-IN axiom (see 4.3.3). Accordingly, this formalization of KF (or rather, the extension representing it) cannot accept any argument that utilizes T-IN.

[11] Notably, for classical logics, the deMorgan rules and conditional contraposition can be proven using the standard building blocks and the meta-inferences of natural deduction. In keeping with the principle of minimality, they need not be explicitly included.

[12] The specific instance of the axiom scheme T-OUT is $True(\langle Liar\rangle) \rightarrow Liar$. For more information on the Liar sentence, see 2.2.2 .

However, it is still possible that a less strict definition of KF may wish to accept some unproblematic instances of T-IN while not accepting the full axiom scheme. To achieve this, we will require that the preference relation $<$ of the argumentation theory regards all instances of the rule $r_{\neg T\_IN(\langle\widehat{p}\rangle)}$ as minimal elements (which cannot successfully rebut an assertion that any instance is accepted).

In his article, before assessing where KF contradicts the Consistency argument, Field presented a preliminary result necessary for his later argument. He showed that while KF must accept the Liar sentence, KF also regards the Liar sentence as not true.[13] We can easily replicate his argument in ASPIC-END:

By the definition of the Liar sentence, KF accepts $Liar$.

E_{1_1}: $\quad \mathsf{Assume}_{\vee}\big(True(\langle Liar\rangle)\big)$

E_{1_2}: $\quad \rightsquigarrow \quad True(\langle Liar\rangle) \to Liar$

E_{1_3}: $\quad \mathrm{E}_{1\_1}, \mathrm{E}_{1\_2} \quad \rightsquigarrow \quad Liar$

E_{1_4}: $\quad \mathsf{Assume}_{\vee}\big(\neg True(\langle Liar\rangle)\big)$

E_{1_5}: $\quad \rightsquigarrow \quad DefinedAs(\langle Liar\rangle, \langle\neg True(\langle Liar\rangle)\rangle)$

E_{1_6}: $\quad \mathrm{E}_{1\_5}, \mathrm{E}_{1\_4} \quad \rightsquigarrow \quad Liar$

E_{1_7}: $\quad \rightsquigarrow \quad True(\langle Liar\rangle) \vee \neg True(\langle Liar\rangle)$

E_{1_8}: $\quad \mathrm{ReasonByCases}(Liar, \mathrm{E}_{1\_3}, \mathrm{E}_{1\_6}, \mathrm{E}_{1\_7})$

Also by the definition of the Liar sentence, KF regards $Liar$ as not true.

E_{1_9}: $\quad \mathrm{E}_{1\_5}, \mathrm{E}_{1\_8} \quad \rightsquigarrow \quad \neg True(\langle Liar\rangle)$

Field then aims to show that KF does not regard all of its axioms as true. We can now present his argument in two steps. In the first step, we will establish the equivalence between $True(\langle Liar\rangle) \to Liar$ (an instance of T-OUT) and $Liar$. In the second step, we will show that within KF $True(\langle Liar\rangle) \to Liar$ must be regarded as not true. Since this is an instance of T-OUT (and therefore, an axiom of KF), the Consistency argument must be blocked in step (1).[14]

[13] We will present Field's argument below for reference. He presented his argument for the Liar sentence in a footnote (as a note, Field used the symbol Q to represent the Liar sentence):

> We have both
>
> If $True(\langle Q\rangle)$, then Q
>
> (by (T-OUT)) and
>
> If $\neg True(\langle Q\rangle)$, then Q
>
> (by the equivalence between Q and $\neg True(\langle Q\rangle)$). Q follows; and by the equivalence, $\neg True(\langle Q\rangle)$ does too.
>
> ([Fie06, p. 572])

[14] For the sake of reference, we will present Field's proof that an instance of T-OUT is not regarded as true. As a reminder, in his paper. Field sketched the proof in the following way (as a note, Field referred to the Liar's sentence by Q):

We will show that $\mathit{True}(\langle \mathit{Liar}\rangle \to \mathit{Liar})$ entails Liar.

E_{2_1}: $\mathsf{Assume}_{\to}\big(\mathit{True}(\langle \mathit{Liar}\rangle) \to \mathit{Liar}\big)$

This implication can be written as a disjunction.

E_{2_2}: $\mathsf{Assume}_{\vee}\big(\mathit{True}(\langle \mathit{Liar}\rangle)\big)$

E_{2_3}: $\mathrm{E}_{2\_2}, \mathrm{E}_{2\_1} \rightsquigarrow \mathit{Liar}$

E_{2_4}: $\mathrm{E}_{2\_3} \rightsquigarrow \neg \mathit{True}(\langle \mathit{Liar}\rangle) \vee \mathit{Liar}$

E_{2_5}: $\mathsf{Assume}_{\vee}\big(\neg \mathit{True}(\langle \mathit{Liar}\rangle)\big)$

E_{2_6}: $\mathrm{E}_{2\_5} \rightsquigarrow \neg \mathit{True}(\langle \mathit{Liar}\rangle) \vee \mathit{Liar}$

E_{2_7}: $\rightsquigarrow \mathit{True}(\langle \mathit{Liar}\rangle) \vee \neg \mathit{True}(\langle \mathit{Liar}\rangle)$

E_{2_8}: $\mathrm{ReasonByCases}\big(\neg \mathit{True}(\langle \mathit{Liar}\rangle) \vee \mathit{Liar}, \mathrm{E}_{2\_4}, \mathrm{E}_{2\_6}, \mathrm{E}_{2\_7}\big)$

Using the definition of the Liar sentence, this disjunction can be simplified.

E_{2_9}: $\mathsf{Assume}_{\vee}\big(\neg \mathit{True}(\langle \mathit{Liar}\rangle)\big)$

E_{2_10}: $\mathrm{E}_{2\_9}, \mathrm{E}_{1\_5} \rightsquigarrow \mathit{Liar}$

E_{2_11}: $\mathsf{Assume}_{\vee}(\mathit{Liar})$

E_{2_12}: $\mathrm{ReasonByCases}(\mathit{Liar}, \mathrm{E}_{2\_10}, \mathrm{E}_{2\_11}, \mathrm{E}_{2\_8})$

Therefore, $\mathit{True}\big(\langle \mathit{Liar}\rangle \to \mathit{Liar}\big)$ entails Liar.

E_{2_13}: $\to\text{-}\mathsf{intro}\big([\mathit{True}(\langle \mathit{Liar}\rangle) \to \mathit{Liar}] \to \mathit{Liar}, \mathrm{E}_{2\_12}\big)$

The proof that Liar entails $\mathit{True}\big(\langle \mathit{Liar}\rangle\big) \to \mathit{Liar}$ is very straightforward.

E_{3_1}: $\mathsf{Assume}_{\to}(\mathit{Liar})$

E_{3_2}: $\mathsf{Assume}_{\to}\big(\mathit{True}(\langle \mathit{Liar}\rangle)\big)$

E_{3_3}: $\to\text{-}\mathsf{intro}\big(\mathit{True}(\langle \mathit{Liar}\rangle) \to \mathit{Liar}, \mathrm{E}_{3\_2}\big)$

E_{3_4}: $\to\text{-}\mathsf{intro}\big(\mathit{Liar} \to [\mathit{True}(\langle \mathit{Liar}\rangle) \to \mathit{Liar}], \mathrm{E}_{3\_3}\big)$

Thus, there is an instance of the axiom T-OUT that is regarded as not true.

E_{4_1}: $\mathrm{E}_{2\_13}, \mathrm{E}_{3\_4}, \mathrm{E}_{1\_9} \Rightarrow \neg \mathit{True}\big(\big\langle \mathit{True}(\langle \mathit{Liar}\rangle) \to \mathit{Liar}\big\rangle\big)$

E_{4_2}: $\rightsquigarrow \mathit{ax}_{T\_OUT}(\langle \mathit{Liar}\rangle) = \big\langle \mathit{True}(\langle \mathit{Liar}\rangle) \to \mathit{Liar}\big\rangle$

E_{4_3}: $\mathrm{E}_{4\_1}, \mathrm{E}_{4\_2} \rightsquigarrow \neg \mathit{True}\big(\mathit{ax}_{T\_OUT}(\langle \mathit{Liar}\rangle)\big)$

> [T]he theory implies the the untruth of certain of its axioms, so it holds that the Consistency Argument goes wrong in Step (1). For instance, the sentence
>
> If $True(\langle Q\rangle)$, then Q
>
> is an instance of (T-OUT), hence an axiom; but the theory implies
>
> (*) $\neg True(\langle True(\langle Q\rangle)\rangle)$.
>
> (The theory takes
>
> (**) If $True(\langle Q\rangle)$ then Q
>
> to be equivalent to
>
> (**) Either not $True(\langle Q\rangle)$, or Q.
>
> This in turn is equivalent to Q, since by the Liar property, the untruth of Q is equivalent to Q. Since the logic takes (**) to be equivalent to Q, and it takes Q not to be true, it is not surprising that it takes (**) not to be true, i.e. that it accepts (*).) ([Fie06, p. 572f])

This is a counterexample to the assumption that all axioms are true.

E_{4_4}:			$\mathsf{Assume}_{\neg}\Big(\forall x.\forall t.\big[[\mathit{Axiom}\big(ax_{T\_OUT}(\langle \mathit{Liar}\rangle),t\big) \wedge \mathit{AdoptTheo}(t)] \rightarrow \mathit{True}\big(ax_{T\_OUT}(\langle \mathit{Liar}\rangle)\big)\big]\Big)$
E_{4_5}:	E_{4_4}	$\rightsquigarrow$	$\forall t.\big[[\mathit{Axiom}\big(ax_{T\_OUT}(\langle \mathit{Liar}\rangle),t\big) \wedge \mathit{AdoptTheo}(t)] \rightarrow \mathit{True}\big(ax_{T\_OUT}(\langle \mathit{Liar}\rangle)\big)\big]$
E_{4_6}:	E_{4_5}	$\rightsquigarrow$	$[\mathit{Axiom}\big(ax_{T\_OUT}(\langle \mathit{Liar}\rangle),KF\big) \wedge \mathit{AdoptTheo}(KF)] \rightarrow \mathit{True}\big(ax_{T\_OUT}(\langle \mathit{Liar}\rangle)\big)$
E_{4_7}:		$\rightsquigarrow$	$\mathit{Axiom}\big(ax_{T\_OUT}(\langle \mathit{Liar}\rangle),KF\big)$
E_{4_8}:		$\Rightarrow$	$\mathit{AdoptTheo}(KF)$
E_{4_9}:	$\mathrm{E}_{4\_7}, \mathrm{E}_{4\_8}$	$\rightsquigarrow$	$\mathit{Axiom}\big(ax_{T\_OUT}(\langle \mathit{Liar}\rangle),KF\big) \wedge \mathit{AdoptTheo}(KF)$
E_{4_10}:	$\mathrm{E}_{4\_9}, \mathrm{E}_{4\_6}$	$\rightsquigarrow$	$\mathit{True}\big(ax_{T\_OUT}(\langle \mathit{Liar}\rangle)\big)$
E_{4_11}:	$\mathrm{E}_{4\_10}, \mathrm{E}_{4\_3}$	$\rightsquigarrow$	$\mathit{True}\big(ax_{T\_OUT}(\langle \mathit{Liar}\rangle)\big) \wedge \neg\,\mathit{True}\big(ax_{T\_OUT}(\langle \mathit{Liar}\rangle)\big)$
E_{4_12}:	E_{4_11}	$\rightsquigarrow$	$\bot$
E_{4_13}:			$\mathrm{ProofByContrad}\Big(\neg\Big[\forall x.\forall t.\big[[\mathit{Axiom}\big(ax_{T\_OUT}(\langle \mathit{Liar}\rangle),t\big) \wedge \mathit{AdoptTheo}(t)] \rightarrow \mathit{True}\big(ax_{T\_OUT}(\langle \mathit{Liar}\rangle)\big)\big]\Big], \mathrm{E}_{4\_12}\Big)$

As noted above, the instance of the T-OUT axiom scheme for the Liar sentence contradicts the induction start (step 1) of the Consistency argument. In particular, E_{4_13} rebuts the assumption that all axioms are true, represented e.g. in C_{1_2}.

Chapter 5: Classical Dialetheisms

In chapter 5, Field discusses the opposite approach to the one analyzed in chapter 4: he proposes that there could be theories that restrict the T-OUT rule scheme, but preserve the T-IN scheme. As mentioned in 2.2.3, Field uses the name "classical dialetheism" for such logics, since for the Liar's sentence Liar both $True(\langle \mathit{Liar}\rangle)$ and $True(\langle\neg \mathit{Liar}\rangle))$ must hold.[15] Apart from the rejection of T-OUT, these theories accept all the standard building blocks of classical logic. Similar to the case with KF, we will present here the strictest form of these theories where all instances of T-OUT are rejected.

$$\begin{array}{rrcl} & classDia(\widetilde{t}) & \rightsquigarrow & Classical(\widetilde{t}) \\ & classDia(\widetilde{t}) & \rightsquigarrow & Axiom\big(ax_{T\_IN}(\langle\widehat{p}\rangle),\widetilde{t}\big) \\ r_{\neg T\_OUT(\langle\widehat{p}\rangle)}: & classDia(\widetilde{t}) & \rightsquigarrow & \neg Axiom\big(ax_{T\_OUT}(\langle\widehat{p}\rangle),\widetilde{t}\big) \end{array}$$

In his argumentation, Field presents an example for the "dialetheic" nature of this logic. Using the rules of classical logic, Field shows how a logic that accepts T-IN must regard the Liar sentence and its negation as true. Using reasoning by cases, he argues

[15] This is in fact not the standard notion of dialetheism since $True(\langle \mathit{Liar}\rangle)$ and $True(\langle\neg \mathit{Liar}\rangle)$ are not actual negations of one another, so one could argue that this approach does not harm actual consistency.

that *Liar* must be true (using either T-IN) or the definition of the Liar sentence). In ASPIC-END, this argument can be formulated very straightforwardly.[16]

Using reasoning by cases, the theory regards *Liar* as true.			
F_{1_1}:		$\rightsquigarrow$	$\mathit{Liar} \lor \neg \mathit{Liar}$
If the theory accepts *Liar*, then it is true by T-IN.			
F_{1_2}:			$\mathsf{Assume}_{\lor}(\mathit{Liar})$
F_{1_3}:		$\rightsquigarrow$	$\mathit{Liar} \rightarrow \mathit{True}(\langle \mathit{Liar} \rangle)$
F_{1_4}:	$\mathrm{F}_{1\_2}, \mathrm{F}_{1\_3}$	$\rightsquigarrow$	$\mathit{True}(\langle \mathit{Liar} \rangle)$
If not, then *Liar* is true by its own definition.			
F_{1_5}:			$\mathsf{Assume}_{\lor}(\neg \mathit{Liar})$
F_{1_6}:			$\mathsf{Assume}_{\neg}\big(\neg \mathit{True}(\langle \mathit{Liar} \rangle)\big)$
F_{1_7}:		$\rightsquigarrow$	$\mathit{DefinedAs}(\langle \mathit{Liar} \rangle, \langle \neg \mathit{True}(\langle \mathit{Liar} \rangle) \rangle)$
F_{1_8}:	$\mathrm{F}_{1\_7}, \mathrm{F}_{1\_6}$	$\rightsquigarrow$	Liar
F_{1_9}:	$\mathrm{F}_{1\_8}, \mathrm{F}_{1\_5}$	$\rightsquigarrow$	$\mathit{Liar} \land \neg \mathit{Liar}$
F_{1_10}:	F_{1_9}	$\rightsquigarrow$	$\bot$
F_{1_11}:			$\mathsf{ProofByContrad}\big(\neg\neg \mathit{True}(\langle \mathit{Liar} \rangle), \mathrm{F}_{1\_10}\big)$
F_{1_12}:	F_{1_11}	$\rightsquigarrow$	$\mathit{True}(\langle \mathit{Liar} \rangle)$
F_{1_13}:			$\mathsf{ReasonByCases}\big(\mathit{True}(\langle \mathit{Liar} \rangle), \mathrm{F}_{1\_4}, \mathrm{F}_{1\_12}, \mathrm{F}_{1\_1}\big)$
By the definition of the Liar sentence, the theory cannot accept *Liar*.			
F_{1_14}:			$\mathsf{Assume}_{\neg}(\mathit{Liar})$
F_{1_15}:	$\mathrm{F}_{1\_7}, \mathrm{F}_{1\_14}$	$\rightsquigarrow$	$\neg \mathit{True}(\langle \mathit{Liar} \rangle)$
F_{1_16}:	$\mathrm{F}_{1\_13}, \mathrm{F}_{1\_15}$	$\rightsquigarrow$	$\mathit{True}(\langle \mathit{Liar} \rangle) \land \neg \mathit{True}(\langle \mathit{Liar} \rangle)$
F_{1_17}:	F_{1_16}	$\rightsquigarrow$	$\bot$
F_{1_18}:			$\mathsf{ProofByContrad}(\neg \mathit{Liar}, \mathrm{F}_{1\_17})$
By T-IN, the theory also regards $\neg$*Liar* as true.			
F_{1_19}:		$\rightsquigarrow$	$\neg \mathit{Liar} \rightarrow \mathit{True}(\langle \neg \mathit{Liar} \rangle)$
F_{1_20}:	$\mathrm{F}_{1\_18}, \mathrm{F}_{1\_19}$	$\rightsquigarrow$	$\mathit{True}(\langle \neg \mathit{Liar} \rangle)$

Notably, F_{1_3} and F_{1_19} are applications of the T-IN axiom scheme. Consequently, while this argument can be made in any classically dialetheic theory, it would be undercut in KF (and is not accepted in the corresponding extension).

Field then proposes two logical theories that he ascribes to this approach: *hyper-*

[16] Field sketches the steps needed to derive these formulae in the following way: "We have that if Q then $\mathit{True}(\langle Q \rangle)$, by (T-IN), and that if $\neg Q$ then $\mathit{True}(\langle Q \rangle$, by the meaning of Q, so $\mathit{True}(\langle Q \rangle)$ either way. But $\mathit{True}(\langle Q \rangle)$ is equivalent to $\neg Q$, so we have $\neg Q$; and by (T-IN) we get $\mathit{True}(\langle Q \rangle)$." ([Fie06, p. 575])

(As a reminder: Field chose to represent the Liar sentence with the symbol Q.)

dialetheism and a more standard classical "*dialetheic*" theory.

Hyperdialetheic Theory
Field describes as hyper-dialetheism the logic where all sentences are true. While he admits that this is as such a very uninteresting approach,[17] it is still formally consistent and thus worth considering.

$$\begin{array}{rcl} & \Rightarrow & \mathit{AdoptTheo}(\mathit{hyperDia}) \\ \mathit{AdoptTheo}(\mathit{hyperDia}) & \rightsquigarrow & \forall x.\mathit{True}(x) \end{array}$$

Notably, for any formula φ the incongruent views $\mathit{True}(\langle\varphi\rangle)$ and $\mathit{True}(\langle\neg\varphi\rangle)$ are held (i.e. all sentences are dialetheia). Notably, these dialetheia by themselves do not entail any inconsistencies, but they nonetheless diminish the usefulness of the truth predicate. Furthermore, these dialetheia provide counter-examples against both versions of step (ii) of the Consistency argument: It is obvious that the first assumption, Field's "elementary property" of truth $(\forall x.\neg[\mathit{True}(x) \wedge \mathit{True}(\mathit{neg}(x))])$, cannot be accepted when every sentence and its negation are true. At the same time, it is equally obvious that the alternative assumption $(\exists x.\neg \mathit{True}(x))$ cannot be accepted in hyperdialetheism either, since no sentence is not true.[18]

`We can consider adopting hyper-dialetheism.`

$$\begin{array}{lrcl} \mathrm{F}_{2\_1}\text{:} & & \Rightarrow & \mathit{AdoptTheo}(\mathit{hyperDia}) \end{array}$$

`In hyper-dialetheism, any sentence and its negation are true.`

$$\begin{array}{lrcl} \mathrm{F}_{2\_2}\text{:} & \mathrm{F}_{2\_1} & \rightsquigarrow & \forall x.\mathit{True}(x) \\ \mathrm{F}_{2\_3}\text{:} & \mathrm{F}_{2\_2} & \rightsquigarrow & \mathit{True}(x) \\ \mathrm{F}_{2\_4}\text{:} & \mathrm{F}_{2\_2} & \rightsquigarrow & \mathit{True}\big(\mathit{neg}(x)\big) \end{array}$$

`Then, Field's ''elementary property of truth'' is not satisfied.`

$$\begin{array}{lrcl} \mathrm{F}_{2\_5}\text{:} & & & \mathsf{Assume}_{\neg}\Big(\forall x.\neg\big[\mathit{True}(x) \wedge \mathit{True}\big(\mathit{neg}(x)\big)\big]\Big) \\ \mathrm{F}_{2\_6}\text{:} & \mathrm{F}_{2\_5} & \rightsquigarrow & \neg[\mathit{True}(x) \wedge \mathit{True}\big(\mathit{neg}(x)\big)] \\ \mathrm{F}_{2\_7}\text{:} & \mathrm{F}_{2\_3}, \mathrm{F}_{2\_4} & \rightsquigarrow & \mathit{True}(x) \wedge \mathit{True}\big(\mathit{neg}(x)\big) \\ \mathrm{F}_{2\_8}\text{:} & \mathrm{F}_{2\_7}, \mathrm{F}_{2\_6} & \rightsquigarrow & [\mathit{True}(x) \wedge \mathit{True}\big(\mathit{neg}(x)\big)] \wedge \neg[\mathit{True}(x) \wedge \mathit{True}\big(\mathit{neg}(x)\big)] \\ \mathrm{F}_{2\_9}\text{:} & \mathrm{F}_{2\_8} & \rightsquigarrow & \bot \\ \mathrm{F}_{2\_10}\text{:} & & & \mathsf{ProofByContrad}\Big(\neg\forall x.\neg\big[\mathit{True}(x) \wedge \mathit{True}\big(\mathit{neg}(x)\big)\big], \mathrm{F}_{2\_9}\Big) \end{array}$$

`Also, no sentence exists that is not true.`

$$\begin{array}{lrcl} \mathrm{F}_{2\_11}\text{:} & & & \mathsf{Assume}_{\neg}\big(\exists x.\neg \mathit{True}(x)\big) \end{array}$$

[17] Field notes that truth loses its meaning in the hyperdialetheic logic, since no sentence is not true in hyperdialetheism.

[18] Field chooses to describe the problem of hyperdialetheism in the following way: "The problem, rather, is that hyperdialetheism blocks the inference from the truth of the theory to its consistency: it takes even inconsistent sentences to be true." ([Fie06, p. 575])

$$\begin{array}{lrcl}
\mathrm{F}_{2\_12}: & \mathrm{F}_{2\_11} & \rightsquigarrow & \neg True\Big(g\big(\langle\neg True(x)\rangle\big)\Big) \\
\mathrm{F}_{2\_13}: & \mathrm{F}_{2\_2} & \rightsquigarrow & True\Big(g\big(\langle\neg True(x)\rangle\big)\Big) \\
\mathrm{F}_{2\_14}: & \mathrm{F}_{2\_13}, \mathrm{F}_{2\_12} & \rightsquigarrow & True\Big(g\big(\langle\neg True(x)\rangle\big)\Big) \wedge \neg True\Big(g\big(\langle\neg True(x)\rangle\big)\Big) \\
\mathrm{F}_{2\_15}: & \mathrm{F}_{2\_14} & \rightsquigarrow & \bot \\
\mathrm{F}_{2\_16}: & & & \mathsf{ProofByContrad}\big(\neg\exists x.\neg True(x), \mathrm{F}_{2\_15}\big)
\end{array}$$

It is easy to see that these arguments are rebuttals to the assumptions Field bases the arguments for step (ii) of the Consistency argument on. F_{2_10} asserts that in hyper-dialetheism not all sentences are not dialetheia (i.e. there exist sentences where both it and the negation are true). This counters and rebuts C_{3_4}, while F_{2_16} on the other hand rebuts C_{4_1}, the assumption for the variant argument. In both cases, Field proposed common (but not unassailable) assumptions that hyper-dialetheism violates. The rebuttals are successful since we presented Field's assumptions as defeasible.[19]

Classical Non-Hyperdialetheic Theories

Besides the rather "flat" hyperdialetheism, Field argues that there might be classical theories where some, but not all sentences are dialetheia. These present a more interesting case and might be more deserving of exploration compared to hyperdialetheism. We will call this type of logic classical non-hyper-dialetheism (shortened to classical nh-dialetheism).

Classical nh-dialetheism can be characterized in ASPIC-END using the following rules (For classical nh-dialetheism, we will assume that there is a sentence that is not true, represented by the constant symbol $nTrueS$):

$$\begin{array}{rcl}
 & \Rightarrow & AdoptTheo(cnhDia) \\
 & \rightsquigarrow & classDia(cnhDia) \\
AdoptTheo(cnhDia) & \rightsquigarrow & \neg True(nTrueS)
\end{array}$$

However, in classical nh-dialetheic logic, Field notes that for any sentence that is not true, an instance of modus ponens can be found that is not truth-preserving. In particular, Field introduces different instances of modus ponens that are not truth-preserving, depending on whether the conditional $\neg Liar \rightarrow nTrueS$ is true or not (essentially reasoning by cases).

On one hand, he shows that if the theory regarded the conditional as true, then for the rule $\neg Liar \wedge [\neg Liar \rightarrow nTrueS] \vdash nTrueS$ (this is an instance of modus ponens), both antecedents are true,[20] while the conclusion cannot be accepted as true.

On the other hand, if it regarded the sentence as not true, then the rule $Liar \wedge \big[Liar \rightarrow [\neg Liar \rightarrow nTrueS]\big] \vdash \neg Liar \rightarrow nTrueS$ would not be truth-preserving. Its

[19] Recall that a rebuttal can only be done on a defeasible rule application.

[20] As shown above, $\neg Liar$ is accepted in any classical dialetheism.

antecedents are unconditionally true in classical nh-dialetheism,[21] but the conclusion is assumed not true.[22]

We can characterize these instances of modus ponens in the following way:

$$\begin{array}{rcl}
Ante\big(a, r_{MP}\langle(\langle\neg Liar\rangle, \langle nTrueS\rangle)\big) & \rightsquigarrow & a = \langle\neg Liar\rangle \vee \\
& & a = \langle\neg Liar \rightarrow nTrueS\rangle \\
& \rightsquigarrow & conc\big(r_{MP}(\langle\neg Liar\rangle, \langle nTrueS\rangle)\big) = \\
& & \langle nTrueS\rangle \\
Ante\big(a, r_{MP}(\langle Liar\rangle, \langle\neg Liar \rightarrow nTrueS\rangle)\big) & \rightsquigarrow & a = \langle Liar\rangle \vee \\
& & a = \langle Liar \rightarrow [\neg Liar \rightarrow nTrueS]\rangle \\
& \rightsquigarrow & conc\big(r_{MP}(\langle Liar\rangle, \langle\neg Liar \rightarrow nTrueS\rangle)\big) = \\
& & \langle\neg Liar \rightarrow nTrueS\rangle
\end{array}$$

Then, we can replicate Field's argument.

`We can consider adopting classical nh-dialetheism.`

$\mathrm{F}_{3,1}$:		$\Rightarrow$	$AdoptTheo(cnhDia)$
$\mathrm{F}_{3,2}$:		$\rightsquigarrow$	$ClassDia(cnhDia)$
$\mathrm{F}_{3,3}$:	$\mathrm{F}_{3,2}$	$\rightsquigarrow$	$Classical(cnhDia)$
$\mathrm{F}_{3,4}$:	$\mathrm{F}_{3,1}$	$\rightsquigarrow$	$\neg True(\langle nTrueS\rangle)$

[21] At the beginning of this chapter, we have shown that $\neg Liar$ must be true in any classical dialetheism, while the double implication is a tautology of classical logic.

[22] Field presents this proof in his paper at greater length. For the sake of reference, we will present the essential pieces of the proof here. As a note, Field chooses to represent the Liar sentence with the symbol Q and the non-true sentence with the symbol $\bot$.

> [A] theory of this kind must entail that modus ponens is not truth-preserving. [...]
> Why is this? We have seen the on such a view [i.e. classical dialetheism], $True(\langle Q\rangle)$, $True(\langle\neg Q\rangle)$ and $\neg True(\bot)$. What about the conditional $Q \rightarrow \bot$? [...] it may seem natural to suppose it should assume $True(\langle Q \rightarrow \bot\rangle)$. But if so, it takes the following instance of modus ponens to have true premisses and a false conclusion:
>
> $\neg Q$
> $\neg Q \rightarrow \bot$
> $\therefore \bot$
>
> So we have an instance of modus ponens that is not truth-preserving
> Perhaps it would be better [...] to declare $\neg True(\langle Q \rightarrow \bot\rangle)$ [...] But this will not help, for now consider the following instance of modus ponens:
>
> Q
> $Q \rightarrow (\neg Q \rightarrow \bot)$
> $\therefore \neg Q \rightarrow \bot$
>
> We have seen that the view takes the first premiss to be true, and we are now supposing that it takes the conclusion not to be true. But it also takes the second premiss to be true, given that that is a truth of classical logic and (T-IN) holds. So here too we have an instance of modus ponens that is not truth-preserving. ([Fie06, p. 576])

By case distinction, we can show that not all rules are truth-preserving.

$\mathrm{F}_{3,5}$: $\rightsquigarrow$ $\mathit{True}(\langle\neg \mathit{Liar} \rightarrow \mathit{nTrueS}\rangle) \vee \neg \mathit{True}(\langle\neg \mathit{Liar} \rightarrow \mathit{nTrueS}\rangle)$

First, consider the case where the conditional is true.

$\mathrm{F}_{4,1}$: $\mathsf{Assume}_{\vee}\big(\mathit{True}(\langle\neg \mathit{Liar} \rightarrow \mathit{nTrueS}\rangle)\big)$

Consider this instance of modus ponens: $\neg \mathit{Liar}, [\neg \mathit{Liar} \rightarrow \mathit{nTrueS}] \vdash \mathit{nTrueS}$.

$\mathrm{F}_{4,2}$: $\mathrm{F}_{3,3}$ $\Rightarrow$ $\mathit{Rule}\big(r_{MP}(\langle\neg \mathit{Liar}\rangle, \langle \mathit{nTrueS}\rangle), \mathit{cnhDia}\big)$

$\mathrm{F}_{4,3}$: $\mathsf{Assume}_{\rightarrow}\Big(\mathit{Ante}\big(r_{MP}(\langle\neg \mathit{Liar}\rangle, \langle \mathit{nTrueS}\rangle), a\big)\Big)$

$\mathrm{F}_{4,4}$: $\mathrm{F}_{4,3}$ $\rightsquigarrow$ $a = \langle\neg \mathit{Liar}\rangle \vee a = \langle\neg \mathit{Liar} \rightarrow \mathit{nTrueS}\rangle$

Both antecedents of this instance are true.

$\mathrm{F}_{4,5}$: $\mathsf{Assume}_{\vee}(a = \langle\neg \mathit{Liar}\rangle)$

$\mathrm{F}_{4,6}$: $\mathrm{F}_{4,5}, \mathrm{F}_{1,18}$ $\rightsquigarrow$ $\mathit{True}(a)$

$\mathrm{F}_{4,7}$: $\mathsf{Assume}_{\vee}(a = \langle\neg \mathit{Liar} \rightarrow \mathit{nTrueS}\rangle)$

$\mathrm{F}_{4,8}$: $\mathrm{F}_{4,7}, \mathrm{F}_{4,1}$ $\rightsquigarrow$ $\mathit{True}(a)$

$\mathrm{F}_{4,9}$: $\mathsf{ReasonByCases}\big(\mathit{True}(a), \mathrm{F}_{4,6}, \mathrm{F}_{4,8}, \mathrm{F}_{4,4}\big)$

$\mathrm{F}_{4,10}$: $\rightarrow\text{-}\mathsf{intro}\Big(\mathit{Ante}\big(r_{MP}(\langle\neg \mathit{Liar}\rangle, \langle \mathit{nTrueS}\rangle), a\big) \rightarrow \mathit{True}(a), \mathrm{F}_{4,9}\Big)$

$\mathrm{F}_{4,11}$: $\forall\text{-}\mathsf{intro}\Big(\forall a.[\mathit{Ante}\big(r_{MP}(\langle\neg \mathit{Liar}\rangle, \langle \mathit{nTrueS}\rangle), a\big) \rightarrow$
$\mathit{True}(a)], \mathrm{F}_{4,10}\Big)$

Then, not all rule instances are truth-preserving.

$\mathrm{F}_{4,12}$: $\rightsquigarrow$ $\mathit{conc}\big(r_{MP}(\langle\neg \mathit{Liar}\rangle, \langle \mathit{nTrueS}\rangle)\big) = \langle \mathit{nTrueS}\rangle$

$\mathrm{F}_{4,13}$: $\mathrm{F}_{4,12}, \mathrm{F}_{3,4}$ $\rightsquigarrow$ $\neg \mathit{True}\Big(\mathit{conc}\big(r_{MP}(\langle\neg \mathit{Liar}\rangle, \langle \mathit{nTrueS}\rangle)\big)\Big)$

$\mathrm{F}_{4,14}$: $\mathsf{Assume}_{\neg}\Big(\mathit{TPres}\big(r_{MP}(\langle\neg \mathit{Liar}\rangle, \langle \mathit{nTrueS}\rangle)\big)\Big)$

$\mathrm{F}_{4,15}$: $\rightsquigarrow$ $\mathit{DefinedAs}\Big(\langle \mathit{TPres}(r_{MP}(\langle\neg \mathit{Liar}\rangle, \langle \mathit{nTrueS}\rangle))\rangle,$
$\big\langle\big[\forall a.[\mathit{Ante}(r_{MP}(\langle\neg \mathit{Liar}\rangle, \langle \mathit{nTrueS}\rangle), a) \rightarrow \mathit{True}(a)]\big] \rightarrow$
$\mathit{True}(\mathit{conc}(r_{MP}(\langle\neg \mathit{Liar}\rangle, \langle \mathit{nTrueS}\rangle)))\big\rangle\Big)$

$\mathrm{F}_{4,16}$: $\mathrm{F}_{4,14}, \mathrm{F}_{4,15}$ $\rightsquigarrow$ $\big[\forall a.[\mathit{Ante}\big(r_{MP}(\langle\neg \mathit{Liar}\rangle, \langle \mathit{nTrueS}\rangle), a\big) \rightarrow \mathit{True}(a)]\big] \rightarrow$
$\mathit{True}\Big(\mathit{conc}\big(r_{MP}(\langle\neg \mathit{Liar}\rangle, \langle \mathit{nTrueS}\rangle)\big)\Big)$

$\mathrm{F}_{4,17}$: $\mathrm{F}_{4,11}, \mathrm{F}_{4,16}$ $\rightsquigarrow$ $\mathit{True}\Big(\mathit{conc}\big(r_{MP}(\langle\neg \mathit{Liar}\rangle, \langle \mathit{nTrueS}\rangle)\big)\Big)$

$\mathrm{F}_{4,18}$: $\mathrm{F}_{4,17}, \mathrm{F}_{4,13}$ $\rightsquigarrow$ $\mathit{True}\Big(\mathit{conc}\big(r_{MP}(\langle\neg \mathit{Liar}\rangle, \langle \mathit{nTrueS}\rangle)\big)\Big) \wedge$
$\neg \mathit{True}\Big(\mathit{conc}\big(r_{MP}(\langle\neg \mathit{Liar}\rangle, \langle \mathit{nTrueS}\rangle)\big)\Big)$

$\mathrm{F}_{4,19}$: $\mathrm{F}_{4,18}$ $\rightsquigarrow$ $\bot$

$\mathrm{F}_{4,20}$: $\mathsf{ProofByContrad}\Big(\neg \mathit{TPres}\big(r_{MP}(\langle\neg \mathit{Liar}\rangle, \langle \mathit{nTrueS}\rangle)\big), \mathrm{F}_{4,19}\Big)$

$\mathrm{F}_{4,21}$: $\mathrm{F}_{4,2}, \mathrm{F}_{4,20}$ $\rightsquigarrow$ $\mathit{Rule}\big(r_{MP}(\langle\neg \mathit{Liar}\rangle, \langle \mathit{nTrueS}\rangle)\big) \wedge$
$\neg \mathit{TPres}\big(r_{MP}(\langle\neg \mathit{Liar}\rangle, \langle \mathit{nTrueS}\rangle)\big)$

$\mathrm{F}_{4,22}$: $\mathrm{F}_{4,21}$ $\rightsquigarrow$ $\exists r.[\mathit{Rule}(r, \mathit{cnhDia}) \land \neg \mathit{TPres}(r)]$

Next, consider the case where the conditional is not true.

$\mathrm{F}_{5,1}$: $\mathsf{Assume}_{\vee}\big(\neg \mathit{True}(\langle \neg \mathit{Liar} \to \mathit{nTrueS}\rangle)\big)$

Consider this instance of modus ponens:

$$\mathit{Liar}, \mathit{Liar} \to [\neg \mathit{Liar} \to \mathit{nTrueS}] \vdash [\neg \mathit{Liar} \to \mathit{nTrueS}].$$

$\mathrm{F}_{5,2}$: $\mathrm{F}_{3,3}$ $\rightsquigarrow$ $\mathit{Rule}\big(r_{MP}(\langle \mathit{Liar}\rangle, \langle \neg \mathit{Liar} \to \mathit{nTrueS}\rangle), \mathit{cnhDia}\big)$

$\mathrm{F}_{5,3}$: $\mathsf{Assume}_{\to}\Big(\mathit{Ante}\big(r_{MP}(\langle \mathit{Liar}\rangle, \langle \neg \mathit{Liar} \to \mathit{nTrueS}\rangle), a\big)\Big)$

$\mathrm{F}_{5,4}$: $\mathrm{F}_{5,3}$ $\rightsquigarrow$ $a = \langle \mathit{Liar}\rangle \lor a = \langle \mathit{Liar} \to [\neg \mathit{Liar} \to \mathit{nTrueS}]\rangle$

As shown earlier, *Liar* is true in any classically dialetheic theory.

$\mathrm{F}_{5,5}$: $\mathsf{Assume}_{\vee}(a = \langle \mathit{Liar}\rangle)$

$\mathrm{F}_{5,6}$: $\mathrm{F}_{5,5}, \mathrm{F}_{1,12}$ $\rightsquigarrow$ $\mathit{True}(a)$

$\mathit{Liar} \to [\neg \mathit{Liar} \to \mathit{nTrueS}]$ is a classical tautology.

$\mathrm{F}_{5,7}$: $\mathsf{Assume}_{\vee}(a = \langle \mathit{Liar} \to [\neg \mathit{Liar} \to \mathit{nTrueS}]\rangle)$

$\mathrm{F}_{5,8}$: $\mathsf{Assume}_{\to}(\mathit{Liar})$

$\mathrm{F}_{5,9}$: $\mathsf{Assume}_{\to}(\neg \mathit{Liar})$

$\mathrm{F}_{5,10}$: $\mathrm{F}_{5,8}, \mathrm{F}_{5,9}$ $\rightsquigarrow$ $\mathit{Liar} \land \neg \mathit{Liar}$

$\mathrm{F}_{5,11}$: $\mathrm{F}_{5,10}$ $\rightsquigarrow$ $\bot$

$\mathrm{F}_{5,12}$: $\mathrm{F}_{5,11}$ $\rightsquigarrow$ nTrueS

$\mathrm{F}_{5,13}$: $\to\text{-}\mathsf{intro}(\neg \mathit{Liar} \to \mathit{nTrueS}, \mathrm{F}_{5,12})$

$\mathrm{F}_{5,14}$: $\to\text{-}\mathsf{intro}(\mathit{Liar} \to [\neg \mathit{Liar} \to \mathit{nTrueS}], \mathrm{F}_{5,13})$

$\mathrm{F}_{5,15}$: $\rightsquigarrow$ $\big[\mathit{Liar} \to [\neg \mathit{Liar} \to \mathit{nTrueS}]\big] \to$ $\mathit{True}(\langle \mathit{Liar} \to [\neg \mathit{Liar} \to \mathit{nTrueS}]\rangle)$

$\mathrm{F}_{5,16}$: $\mathrm{F}_{5,14}, \mathrm{F}_{5,15}$ $\rightsquigarrow$ $\mathit{True}(\langle \mathit{Liar} \to [\neg \mathit{Liar} \to \mathit{nTrueS}]\rangle)$

$\mathrm{F}_{5,17}$: $\mathrm{F}_{5,16}, \mathrm{F}_{5,7}$ $\rightsquigarrow$ $\mathit{True}(a)$

Thus, both antecedents of the instance of modus ponens are true.

$\mathrm{F}_{5,18}$: $\mathsf{ReasonByCases}\big(\mathit{True}(a), \mathrm{F}_{5,6}, \mathrm{F}_{5,17}, \mathrm{F}_{5,4}\big)$

$\mathrm{F}_{5,19}$: $\to\text{-}\mathsf{intro}\Big(\mathit{Ante}\big(r_{MP}(\langle \mathit{Liar}\rangle, \langle \neg \mathit{Liar} \to \mathit{nTrueS}\rangle), a\big) \to$ $\mathit{True}(a), \mathrm{F}_{5,18}\Big)$

$\mathrm{F}_{5,20}$: $\forall\text{-}\mathsf{intro}\Big(\forall a.[\mathit{Ante}\big(r_{MP}(\langle \mathit{Liar}\rangle, \langle \neg \mathit{Liar} \to \mathit{nTrueS}\rangle), a\big) \to$ $\mathit{True}(a)], \mathrm{F}_{5,19}\Big)$

Then, not all rule instances are truth-preserving.

$\mathrm{F}_{5,21}$: $\rightsquigarrow$ $\mathit{conc}\big(r_{MP}(\langle \mathit{Liar}\rangle, \langle \neg \mathit{Liar} \to \mathit{nTrueS}\rangle)\big) = \langle \neg \mathit{Liar} \to$ $\mathit{nTrueS}\rangle$

$\mathrm{F}_{5,22}$: $\mathrm{F}_{5,21}, \mathrm{F}_{5,1}$ $\rightsquigarrow$ $\neg \mathit{True}\Big(\mathit{conc}\big(r_{MP}(\langle \mathit{Liar}\rangle, \langle \neg \mathit{Liar} \to \mathit{nTrueS}\rangle)\big)\Big)$

$\mathrm{F}_{5,23}$: $\mathsf{Assume}_{\neg}\Big(\mathit{TPres}\big(r_{MP}(\langle \mathit{Liar}\rangle, \langle \neg \mathit{Liar} \to \mathit{nTrueS}\rangle)\big)\Big)$

$F_{5,24}$:		$\leadsto$	$\mathit{DefinedAs}\Big(\langle \mathit{TPres}(r_{MP}(\langle \mathit{Liar}\rangle, \langle \neg \mathit{Liar} \to nTrueS\rangle)))\rangle, \big\langle \big[\forall a.[\mathit{Ante}(r_{MP}(\langle \mathit{Liar}\rangle, \langle \neg \mathit{Liar} \to nTrueS\rangle), a) \to \mathit{True}(a)]\big] \to \mathit{True}(\mathit{conc}(r_{MP}(\langle \mathit{Liar}\rangle, \langle \neg \mathit{Liar} \to nTrueS\rangle))))\big\rangle\Big)$
$F_{5,25}$:	$F_{5,23}, F_{5,24}$	$\leadsto$	$\big[\forall a.[\mathit{Ante}\big(r_{MP}(\langle \mathit{Liar}\rangle, \langle \neg \mathit{Liar} \to nTrueS\rangle), a\big) \to \mathit{True}(a)]\big] \to \mathit{True}\Big(\mathit{conc}\big(r_{MP}(\langle \mathit{Liar}\rangle, \langle \neg \mathit{Liar} \to nTrueS\rangle)\big)\Big)$
$F_{5,26}$:	$F_{5,20}, F_{5,25}$	$\leadsto$	$\mathit{True}\Big(\mathit{conc}\big(r_{MP}(\langle \mathit{Liar}\rangle, \langle \neg \mathit{Liar} \to nTrueS\rangle)\big)\Big)$
$F_{5,27}$:	$F_{5,26}, F_{5,22}$	$\leadsto$	$\mathit{True}\Big(\mathit{conc}\big(r_{MP}(\langle \mathit{Liar}\rangle, \langle \neg \mathit{Liar} \to nTrueS\rangle)\big)\Big) \wedge \neg \mathit{True}\Big(\mathit{conc}\big(r_{MP}(\langle \mathit{Liar}\rangle, \langle \neg \mathit{Liar} \to nTrueS\rangle)\big)\Big)$
$F_{5,28}$:	$F_{5,27}$	$\leadsto$	$\bot$
$F_{5,29}$:			$\mathsf{ProofByContrad}\Big(\neg TPres\big(r_{MP}(\langle \mathit{Liar}\rangle, \langle \neg \mathit{Liar} \to nTrueS\rangle)\big), F_{5,28}\Big)$
$F_{5,30}$:	$F_{5,2}, F_{5,29}$	$\leadsto$	$\mathit{Rule}(r_{MP}\big(\langle \mathit{Liar}\rangle, \langle \neg \mathit{Liar} \to nTrueS\rangle)\big) \wedge \neg TPres\big(r_{MP}(\langle \mathit{Liar}\rangle, \langle \neg \mathit{Liar} \to nTrueS\rangle)\big)$
$F_{5,31}$:	$F_{5,30}$	$\leadsto$	$\exists r.[\mathit{Rule}(r, \mathit{cnhDia}) \wedge \neg TPres(r)]$

`In either case, there is a rule that is not truth-preserving.`

$F_{6,1}$:			$\mathsf{ReasonByCases}\big(\exists r.[\mathit{Rule}(r, \mathit{cnhDia}) \wedge \neg TPres(r)], F_{4,22}, F_{5,31}, F_{3,5}\big)$
$F_{6,2}$:	$F_{6,1}$	$\leadsto$	$\mathit{Rule}\Big(g\big(\langle \mathit{Rule}(r, \mathit{cnhDia}) \wedge \neg TPres(r)\rangle\big), \mathit{cnhDia}\Big) \wedge \neg TPres\Big(g\big(\langle \mathit{Rule}(r, \mathit{cnhDia}) \wedge \neg TPres(r)\rangle\big)\Big)$

`This contradicts the assumption that all rules are truth-preserving.`

$F_{6,3}$:			$\mathsf{Assume}_{\neg}\big(\forall r.\forall t.\big[[\mathit{Rule}(r,t) \wedge \mathit{AdoptTheo}(t)] \to TPres(r)\big]\big)$
$F_{6,4}$:	$F_{6,3}$	$\leadsto$	$\forall t.\Big[\Big[\mathit{Rule}\Big(g\big(\langle \mathit{Rule}(r, \mathit{cnhDia}) \wedge \neg TPres(r)\rangle\big), t\Big) \wedge AdoptTheo(t)\Big] \to TPres\Big(g\big(\langle \mathit{Rule}(r, \mathit{cnhDia}) \wedge \neg TPres(r)\rangle\big)\Big)\Big]$
$F_{6,5}$:	$F_{6,4}$	$\leadsto$	$[\mathit{Rule}\Big(g\big(\langle \mathit{Rule}(r, \mathit{cnhDia}) \wedge \neg TPres(r)\rangle\big), \mathit{cnhDia}\Big) \wedge AdoptTheo(\mathit{cnhDia})] \to TPres\Big(g\big(\langle \mathit{Rule}(r, \mathit{cnhDia}) \wedge \neg TPres(r)\rangle\big)\Big)$
$F_{6,6}$:	$F_{6,2}$	$\leadsto$	$\mathit{Rule}\Big(g\big(\langle \mathit{Rule}(r, \mathit{cnhDia}) \wedge \neg TPres(r)\rangle\big), \mathit{cnhDia}\Big)$
$F_{6,7}$:	$F_{6,6}, F_{3,1}$	$\leadsto$	$\mathit{Rule}\Big(g\big(\langle \mathit{Rule}(r, \mathit{cnhDia}) \wedge \neg TPres(r)\rangle\big), \mathit{cnhDia}\Big) \wedge AdoptTheo(\mathit{cnhDia})$
$F_{6,8}$:	$F_{6,7}, F_{6,5}$	$\leadsto$	$TPres\Big(g\big(\langle \mathit{Rule}(r, \mathit{cnhDia}) \wedge \neg TPres(r)\rangle\big)\Big)$
$F_{6,9}$:	$F_{6,2}$	$\leadsto$	$\neg TPres\Big(g\big(\langle \mathit{Rule}(r, \mathit{cnhDia}) \wedge \neg TPres(r)\rangle\big)\Big)$

$$\begin{array}{llcl}
\mathrm{F}_{6,10}: & \mathrm{F}_{6,8}, \mathrm{F}_{6,9} & \rightsquigarrow & TPres\Big(g\big(\langle Rule(r, cnhDia) \wedge \neg TPres(r)\rangle\big)\Big) \wedge \\
 & & & \neg TPres\Big(g\big(\langle Rule(r, cnhDia) \wedge \neg TPres(r)\rangle\big)\Big) \\
\mathrm{F}_{6,11}: & \mathrm{F}_{6,10} & \rightsquigarrow & \bot \\
\mathrm{F}_{6,12}: & & & \mathsf{ProofByContrad}\Big(\neg\forall r.\forall t.\big[[Rule(r,t) \wedge AdoptTheo(t)] \rightarrow \\
 & & & TPres(r)\big], \mathrm{F}_{6,11}\Big)
\end{array}$$

It is easy to see that this argument contradicts the general assumption for proof-theoretic systems that all rules of the adopted theory preserve truth. In particular, $\mathrm{F}_{6,12}$ is a successful rebuttal against $\mathrm{C}_{2,2}$ of the Consistency argument.

5.3.3 Non-Classical Logic

After discussing theories of classical logics in chapters 4 – 8, Field shifts his focus towards non-classical logics. Here, there are two groups of logics he presents and assesses separately: on the one hand paracomplete logics that allow for truth-value gaps and on the other hand dialetheic logics that allow for truth-value gluts (also called dialetheia). We have chosen to take a closer look at paracomplete logics in particular, which Field discusses in chapters 9 and 10 of his paper.

Paracomplete Theories

In chapter 9 and 10, Field introduces the paracomplete logical theories as an alternative to the classical theories under consideration so far. As explained in 2.2.3, paracomplete theories aim to solve the Liar paradox by rejecting the Law of the Excluded Middle. They preserve (most) other "building blocks" of the philosophy of logic (an explicit definition follows below).

Field defines his ideal properties for a paracomplete logic (see [Fie06, p. 586]) as follows: He wishes to keep *reasoning by cases*, *Intersubstitutivity*, the *T-scheme*, *modus ponens*, the *Principle of explosion*, *double negation elimination* as well as the deMorgan laws and conditional contraposition. This can be used to define Field's paracomplete theories in ASPIC-END straightforwardly. (Notably, the deMorgan laws and contraposition were entailed by the rules of classical logic. Consequently, paracomplete logic is the first theory in this thesis to use them explicitly. See 4.3.2 for their definitions.)

$$
\begin{array}{ll}
\Rightarrow & AdoptTheo(paracomp) \\
\rightsquigarrow & Axiom\big(ax_{T\_IN}(\langle\widehat{p}\rangle), paracomp\big) \\
\rightsquigarrow & Axiom\big(ax_{T\_OUT}(\langle\widehat{p}\rangle), paracomp\big) \\
\rightsquigarrow & Rule\big(r_{MP}(\langle\widehat{p}\rangle, \langle\widehat{q}\rangle), paracomp\big) \\
\rightsquigarrow & Rule\big(r_{PE}(\langle\widehat{p}\rangle), paracomp\big) \\
\rightsquigarrow & Rule\big(r_{DNE}(\langle\widehat{p}\rangle), paracomp\big) \\
\rightsquigarrow & Rule(dml, paracomp) \\
\rightsquigarrow & Rule(r_{cCP}(\langle\widehat{p}\rangle, \langle\widehat{q}\rangle), paracomp)
\end{array}
$$

As mentioned in 2.2.1, the principle of Intersubstitutivity can be used to express the full T-scheme in the presence of the (relatively uncontroversial) conditional $p \rightarrow p$. This conditional is a given in all classical theories where →-introduction can be used, but as explained below, this is not the case in paracomplete logic. Instead, paracomplete logic states this conditional explicitly.

$$AdoptTheo(paracomp) \quad \rightsquigarrow \quad \widehat{p} \rightarrow \widehat{p}$$

Then, instances of T-IN and T-OUT can easily be inferred in paracomplete logic.[23] Unfortunately, without →-introduction, this inference cannot be done in a general manner in ASPIC-END. Instead, we will provide an example argument for a "generic" formula P.

$$
\begin{array}{lrcl}
\mathrm{G}_{1\_1}: & & \rightsquigarrow & P \rightarrow P \\
\mathrm{G}_{1\_2}: & & \rightsquigarrow & Alike_T\big(\langle P\rangle, \langle True(\langle P\rangle)\rangle\big) \\
\mathrm{G}_{1\_3}: & \mathrm{G}_{1\_2} & \rightsquigarrow & Alike_T\big(\langle P \rightarrow P\rangle, \langle P \rightarrow True(\langle P\rangle)\rangle\big) \\
\mathrm{G}_{1\_4}: & \mathrm{G}_{1\_2} & \rightsquigarrow & Alike_T\big(\langle P \rightarrow P\rangle, \langle True(\langle P\rangle) \rightarrow P\rangle\big) \\
\mathrm{G}_{1\_5}: & \mathrm{G}_{1\_3}, \mathrm{G}_{1\_1} & \rightsquigarrow & P \rightarrow True(\langle P\rangle) \\
\mathrm{G}_{1\_6}: & \mathrm{G}_{1\_4}, \mathrm{G}_{1\_1} & \rightsquigarrow & True(\langle P\rangle) \rightarrow P
\end{array}
$$

Here, G_{1_5} presents an instance of T-IN and G_{1_6} an instance of T-OUT (for the placeholder P). For all other instances, similar arguments can be constructed.

In exchange for these positive properties, paracomplete theories cannot keep the Law of the Excluded Middle, →-Introduction or proof by contradiction for pathological sentences (we will assess Field's argument against →-Introduction in more detail below). For now, it should be noted that Field accepts these inferences in all cases without the *True* predicate.[24] To represent this properly in ASPIC-END, we will first introduce a

[23] In order to keep to Field's arguments, we have decided to still include them formally in the definition of paracomplete logic above.

[24] Since, as Field argues "[i]n mathematics, physics, etc. logic is 'effectively classical' since all 'true'-free

definition for such "pathological" cases: any formula that directly contains the $True$ predicate or relies on a pathological sentence for its own definition should be seen as pathological. (As explained in 4.4, we do not replicate the syntactical construction necessary to define the Liar sentence. Accordingly, even though in our modeling the names $Liar$ and $Curry$ may not seem pathological since they do not contain truth, they should still be included.)

$$\begin{aligned} &\rightsquigarrow\quad Patho(\langle True(\langle\widehat{p}\rangle)\rangle) \\ Patho(\langle\widehat{p}\rangle) \vee Patho(\langle\widehat{q}\rangle) \quad &\rightsquigarrow\quad Patho(\langle\widehat{p} \wedge \widehat{q}\rangle) \end{aligned}$$

The last rule should be done analogously for the connectives $\neg$, $\vee$, $\rightarrow$, $\forall\widetilde{x}.$ and $\exists\widetilde{x}.$ as well.

In order to include the semantic paradoxes as well as any other properties (such as truth-preservation) that rely on a truth-assertion, we will require that the predicate *DefinedAs* extends pathologicality.

$$\mathit{DefinedAs}(\langle\widehat{p}\rangle, \langle\widehat{q}\rangle), Patho(\langle\widehat{q}\rangle) \quad \rightsquigarrow \quad Patho(\langle\widehat{p}\rangle)$$

Then, we can restrict the problematic inferences only for pathological cases:

$$\begin{aligned} Patho(\langle\widehat{p}\rangle) \quad &\rightsquigarrow\quad \neg Axiom\big(ax_{LEM}(\langle\widehat{p}\rangle), paracomp\big) \\ AdoptTheo(paracomp), Patho(\langle\widehat{p}\rangle) \quad &\rightsquigarrow\quad \neg\mathsf{Assumable}_{\rightarrow}(\widehat{p}) \\ AdoptTheo(paracomp), Patho(\langle\widehat{p}\rangle) \quad &\rightsquigarrow\quad \neg\mathsf{Assumable}_{\neg}(\widehat{p}) \end{aligned}$$

In ASPIC-END, this restriction on the LEM allows undercut attacks on any pathological usage of the LEM (e.g. E_{1_7} or F_{1_1}). Consequently, paracomplete theories do not accept the standard proofs for different Liar sentence results of earlier theories.

Field motivates his rejection of →-introducation as follows: he notes that for the Curry paradox, an argument can be constructed that relies on Intersubstitutivity, modus ponens and the →-Introduction. Field argues that any logic that aims to keep the first two principles, such as paracomplete logic, cannot keep →-Introductions as well. He goes on to note how other laws of implication would have to be restricted as well, which as he shows can be achieved by restricting the Law of the Excluded Middle.

It should be noted here that while both the induction start and induction step of the Consistency argument rely on →-introduction in this thesis (and it could be argued that the induction step in Field's formalization does as well[25]), in both cases the assumptions are made on non-pathological sentences (and the truth-assertions are introduced at a later point). Accordingly, the technicalities of the induction are not suspect for paracomplete theories.

instance of excluded middle can be taken as (non-logical) axioms" ([Fie06, p. 587]).

[25] Field formulates simple induction on natural numbers as follow: "$A(0)\&\forall n(A(n) \rightarrow A(n+1)) \vdash \forall n A(n)$" ([Fie06, p. 587]).

Field also assesses where the Consistency argument breaks down. It can be easily seen that the induction start is not problematic: Using T-IN, one can always take the step from any axiom to the truth of that axiom. Furthermore, paracomplete logic does not accept both a sentence and its negation as true, so the standard argument for step (ii) can be made as well (C_{3_1} - C_{3_16}).

Field then goes on to the last possibility: there are rules of the paracomplete theory that do not preserve truth. In particular, he distinguishes between rules unrestrictedly preserving truth and rules preserving truth when restricted to *legitimately assertible* formulae (i.e. theorems). However, he notes that neither form of truth preservation can be proven in a paracomplete theory.[26]

Field proves that some rules fail to preserve truth in the following way: he notes that the conditional $A \rightarrow B$ corresponding to a rule $A \vdash B$ can be transformed into an assertion of truth-preservation (with the help of Intersubstitutivity) and vice versa. He then notes that some rules cannot accept the conditional, so they cannot accept truth-preservation either. (As a note: because paracomplete logic rejects →-Introduction, the conditional is not automatically entailed by the existence of the rule.)

As a first step for sketching this proof, we will show how Field derives truth-preservation from the corresponding conditional for a rule. For the sake of simplicity, we will sketch only the direction from truth-preservation to the conditional in ASPIC-END, but the other direction can be performed in exactly the same manner. Notably, paracomplete logics reject →-Introduction, so we can only treat this argument as a generic snippet of a larger argument where the conditional for truth-preservation $True(\langle A\rangle) \rightarrow True(\langle B\rangle)$ is already shown (with A and B standing for arbitrary formulae):

$$
\begin{array}{llcl}
\mathrm{H}_{1\_0}: & & & True\big(\langle A\rangle\big) \rightarrow True\big(\langle B\rangle\big) \\
\mathrm{H}_{1\_1}: & & \rightsquigarrow & Alike_T\Big(\langle A\rangle, \big\langle True\big(\langle A\rangle\big)\big\rangle\Big) \\
\mathrm{H}_{1\_2}: & \mathrm{H}_{1\_1} & \rightsquigarrow & Alike_T\Big(\langle A \rightarrow B\rangle, \big\langle True\big(\langle A\rangle\big) \rightarrow B\big\rangle\Big) \\
\mathrm{H}_{1\_3}: & & \rightsquigarrow & Alike_T\big(\langle B\rangle, \big\langle True(\langle B\rangle)\big\rangle\big) \\
\mathrm{H}_{1\_4}: & \mathrm{H}_{1\_3} & \rightsquigarrow & Alike_T\Big(\big\langle True(\langle A\rangle) \rightarrow B\big\rangle, \big\langle True(\langle A\rangle) \rightarrow True(\langle B\rangle)\big\rangle\Big) \\
\mathrm{H}_{1\_5}: & \mathrm{H}_{1\_4}, \mathrm{H}_{1\_0} & \rightsquigarrow & True(\langle A\rangle) \rightarrow B \\
\mathrm{H}_{1\_6}: & \mathrm{H}_{1\_2}, \mathrm{H}_{1\_5} & \rightsquigarrow & A \rightarrow B
\end{array}
$$

As mentioned above, Field then also proves that there are rules where it would be paradoxical to accept the conditional corresponding to that rule, and for which the paracomplete theory therefore cannot assert truth-preservation. He presents two rules as examples of this failure to assert the corresponding conditionals: on the one hand, any instance of the principle of explosion entails an instance of LEM using contraposition (and thus cannot be accepted for pathological instances); while on the other hand, the

[26] Field notes that (similar to the variant reasoning for step (ii)) showing that any "interesting" formula is not a theorem, and therefore not legitimately assertible, amounts to assuming the consistency of the theory (which is the aim of the Consistency argument) and should not be assumed.

instance of modus ponens for a Curry sentence is problematic since it entails triviality (while the conditional entails the Curry sentence at the same time. We will present an argument for the first example (the argument for the Curry instance of modus ponens presents much the same challenges and might not be very insightful to reproduce).

Conditionals for the Principle of Explosion

Field notes that paracomplete logic cannot contain the conditionals corresponding to any pathological instances of the principle of explosion. He shows this in the following way: If it did accept the corresponding conditional to any pathological sentence, it would also accept a rejected instance of the Law of the Excluded Middle (using MP, contraposition and DNE). Thus, since it does not accept the corresponding conditional, it cannot accept the (equivalent by Intersubstitutivity) statement of truth-preservation.[27]

Notably, while Field uses the sentence $0 = 1$ as the conclusion of the principle of explosion, he has introduced this sentence as a sentence regarded as not true. Importantly, the conclusion of the principle of explosion should be triviality, i.e. all sentences. Instead of the false sentence we will here use $\bot$, since it represents triviality in ASPIC-END. However, similarly to the false sentence, it is also necessary to explicitly state that logics reject triviality:

$$\rightsquigarrow \quad Axiom(\langle \neg\bot \rangle, \widetilde{t})$$

This assertion should not cause any problems. A non-pathological reasoning by cases can show that any logic accepts non-triviality. Furthermore, it does not introduce any new inadmissibilities (since any logic that infers $\bot$ already infers all contradictions) and other arguments that introduced $\bot$ have so far been based on the principle of explosion (which is an intuitively strict rule, and thus would not rebuttable).

Then, Field's argument can be realized as two proofs by contradiction. The first proof argument shows that the conditional $C \wedge \neg C \to \bot$ cannot be accepted in paracomplete logic for a pathological C, while the second proof shows that the associated conditional representing truth-preservation $True(\langle C \wedge \neg C \rangle) \to True(\langle \bot \rangle)$ should not be accepted either. The second argument is based on the observation (shown above) that with Intersubstitutivity, truth-preservation can be introduced or removed for conditionals.

Importantly, Field here performs *meta-level reasoning* on why paracomplete logic should not accept the conditional and why it should then not accept truth-preservation for the associated instance of PE. Notably, both these arguments may seem like proofs

[27] Field expresses this argument in the following way:

> [... C]onsider the rule of explosion: $C\&\neg C \vdash 0 = 1$ (for arbitrary C). Paracomplete theories of the kind considered in the previous section contain this rule, but they do not contain the corresponding conditional $C\&\neg C \to 0 = 1$ except for those C where excluded middle can be assumed. (The reason is that $C\&\neg C \to 0 = 1$ is equivalent to $\neg(0 = 1) \to \neg(C\&\neg C)$, which by modus ponens implies $\neg(C\&\neg C)$, which is equivalent to the instance of excluded middle $\neg C \vee C$.) [...] Since we do not have $C\&\neg C \to 0 = 1$, we *should not expect* [emphasis in original] $True(\langle C\&\neg C \rangle) \to True(\langle 0 = 1 \rangle)$. ([Fie06, p. 589])

by contradiction for a pathological C. For example, the first argument reasons that the conditional $C \wedge \neg C \rightarrow \bot$ should not be accepted in the theory, since if it were accepted, then a rejected instance of LEM would follow. However, we would like to argue that Field is intending to dispute whether the conditional can be *a theorem* of paracomplete logic (which would not be a pathological statement). Similarly, the second argument is equally moved to a meta-level, where it does not use any directly pathological statements. We will point out this distinction for the concrete cases below.

In order to provide an ASPIC-END representation for these two arguments, we will have to first introduce the pathological individual C and the rules Field uses in his meta-level reasoning. Notably, C is not a variable, instead it is a nullary predicate acting as the representative for pathological sentences not satisfying LEM. It can be characterized as follows:

$$\begin{aligned}
&\rightsquigarrow \quad Patho(\langle C \rangle)\\
&\Rightarrow \quad \neg Theorem(\langle C \vee \neg C \rangle, paracomp)\\
&\Rightarrow \quad \neg Theorem(\langle \neg C \vee \neg\neg C \rangle, paracomp)
\end{aligned}$$

Notably, the disjunction $\neg C \vee \neg\neg C$ is equivalent to $C \vee \neg C$, using DNE and reasoning by cases (a fact that Field explicitly acknowledges). However, within paracomplete logic, this fact cannot be shown on the object-level (i.e. outside of meta-level reasoning), since it uses reasoning by cases. At the same time, we have not introduced sufficient rules for meta-level reasoning to make these kinds of inferences possible. Unfortunately fully realizing meta-level reasoning is beyond the scope of this thesis, although we will discuss solutions in 6.2.5. As a workaround solution in this thesis, we have introduced the disjunct $\neg C \vee \neg\neg C$ above separately. We will discuss a way to realize meta-level reasoning for the meta-inferences of natural deduction in 6.2.5.

Furthermore, in the first argument, Field uses an instance of conditional contraposition, modus ponens and of the deMorgan laws. These instances can be characterized in ASPIC-END as follows:

$$\begin{aligned}
Ante(a, r_{cCP}(\langle C \wedge \neg C \rangle, \langle \bot \rangle)) \quad &\rightsquigarrow \quad a = \langle C \wedge \neg C \rightarrow \bot \rangle\\
&\rightsquigarrow \quad conc(r_{cCP}(\langle C \wedge \neg C \rangle, \langle \bot \rangle)) =\\
&\qquad \langle \neg\bot \rightarrow \neg[C \wedge \neg C] \rangle\\
Ante(a, r_{MP}(\langle \neg\bot \rangle, \langle \neg[C \wedge \neg C] \rangle)) \quad &\rightsquigarrow \quad a = \langle \neg\bot \rangle \vee a = \langle \neg\bot \rightarrow \neg[C \wedge \neg C] \rangle\\
&\rightsquigarrow \quad conc(r_{MP}(\langle \neg\bot \rangle, \langle \neg[C \wedge \neg C] \rangle)) =\\
&\qquad \langle \neg[C \wedge \neg C] \rangle\\
Ante(a, r_{DML\_4}(\langle C \rangle, \langle \neg C \rangle)) \quad &\rightsquigarrow \quad a = \langle \neg[C \wedge \neg C] \rangle\\
&\rightsquigarrow \quad conc(r_{DML\_4}(\langle C \rangle, \langle \neg C \rangle)) =\\
&\qquad \langle \neg C \vee \neg\neg C \rangle
\end{aligned}$$

Then, Field's first argument can be presented as a form of meta-level reasoning:

Let C be a pathological sentence.

H_{2_1}: $\qquad \Rightarrow \quad Patho(\langle C\rangle)$

By proof by contradiction, $C \wedge \neg C \rightarrow \bot$ is not a theorem.

H_{2_2}: $\qquad \mathsf{Assume}_{\neg}(Theorem(\langle C \wedge \neg C \rightarrow \bot\rangle, paracomp))$

By contraposition, this can infer $\neg\bot \rightarrow \neg[C \wedge \neg C]$.

H_{2_3}: $\qquad \mathsf{Assume}_{\rightarrow}(Ante(a, r_{cCP}(\langle C \wedge \neg C\rangle, \langle\bot\rangle)))$

H_{2_4}: $\quad \mathrm{H}_{2\_3} \quad \rightsquigarrow \quad a = \langle C \wedge \neg C \rightarrow \bot\rangle$

H_{2_5}: $\quad \mathrm{H}_{2\_2}, \mathrm{H}_{2\_4} \quad \rightsquigarrow \quad Theorem(a, paracomp)$

H_{2_6}: $\qquad \rightarrow\text{-}\mathsf{intro}(Ante(a, r_{cCP}(\langle C \wedge \neg C\rangle, \langle\bot\rangle)) \rightarrow$
$Theorem(a, paracomp), \mathrm{H}_{2\_5})$

H_{2_7}: $\qquad \forall\text{-}\mathsf{intro}(\forall a.[Ante(a, r_{cCP}(\langle C \wedge \neg C\rangle, \langle\bot\rangle)) \rightarrow$
$Theorem(a, paracomp)], \mathrm{H}_{2\_6})$

H_{2_8}: $\quad \mathrm{H}_{2\_7} \quad \rightsquigarrow \quad Theorem(conc(r_{cCP}(\langle C \wedge \neg C\rangle, \langle\bot\rangle)), paracomp)$

H_{2_9}: $\qquad \rightsquigarrow \quad conc(r_{cCP}(\langle C \wedge \neg C\rangle, \langle\bot\rangle)) = \langle\neg\bot \rightarrow \neg[C \wedge \neg C]\rangle$

H_{2_10}: $\quad \mathrm{H}_{2\_9}, \mathrm{H}_{2\_8} \quad \rightsquigarrow \quad Theorem(\langle\neg\bot \rightarrow \neg[C \wedge \neg C]\rangle, paracomp)$

Then by modus ponens, $\neg[C \wedge \neg C]$ is a theorem.

H_{2_11}: $\qquad \mathsf{Assume}_{\rightarrow}(Ante(a, r_{MP}(\langle\neg\bot\rangle, \langle\neg[C \wedge \neg C]\rangle)))$

H_{2_12}: $\quad \mathrm{H}_{2\_11} \quad \rightsquigarrow \quad a = \langle\neg\bot\rangle \vee a = \langle\neg\bot \rightarrow \neg[C \wedge \neg C]\rangle$

H_{2_13}: $\qquad \mathsf{Assume}_{\vee}(a = \langle\neg\bot\rangle)$

H_{2_14}: $\qquad \rightsquigarrow \quad Axiom(\langle\neg\bot\rangle, paracomp)$

H_{2_15}: $\quad \mathrm{H}_{2\_14} \quad \rightsquigarrow \quad Theorem(\langle\neg\bot\rangle, paracomp)$

H_{2_16}: $\quad \mathrm{H}_{2\_15}, \mathrm{H}_{2\_13} \quad \rightsquigarrow \quad Theorem(a, paracomp)$

H_{2_17}: $\qquad \mathsf{Assume}_{\vee}(a = \langle\neg\bot \rightarrow \neg[C \wedge \neg C]\rangle)$

H_{2_18}: $\quad \mathrm{H}_{2\_10}, \mathrm{H}_{2\_17} \quad \rightsquigarrow \quad Theorem(a, paracomp)$

H_{2_19}: $\qquad \mathsf{ReasonByCases}(Theorem(a, paracomp), \mathrm{H}_{2\_16}, \mathrm{H}_{2\_18}, \mathrm{H}_{2\_12})$

H_{2_20}: $\qquad \rightarrow\text{-}\mathsf{intro}(Ante(a, r_{MP}(\langle\neg\bot\rangle, \langle\neg[C \wedge \neg C]\rangle)) \rightarrow$
$Theorem(a, paracomp), \mathrm{H}_{2\_19})$

H_{2_21}: $\qquad \forall\text{-}\mathsf{intro}(\forall a.[Ante(a, r_{MP}(\langle\neg\bot\rangle, \langle\neg[C \wedge \neg C]\rangle)) \rightarrow$
$Theorem(a, paracomp)], \mathrm{H}_{2\_20})$

H_{2_22}: $\quad \mathrm{H}_{2\_21} \quad \rightsquigarrow \quad Theorem(conc(r_{MP}(\langle\neg\bot\rangle, \langle\neg[C \wedge \neg C]\rangle)), paracomp)$

H_{2_23}: $\qquad \rightsquigarrow \quad conc(r_{MP}(\langle\neg\bot\rangle, \langle\neg[C \wedge \neg C]\rangle)) = \langle\neg[C \wedge \neg C]\rangle$

H_{2_24}: $\quad \mathrm{H}_{2\_22}, \mathrm{H}_{2\_23} \quad \rightsquigarrow \quad Theorem(\langle\neg[C \wedge \neg C]\rangle, paracomp)$

Then by deMorgan, $\neg C \vee \neg\neg C$ is a theorem.

H_{2_25}: $\qquad \mathsf{Assume}_{\rightarrow}(Ante(a, r_{DML\_4}(\langle C\rangle, \langle\neg C\rangle)))$

H_{2_26}: $\quad \mathrm{H}_{2\_25} \quad \rightsquigarrow \quad a = \langle\neg[C \wedge \neg C]\rangle$

H_{2_27}:	H_{2_24}, H_{2_26}	$\leadsto$	$Theorem(a, paracomp)$
H_{2_28}:			$\to$-$\mathsf{intro}(Ante(a, r_{DML\_4}(\langle C\rangle, \langle\neg C\rangle)) \to Theorem(a, paracomp), H_{2\_27})$
H_{2_29}:			$\forall$-$\mathsf{intro}(\forall a.[Ante(a, r_{DML\_4}(\langle C\rangle, \langle\neg C\rangle)) \to Theorem(a, paracomp)], H_{2\_28})$
H_{2_30}:	H_{2_29}	$\leadsto$	$Theorem(conc(r_{DML\_4}(\langle C\rangle, \langle\neg C\rangle)), paracomp)$
H_{2_31}:		$\leadsto$	$conc(r_{DML\_4}(\langle C\rangle, \langle\neg C\rangle)) = \langle\neg C \vee \neg\neg C\rangle$
H_{2_32}:	H_{2_31}, H_{2_30}	$\leadsto$	$Theorem(\langle\neg C \vee \neg\neg C\rangle, paracomp)$

`This is LEM for a pathological sentence.`

H_{2_33}:		$\Rightarrow$	$\neg Theorem(\langle\neg C \vee \neg\neg C\rangle, paracomp)$
H_{2_34}:	H_{2_32}, H_{2_33}	$\Rightarrow$	$Theorem(\langle\neg C \vee \neg\neg C\rangle, paracomp) \wedge \neg Theorem(\langle\neg C \vee \neg\neg C\rangle, paracomp)$
H_{2_35}:	H_{2_34}	$\leadsto$	$\bot$
H_{2_36}:			$\mathsf{ProofByContrad}(\neg Theorem(\langle C \wedge \neg C \to \bot\rangle, paracomp))$

As mentioned above, Field avoids pathological sentences by performing meta-level reasoning. In particular, in H_{2_2} he does not assume that the pathological sentence $C \wedge \neg C \to \bot$ holds, but instead assumes that this sentence could be a theorem.

The $\to$-Introductions undertaken in H_{2_3}, H_{2_11} and H_{2_25} do not pose a problem either since they are based on non-pathological assumption (i.e. that the antecedents of several different rules be called a). Similarly, the rule leading to H_{2_35} is an instance of PE for non-pathological sentences (in particular, theorem assertions).

The second part of this argument is done in two steps: first, it is shown that the conditional $True(\langle C \wedge \neg C\rangle) \to True(\langle\bot\rangle)$ cannot be a theorem of paracomplete logic, since it is equivalent to the conditional rejected in H_{2_36} (we have sketched this for another case with $H_{1\_0} - H_{1\_6}$). Notably, this conditional does not yet technically assert truth-preservation (though the two statements are functionally equivalent), since it does universally quantify over all antecedents. Truth-preservation could be inferred using $\to$-introduction and $\forall$-introduction, which cannot be performed on the object level. However, as mentioned above, we have decided against providing rules for applying the meta-inferences of natural deduction on the meta-level. Instead, as a workaround solution we will introduce a rule that replicates this specific step.

$$Theorem(\langle TPres(r_{PE}(\langle C\rangle))\rangle, \widetilde{t}) \quad \leadsto \quad Theorem(\langle True(\langle C \wedge \neg C\rangle) \to True(\langle\bot\rangle)\rangle, \widetilde{t})$$

Furthermore, in his argument Field only notes that “we should not expect” the conditional for truth-preservation ([Fie06, p. 589]). Considering the fact that his argument expresses a form of meta-level reasoning, we believe that this careful wording expresses the believe that a theory does not accept the general assumption of truth-preservation if it does not assert of its rules (as meta-level reasoning) that they are truth-preserving. Field also notes in his paper that the truth-preservation of pathological instances is not explicitly rejected, but instead neither asserted nor denied.[28] Again, without rules for meta-inferences in the context of theorems, we are only able to provide a workaround solution:

$$\mathit{AdoptedTheo}(\widetilde{t}), \neg \mathit{Theorem}(\langle \mathit{TPres}(\widetilde{r})\rangle, \widetilde{t}) \quad \rightsquigarrow \quad \neg n(r_{\mathit{AssumeTPres}})$$

Furthermore, the second step of this argument uses an instance of Intersubstitutivity as well as the corresponding instances of the axioms and rules for Alike_T. They can be represented in ASPIC-END with the following rules (for the sake of readability, we will abbreviate $r_{\mathit{Intersub\_2}}(\langle C \wedge \neg C \rightarrow \mathit{True}(\langle \bot \rangle)\rangle, \langle \mathit{True}(\langle C \wedge \neg C\rangle) \rightarrow \mathit{True}(\langle \bot \rangle)\rangle)$ by $r_{\mathit{InSub\_1}}$ and $r_{\mathit{Intersub\_2}}(\langle C \wedge \neg C \rightarrow \bot\rangle, \langle C \wedge \neg C \rightarrow \mathit{True}(\langle \bot \rangle)\rangle)$ by $r_{\mathit{InSub\_2}}$):

$$\begin{array}{rcl}
 & \rightsquigarrow & \mathit{Axiom}\Big(\Big\langle \mathit{Alike}_T(\langle C \wedge \neg C\rangle, \langle \mathit{True}(\langle C \wedge \neg C\rangle)\rangle)\Big\rangle, \widetilde{t}\Big) \\
 & \rightsquigarrow & \mathit{Axiom}\Big(\Big\langle \mathit{Alike}_T(\langle \bot \rangle, \langle \mathit{True}(\langle \bot \rangle)\rangle)\Big\rangle, \widetilde{t}\Big) \\
\mathit{Ante}(a, r_{\mathit{Alike\_1}}) & \rightsquigarrow & a = \Big\langle \mathit{Alike}_T\Big(\langle C \wedge \neg C\rangle, \langle \mathit{True}(\langle C \wedge \neg C\rangle)\rangle\Big)\Big\rangle \\
 & \rightsquigarrow & \mathit{conc}(r_{\mathit{Alike\_1}}) = \Big\langle \mathit{Alike}_T\Big(\langle C \wedge \neg C \rightarrow \mathit{True}(\langle \bot \rangle)\rangle, \\
 & & \langle \mathit{True}(\langle C \wedge \neg C\rangle) \rightarrow \mathit{True}(\langle \bot \rangle)\rangle\Big)\Big\rangle \\
\mathit{Ante}(a, r_{\mathit{Alike\_2}}) & \rightsquigarrow & a = \Big\langle \mathit{Alike}_T\Big(\langle \bot \rangle, \langle \mathit{True}(\langle \bot \rangle)\rangle\Big)\Big\rangle \\
 & \rightsquigarrow & \mathit{conc}(r_{\mathit{Alike\_2}}) = \Big\langle \mathit{Alike}_T\Big(\langle C \wedge \neg C \rightarrow \bot \rangle, \\
 & & \langle C \wedge \neg C \rightarrow \mathit{True}(\langle \bot \rangle)\rangle\Big)\Big\rangle \\
\mathit{Ante}(a, r_{\mathit{InSub\_1}}) & \rightsquigarrow & a = \Big\langle \mathit{True}(\langle C \wedge \neg C\rangle) \rightarrow \mathit{True}(\langle \bot \rangle)\Big\rangle \vee \\
 & & a = \Big\langle \mathit{Alike}_T(\langle C \wedge \neg C \rightarrow \mathit{True}(\langle \bot \rangle)\rangle, \\
 & & \langle \mathit{True}(\langle C \wedge \neg C\rangle) \rightarrow \mathit{True}(\langle \bot \rangle)\rangle)\Big\rangle \\
 & \rightsquigarrow & \mathit{conc}(r_{\mathit{InSub\_1}}) = \langle C \wedge \neg C \rightarrow \mathit{True}(\langle \bot \rangle)\rangle \\
\mathit{Ante}(a, r_{\mathit{InSub\_2}}) & \rightsquigarrow & a = \langle C \wedge \neg C \rightarrow \mathit{True}(\langle \bot \rangle)\rangle \vee \\
 & & a = \Big\langle \mathit{Alike}_T\Big(\langle C \wedge \neg C \rightarrow \bot \rangle, \\
 & & \langle C \wedge \neg C \rightarrow \mathit{True}(\langle \bot \rangle)\rangle\Big)\Big\rangle \\
 & \rightsquigarrow & \mathit{conc}(r_{\mathit{InSub\_2}}) = \langle C \wedge \neg C \rightarrow \bot \rangle
\end{array}$$

Then, the second part of Field's argument can be constructed as follows:

[28] See [Fie06, p. 591]: "[I]n the paracomplete case one never denies that one's rules are unrestrictedly truth-preserving; rather, one *neither asserts nor denies* [emphasis Field] that they are unrestrictedly truth-preserving."

Let the theory accept truth-preservation for the instance of PE.

H_{3_1}: $\quad \mathsf{Assume}_{\neg}(\mathit{Theorem}(\langle \mathit{TPres}(r_{PE}(\langle C \rangle))\rangle, \mathit{paracomp}))$

H_{3_2}: $\quad \mathrm{H}_{3\_1} \rightsquigarrow \mathit{Theorem}(\langle \mathit{True}(\langle C \wedge \neg C\rangle \rightarrow \mathit{True}(\langle \bot \rangle))\rangle, \mathit{paracomp})$

$C \wedge \neg C \rightarrow \mathit{True}(\langle \bot \rangle)$ and the formula for PE truth-preservation are alike.

H_{3_3}: $\quad \rightsquigarrow \mathit{Axiom}\Big(\Big\langle \mathit{Alike}_T\big(\langle C \wedge \neg C\rangle, \big\langle \mathit{True}(\langle C \wedge \neg C\rangle)\big\rangle\big)\Big\rangle, \mathit{paracomp}\Big)$

H_{3_4}: $\quad \mathrm{H}_{3\_3} \rightsquigarrow \mathit{Theorem}\Big(\Big\langle \mathit{Alike}_T\big(\langle C \wedge \neg C\rangle, \big\langle \mathit{True}(\langle C \wedge \neg C\rangle)\big\rangle\big)\Big\rangle, \mathit{paracomp}\Big)$

H_{3_5}: $\quad \mathsf{Assume}_{\rightarrow}(\mathit{Ante}(a, r_{\mathit{Alike}\_1}))$

H_{3_6}: $\quad \mathrm{H}_{3\_5} \rightsquigarrow a = \Big\langle \mathit{Alike}_T\Big(\big\langle C \wedge \neg C\big\rangle, \big\langle \mathit{True}(\langle C \wedge \neg C\rangle)\big\rangle\Big)\Big\rangle$

H_{3_7}: $\quad \mathrm{H}_{3\_4}, \mathrm{H}_{3\_6} \rightsquigarrow \mathit{Theorem}(a, \mathit{paracomp})$

H_{3_8}: $\quad \rightarrow\text{-}\mathsf{intro}(\mathit{Ante}(a, r_{\mathit{Alike}\_1}) \rightarrow \mathit{Theorem}(a, \mathit{paracomp}), \mathrm{H}_{3\_7})$

H_{3_9}: $\quad \forall\text{-}\mathsf{intro}(\forall a.[\mathit{Ante}(a, r_{\mathit{Alike}\_1}) \rightarrow \mathit{Theorem}(a, \mathit{paracomp})], \mathrm{H}_{3\_8})$

H_{3_10}: $\quad \mathrm{H}_{3\_9} \rightsquigarrow \mathit{Theorem}(\mathit{conc}(r_{\mathit{Alike}\_1}), \mathit{paracomp})$

H_{3_11}: $\quad \rightsquigarrow \mathit{conc}(r_{\mathit{Alike}\_1}) = \Big\langle \mathit{Alike}_T\Big(\big\langle C \wedge \neg C \rightarrow \mathit{True}(\langle \bot \rangle)\big\rangle, \big\langle \mathit{True}(\langle C \wedge \neg C\rangle) \rightarrow \mathit{True}(\langle \bot \rangle)\big\rangle\Big)\Big\rangle$

H_{3_12}: $\quad \mathrm{H}_{3\_10}, \mathrm{H}_{3\_11} \rightsquigarrow \mathit{Theorem}\Big(\Big\langle \mathit{Alike}_T\big(\big\langle C \wedge \neg C \rightarrow \mathit{True}(\langle \bot \rangle)\big\rangle, \big\langle \mathit{True}(\langle C \wedge \neg C\rangle) \rightarrow \mathit{True}(\langle \bot \rangle)\big\rangle\big)\Big\rangle, \mathit{paracomp}\Big)$

The formulae $C \wedge \neg C \rightarrow \bot$ and $C \wedge \neg C \rightarrow \mathit{True}(\langle \bot \rangle)$ are alike.

H_{3_13}: $\quad \rightsquigarrow \mathit{Axiom}\Big(\Big\langle \mathit{Alike}_T\big(\langle \bot \rangle, \big\langle \mathit{True}(\langle \bot \rangle)\big\rangle\big)\Big\rangle, \mathit{paracomp}\Big)$

H_{3_14}: $\quad \mathrm{H}_{3\_13} \rightsquigarrow \mathit{Theorem}\Big(\Big\langle \mathit{Alike}_T\big(\langle \bot \rangle, \big\langle \mathit{True}(\langle \bot \rangle)\big\rangle\big)\Big\rangle, \mathit{paracomp}\Big)$

H_{3_15}: $\quad \mathsf{Assume}_{\rightarrow}(\mathit{Ante}(a, r_{\mathit{Alike}\_2}))$

H_{3_16}: $\quad \mathrm{H}_{3\_15} \rightsquigarrow a = \Big\langle \mathit{Alike}_T\Big(\big\langle \bot \big\rangle, \big\langle \mathit{True}(\langle \bot \rangle)\big\rangle\Big)\Big\rangle$

H_{3_17}: $\quad \mathrm{H}_{3\_14}, \mathrm{H}_{3\_16} \rightsquigarrow \mathit{Theorem}(a, \mathit{paracomp})$

H_{3_18}: $\quad \rightarrow\text{-}\mathsf{intro}(\mathit{Ante}(a, r_{\mathit{Alike}\_2}) \rightarrow \mathit{Theorem}(a, \mathit{paracomp}), \mathrm{H}_{3\_17})$

H_{3_19}: $\quad \forall\text{-}\mathsf{intro}(\forall a.[\mathit{Ante}(a, r_{\mathit{Alike}\_2}) \rightarrow \mathit{Theorem}(a, \mathit{paracomp})], \mathrm{H}_{3\_18})$

H_{3_20}: $\quad \mathrm{H}_{3\_19} \rightsquigarrow \mathit{Theorem}(\mathit{conc}(r_{\mathit{Alike}\_2}), \mathit{paracomp})$

H_{3_21}: $\quad \rightsquigarrow \mathit{conc}(r_{\mathit{Alike}\_2}) = \Big\langle \mathit{Alike}_T\Big(\big\langle C \wedge \neg C \rightarrow \bot \big\rangle, \big\langle C \wedge \neg C \rightarrow \mathit{True}(\langle \bot \rangle)\big\rangle\Big)\Big\rangle$

H_{3_22}: $\quad \mathrm{H}_{3\_20}, \mathrm{H}_{3\_21} \rightsquigarrow \mathit{Theorem}\Big(\Big\langle \mathit{Alike}_T\big(\big\langle C \wedge \neg C \rightarrow \bot \big\rangle, \big\langle C \wedge \neg C \rightarrow \mathit{True}(\langle \bot \rangle)\big\rangle\big)\Big\rangle, \mathit{paracomp}\Big)$

By Intersubstitutivity, $C \wedge \neg C \rightarrow \mathit{True}(\langle \bot \rangle)$ must be a theorem.

H_{3_23}: $\quad \mathsf{Assume}_{\rightarrow}(\mathit{Ante}(a, r_{\mathit{InSub}\_1}))$

H_{3_24}: H_{3_23} $\rightsquigarrow$ $a = \Big\langle True(\langle C \wedge \neg C\rangle) \to True(\langle \bot \rangle)\Big\rangle \vee a = \Big\langle Alike_T\big(\langle C \wedge \neg C \to True(\langle \bot \rangle)\rangle, \langle True(\langle C \wedge \neg C\rangle) \to True(\langle \bot \rangle)\rangle\big)\Big\rangle$

H_{3_25}: $\mathsf{Assume}_{\vee}(a = \Big\langle True(\langle C \wedge \neg C\rangle) \to True(\langle \bot \rangle)\Big\rangle)$

H_{3_26}: H_{3_2}, H_{3_25} $\rightsquigarrow$ $Theorem(a, paracomp)$

H_{3_27}: $\mathsf{Assume}_{\vee}(a = \Big\langle Alike_T\big(\langle C \wedge \neg C \to True(\langle \bot \rangle)\rangle, \langle True(\langle C \wedge \neg C\rangle) \to True(\langle \bot \rangle)\rangle\big)\Big\rangle)$

H_{3_28}: H_{3_12}, H_{3_27} $\rightsquigarrow$ $Theorem(a, paracomp)$

H_{3_29}: $\mathsf{ReasonByCases}(Theorem(a, paracomp), H_{3\_26}, H_{3\_28}, H_{3\_24})$

H_{3_30}: $\to$-$\mathsf{intro}(Ante(a, r_{InSub\_1}) \to Theorem(a, paracomp), H_{3\_29})$

H_{3_31}: $\forall$-$\mathsf{intro}(\forall a.[Ante(a, r_{InSub\_1}) \to Theorem(a, paracomp)], H_{3\_30})$

H_{3_32}: H_{3_31} $\rightsquigarrow$ $Theorem(conc(r_{InSub\_1}), paracomp)$

H_{3_33}: $\rightsquigarrow$ $conc(r_{InSub\_1}) = \langle C \wedge \neg C \to True(\langle \bot \rangle)\rangle$

H_{3_34}: H_{3_32}, H_{3_33} $\rightsquigarrow$ $Theorem\Big(\langle C \wedge \neg C \to True(\langle \bot \rangle)\rangle, paracomp\Big)$

`By Intersubstitutivity,` $C \wedge \neg C \to \bot$ `must be a theorem.`

H_{3_35}: $\mathsf{Assume}_{\to}(Ante(a, r_{InSub\_2}))$

H_{3_36}: H_{3_35} $\rightsquigarrow$ $a = \langle C \wedge \neg C \to True(\langle \bot \rangle)\rangle \vee a = \Big\langle Alike_T\Big(\langle C \wedge \neg C \to \bot\rangle, \langle C \wedge \neg C \to True(\langle \bot \rangle)\rangle\Big)\Big\rangle$

H_{3_37}: $\mathsf{Assume}_{\vee}(a = \langle C \wedge \neg C \to True(\langle \bot \rangle)\rangle)$

H_{3_38}: H_{3_34}, H_{3_37} $\rightsquigarrow$ $Theorem(a, paracomp)$

H_{3_39}: $\mathsf{Assume}_{\vee}(a = \Big\langle Alike_T\Big(\langle C \wedge \neg C \to \bot\rangle, \langle C \wedge \neg C \to True(\langle \bot \rangle)\rangle\Big)\Big\rangle)$

H_{3_40}: H_{3_22}, H_{3_39} $\rightsquigarrow$ $Theorem(a, paracomp)$

H_{3_41}: $\mathsf{ReasonByCases}(Theorem(a, paracomp), H_{3\_38}, H_{3\_40}, H_{3\_36})$

H_{3_42}: $\to$-$\mathsf{intro}(Ante(a, r_{InSub\_2}) \to Theorem(a, paracomp), H_{3\_41})$

H_{3_43}: $\forall$-$\mathsf{intro}(\forall a.[Ante(a, r_{InSub\_2}) \to Theorem(a, paracomp)], H_{3\_42})$

H_{3_44}: H_{3_43} $\rightsquigarrow$ $Theorem(conc(r_{InSub\_2}), paracomp)$

H_{3_45}: $\rightsquigarrow$ $conc(r_{InSub\_2}) = \langle C \wedge \neg C \to \bot\rangle$

H_{3_46}: H_{3_44}, H_{3_45} $\rightsquigarrow$ $Theorem\Big(\langle C \wedge \neg C \to \bot\rangle, paracomp\Big)$

`However, this contradicts the first part of the argument.`

H_{4_1}: H_{3_46}, H_{2_36} $\rightsquigarrow$ $\bot$

H_{4_2}: $\mathsf{ProofByContrad}(\neg Theorem(\langle TPres(r_{PE}(\langle C\rangle))\rangle, paracomp), H_{4\_1})$

`This undercuts the general assumption that all rules preserve truth.`

$$\begin{array}{llll}
\mathrm{H}_{4\_3}\text{:} & & \Rightarrow & AdoptTheo(paracomp) \\
\mathrm{H}_{4\_4}\text{:} & \mathrm{H}_{4\_3}, \mathrm{H}_{4\_2} & \rightsquigarrow & \neg n(r_{AssumeTPres})
\end{array}$$

As in the first step of the argument, Field here uses meta-level reasoning to avoid pathological statements. Instead of the pathological truth-preservation, in H_{3_1} Field argues from the assumption that truth-preservation for the problematic instance of PE must be a *theorem*.

The other instances of the rejected meta-inferences (i.e. H_{3_5}, H_{3_15}, H_{3_23} and H_{3_35}) do not assert pathological sentences and are therefore not problematic for Field's paracomplete theory.

Furthermore, this argument ultimately rejects the assumption that all rules must be truth-preserving. As in Field, this is an attack on step (2) of the Consistency Argument, specifically an undercut on the rule used for C_{2_2}. Notably, the assumption cannot be rebutted since (as Field argues) truth-preservation itself is neither accepted nor denied (we have rejected it only within meta-level reasoning).

Chapter 6

Discussion

6.1 Results

In the last chapters, we have presented a framework for the philosophy of logic and we have replicated some of the argumentation made by Hartry Field in his article "Truth and the Unprovability of Consistency." For this, we have presented both our modeling methodology and formalized the syntactic construction essential for discussing the advantages and shortcomings of various theories of truth. While this modeling is by no means *the* definitive approach to philosophy of logic, we believe that it is a first attempt worth examination. As mentioned in 4.1, the aim of this thesis was to provide proof that ASPIC-END is a good fit for philosophical discussions about logic and truth, and to survey the challenges that such modeling in ASPIC-END would have to face. Wherever possible, we have suggested possible solutions for these challenges.

All things considered, we have found that argumentation theory (and ASPIC-END in particular) are well-suited for formalizing discussions in philosophy of logic. The proof-theoretic basis of argumentation theory is central for the meta-terms, which were instrumental in representing the truth predicate. The fact that ASPIC-style frameworks can represent multiple contradictory approaches simultaneously allows for the representation of the different theories of truth on the same level, while ASPIC-END's way of handling intuitively strict rules is important for the formalization of logical building blocks that may need to be rejected.

In this chapter, we will review and discuss the modeling done in the last chapters. First, we will discuss how effective the modeling principles outlined in 4.1 were at guiding the modeling process. Then, we will discuss meta-terms and rule schemes, the new elements we introduced into the ASPIC-END formalization. Lastly, we will focus on the actual modeling: first on the general elements introduced in Chapter 4 and then on the formalizations of the both the Consistency argument and the different theories of truth Field considers.

Furthermore, we will also provide an outlook beyond this thesis. This thesis encountered both some issues that we were not yet able to provide a satisfying solution for and some areas were further research would have been beyond the scope of this thesis. For some of these areas, we will outline possible solutions we believe would prove feasible and insightful.

6.1.1 Methodology

In 4.1.1, we have outlined principles to guide the modeling process: *minimality*, *authenticity* and *explicitness* as well as an inclination towards using the inferences of natural deduction. While they are all useful to focus the modeling efforts, not all of them should be used without care, though we believe that their use within this thesis was warranted.

Within the limitations of this thesis – modeling Field's arguments and assessing their consequences by hand – it was beneficial to strive for minimality. While the argumentation theory we have presented here is large and consequently difficult to fully grasp, minimizing the occurrence of redundant rules leaves the corpus of rules easier to parse and reduces the chance of unintended consequences that went unnoticed.

Authenticity and explicitness are essential goal posts to strive for when replicating the arguments of another individual, such as was done with Field's arguments in his article. However, these goals impose a limit on the applicability of this thesis: it aims to be authentic towards Field's arguments, and therefore might choose his method of representing arguments and knowledge about philosophy of logic over others or contain any other personal biases. When using argumentation theory (or another formalism) as a platform to represent multiple logician's perspectives, it would be important to be mindful of the balancing act between how closely the argumentation theory should be able to reproduce the individual logician's arguments (and be subject to their inherent biases) and how closely the argumentation theory should represent the common knowledge of the field of their discussion. However, within this thesis the intention was to replicate only Field's perspective, leaving authenticity and explicitness as useful goals.

6.1.2 The ASPIC-END Formalization

Meta-Terms

One of the challenges of modeling in the philosophy of logic was the existence of the truth predicate and the self-referentiality that predicate required: Without the ability to refer to a formula within another formula of the same logical language (i.e. Gödel codes), the truth predicate cannot be adequately formalized. Thus, the formalization of *meta-terms* in the logical language of this thesis' argumentation theory was essential to the modeling process.

However, the definition of meta-terms as presented in 3.3.1 presents a versatile tool with its own advantages and challenges. Unfortunately, the use of meta-terms leads to a countably infinite argumentation theory in most situations. This, unfortunately, makes it impossible to fully enumerate the resulting argumentation framework – or even the argumentation theory. While, as noted in 1.1, it might not be feasible to fully analyze an argumentation framework powerful and complex enough to express multiple theories of truth, it is still another obstacle for any analysis of the theory.

At the very least, while the resulting logical language is large and complex, it should not be difficult to decide whether a given string is part of the logical language, meaning that the resulting language is well-defined and therefore feasible for use with proof-theoretic formalisms (such as argumentation theory).[1]

Furthermore, while within this thesis we mostly use meta-terms to represent formulae (in the context of truth or theoremship), the definition provided also contains the symbols of the underlying logical language. It is easy to imagine that the meta-term construction could be used to define a *syntax theory*[2] for the language it is representing within the language. We will discuss the advantages of syntax theory in greater detail in 6.2.

Rule Schemes

Besides the introduction of the syntactically necessary meta-terms, we introduced *rule schemes* as a means of simplifying sets of rules with similar structure. Wile this expresses a very common intuition, it is nonetheless necessary to look at it with a critical eye. The three types of rule schemes we defined produce a countably infinite rule set by instantiating every rule into separate instances for every possible value of the occurring schematic symbols (of which only a small portion may be relevant to the logic). This, together with the blow-up caused by meta-terms, makes a full exploration of the argument space difficult.

Similarly to meta-terms, it is important to ask whether this new construction is well-defined and therefore usable in defining the argumentation theory of this thesis. And it is: given a rule, the question whether this rule is an instance of a rule scheme is decidable. We will not go into great detail for this since it is not the focus of this thesis, but we believe a modified unification algorithm may be able to provide instantiations for all schematic terms, formulae and their arguments.

Another small disappointment with the "rule schemes" approach is the fact that any set of connected rules is instantiated as a set of almost completely disparate rules. This is not a problem for referencing these rules, since it is easily possible to use schematic variables in the names of these rules as well. However, the *neg* function shows an example for an inefficiency that arises from this disparateness: While it might intuitively seem simpler – and less redundant – to express a negated occurrence of a formula using rule schemes and the negation symbol $\neg$, this would result in introducing separate instances. In order to formulate some rules "universally" (i.e. quantified), we can only use variables, which cannot be negated. A possible solution for this inefficiency would be the introduction of a syntax theory for the logical language itself within the argumentation theory, where negating any term could be achieved by applying a concatenation function to add the $\neg$ symbol. We will discuss the possibilities of a syntax theory in greater detail in 6.2.2.

[1] An algorithm checking whether a given string belongs to this language can be easily imagined, recursively checking that all brackets form proper pairs, that the meta-terms only occur in the place of regular terms and that their contents are simple symbols or again formulae or terms. Thus membership to this logical language is decidable.

[2] A syntax theory defines all well-formed sentences within a language.

While it might seem like rule schemes result mostly in downsides for the theory, we would argue against this since the alternative, i.e. providing only the "relevant" rule instances, runs into another problem: that of how to determine relevance. If only those instances strictly necessary to formulate a given argument are included, then the argumentation theory may not be able to recognize any arguments outside of its specific context and may overlook legitimate arguments. This, in our opinion, would prove a detriment to the definition of a general model, i.e. a model used for more than the replication of any individual philosopher's argument). A balance needs to be struck between keeping the rule corpus as small as possible and still including a rich enough set of rule instances to make general discussion insightful. It is an open question how (and if) this can be achieved without the "rule schemes" approach, i.e. including all possible instances.

6.1.3 Concepts of the Philosophy of Logic

Quantifier Symbols

The rules for the $\forall$ and $\exists$ symbols were relatively straightforward to model after their intuition (as explained in 4.2.3, we referenced [Vel06] for a useful characterization of their intuitions). While the $\forall$-introduction is part of the definition of ASPIC-END (as it needs to restrict which variables can be introduced in the universal quantifier), we managed to express the other properties of the quantifiers as rules.

In particular, the rule for $\exists$-elimination can be formulated effectively using meta-terms and a special function symbol g to uniquely introduce the specific element that the existential quantification referred to. While this construction is somewhat problematic in terms of readability (it requires the full quantified formula as an argument of g), it is still useful since it offers specificity and does not require any additional formalization.

The Equality Predicate

In comparison to the relatively unproblematic handling of the quantifiers, our realization of the equality predicate $=$ might require a more careful consideration. Notably, most mathematical logics do not contain the formal underpinning to characterize equality,[3] so some logics admit a special predefined predicate symbol for $=$. Importantly, in mathematical logics the presence of an equality symbol has an influence on the expressive power of the logic. In this argumentation theory on the other hand, the AC0 properties can be fully formalized using rule schemes; allowing all theories of truth to use the equality predicate. We would however like to argue that this is an intended result: reasoning, especially in natural language, is easily capable of characterizing equality, and should therefore always admit an equality predicate.

We have found equality most useful in defining rules using *Ante* and *conc*: antecedents and conclusion of the rule can be expressed individually and inserted into

[3] Equality is usually characterized with the AC0 properties: reflexivity, symmetry, transitivity and the substitution property. See 4.2.2 for details.

other arguments using the substitution property.

Induction

In this thesis, we have presented a formulation of structural induction on theorems (in particular, on the structure of their proofs). Both induction start and induction step are here done as antecedents of a single rule. While formally, induction is usually presented as an axiom scheme, in his article Field presented induction as an inference rule and we have decided to follow that representation. As argued in 4.3.5, this distinction is not significant for induction.

Although Field's arguments did not utilize induction very frequently, it is still an essential element of the Consistency argument. Our rule formulation has proven straightforward to use in this context and it is easy to imagine this rule used in another structural induction (based on theorems) just as effectively.

Other formulations of induction, either based on other structures or on natural numbers, are just as straightforward to formalize as the induction we have provided in this thesis. In fact, we have found that the Consistency argument can be formulated very similarly using induction over the height of theorem proof trees (i.e. natural numbers). We have decided against presenting this formulation in this thesis since it requires introducing several concepts (e.g. natural numbers, theorem proof trees) that structural induction avoids.

The Truth Predicate

In 4.3.1, we have introduced the *True* predicate for ASPIC-END. It reflects the intuition of the philosophy of logic very accurately in that it uses Gödel codes (i.e. meta-terms) to refer to those sentences the extension generally regards as true. The T-scheme, essential for characterizing truth, can be presented very straightforwardly using this predicate. Furthermore, we have used the truth predicate extensively both in the Consistency Argument and in the discussions of the individual theories of truth, finding it to effective in all of these cases.

As an alternative for the T-scheme, Field introduces the Intersubstitutivity principle. The ASPIC-END modeling presented in 4.3.4 reproduces the intuition behind Field's , although it might disagree on the details to some extent. The Intersubstitutivity as introduced by Field might be characterized as a replacement (i.e. substitution) of "alike" formulae within any context. The naive approach to reproducing this might use the following rules:

$$\begin{array}{rcl} \widehat{p}(\langle\widehat{q}\rangle) & \rightsquigarrow & \widehat{p}(\langle True(\langle\widehat{q}\rangle)\rangle) \\ \widehat{p}(\langle True(\langle\widehat{q}\rangle)\rangle) & \rightsquigarrow & \widehat{p}(\langle\widehat{q}\rangle) \end{array}$$

However, there are multiple problems with this. First, this only reproduces the replacement *within a context* $\widehat{p}$. For the case without a context (i.e. when inferring

$True(\langle\widehat{q}\rangle)$ just from $\widehat{q}$), this approach would additionally need the rules T-Elim and T-Introd, which Field does not necessarily assert for Intersubstitutivity. And even more devastatingly, this approach can only replace the meta-term $\langle True(\langle\widehat{q}\rangle)\rangle$, not the truth-assertion.

The approach we have introduced solves both of these problems. It calculates the results of this replacement with the $Alike_T$ predicate in a bottom-up manner. While this might require several steps to show this alikeness for more complex contexts, it can be done in a consistent manner and includes both the $\widehat{q}$ assertions as well as their meta-terms.

The Semantic Paradoxes

As noted in an earlier chapter, the semantic paradoxes or mainly based on the ability to self-reference. Both Liar and Curry's paradox refer to themselves from within their definition. In this thesis, instead of providing a syntax theory that would allow the explicit construction of these paradoxes, we have used the *DefinedAs* predicate (based on the meta-terms) to define these paradoxical sentences. It is easy to see that this definition is effective in bringing forth the paradoxical nature of these sentences: For the Liar sentence, we have presented the arguments E_{1_1} to E_{1_9}, using T-OUT to show that the Liar sentence must hold, and F_{1_1} to F_{1_16}, using T-IN to show that the Liar sentence must not hold. Consequently, this argumentation theory can show that any classical logic that accepts the full T-scheme accepts inconsistent theorems. Notably, while we have not provided arguments for the Curry paradox, the proof-theoretic argument sketched in 2.2.2 can be implemented rather straightforwardly.

Infrastructure for Theories of Truth

The "building blocks" of theories of truth, i.e. the typical axioms and rules of inference, were introduced in 4.3.2. They are represented by intuitively strict rules, where axioms are represented as rules without antecedents.

Notably, these rules are not inherently restricted to specific theories. Instead, these rules are included generally within the argumentation theory and can be used to construct any arguments. However, with the *Rule* and *Axiom* predicate, these rules can be accepted and rejected "locally," i.e. within specific theories of truth and their corresponding extensions, resulting in an undercut for any locally rejected rule application. For that reason, the rule application cannot occur within any admissible extension that rejects that rule. We will discuss the specifics of the separation into different theories of truth below, but on the topic of the building blocks, it suffices to note that extensions are in fact capable of rejecting building blocks (and therefore can represent theories of truth).

Importantly, these rules corresponding to the building blocks are accepted by "default" (i.e. they can be used without problem so long as they are not explicitly rejected), producing a notable asymmetry between these two cases. Fortunately, this does not present a problem within argumentation theory.

Furthermore, we have made a clear distinction between axioms and rules of inference. While within proof theory, there is a definitive difference between these two concepts, it is worth considering whether the same holds for the argumentation theory. In fact, while by definition separate concepts, we have treated axioms and rules of inference very similarly: both rules and axioms are accepted or rejected using an *Accept* predicate, leading to undercuts in either case. Furthermore, both concepts are realized using intuitively strict rules and can (by default) be applied in any extension. In fact, it would be possible to simplify the current argumentation theory by considering axioms as a special type of rule of inference that has no antecedents. While this treatment might be worth consideration, we have decided against it since it is not authentic to the philosophy of logic.

Besides the "building blocks," in 4.3.1 we introduced the "infrastructure" (i.e. predicates and rules to characterize them) for the different theories of truth. In particular, we introduced the predicates *AdoptTheo* and *Theorem* to distinguish them and their characteristics.

Ultimately, the goal of this thesis is to produce an argumentation theory in which every extension represents a separate theory of truth. While, as noted in the introduction, it is impractical to fully analyze the argumentation theory for philosophy of logic due to its size and complexity, we believe that we have succeeded at enforcing separate extensions for the different theories of truth: Proof by Contradiction can be used to reject any incongruous theories (that have contradictory rule or axiom assertions) in the presence of PE and since these *AdoptTheo* instances are produced with defeasible rules, they can be rebutted. Below, this argument is sketched for two theories t_1 and t_2 that disagree on any rule $r*$.

$$\begin{array}{llll}
\multicolumn{4}{l}{\texttt{We adopt a theory } t_1\texttt{.}} \\
\mathrm{I}_1: & & \Rightarrow & \mathit{AdoptTheo}(t_1) \\
\mathrm{I}_2: & & \rightsquigarrow & \mathit{Rule}(r*, t_1) \\
\mathrm{I}_3: & \mathrm{I}_1, \mathrm{I}_2 & \rightsquigarrow & \mathit{Accept}_r(r*) \\
\multicolumn{4}{l}{\texttt{Then, we reject any other incongruous theories.}} \\
\mathrm{I}_4: & & & \mathrm{Assume}_\neg(\mathit{AdoptTheo}(t_2)) \\
\mathrm{I}_5: & & \rightsquigarrow & \neg \mathit{Rule}(r*, t_2) \\
\mathrm{I}_6: & \mathrm{I}_4, \mathrm{I}_5 & \rightsquigarrow & \neg \mathit{Accept}_r(r*) \\
\mathrm{I}_7: & \mathrm{I}_3, \mathrm{I}_6 & \rightsquigarrow & \bot \\
\mathrm{I}_8: & & & \mathrm{ProofByContrad}(\neg \mathit{AdoptTheo}(t_2), \mathrm{I}_7)
\end{array}$$

Importantly, this counterargument for adopting other theories can be constructed symmetrically (i.e. when adopting t_2, one can construct a counterargument rejecting the adoption of t_1). Fortunately, these symmetrical attacks between extensions corresponding to different theories do not make any extension inadmissible. Any such attack is automatically defended against by the symmetrical counterattack.

Notably, this infrastructure accepts all (consistent) theories of truth, regardless of how "absurd" they may seem. In fact, for some of the theories Field considers, he himself notes that they do not have any proponents or can be considered absurd. We believe this speaks to the larger question of how to find the theory of truth representing "correct" reasoning amongst several internally consistent and non-trivial theories. The property of absurdity (or other disadvantageous properties of theories) can be realized in ASPIC-END by additional rules or predicates, which may be able to make extensions that fulfill these properties inadmissible.

Notably, this outlined way of checking theories for absurdity would need the arguments establishing absurdity to be formulated using only rules the theory itself accepts, since this argument itself could be attacked otherwise. Importantly, we have not found a philosophical text discussing this proof-theoretical requirement for theory rejection in detail. We believe this condition could be an important new perspective on discourse in philosophy of logic, resulting from the structured argumentation used in this thesis.

6.1.4 The Theories of Truth

The focus of Field's article was on the relationship several theories of truth have to the Consistency Argument, a generically constructed argument for how a theory should be able to show that it is consistent. In this thesis, we have both presented a representation for the Consistency Argument as well as the different theories and their respective solutions to the argument in ASPIC-END. Below, we will discuss our approach for modeling Field's article and outline areas that merit further consideration.

The Consistency Argument

In 5.2, we have presented a formulation of Field's Consistency Argument in ASPIC-END. The presented argument follows Field's argumentative structure, authentically reproducing the steps and assumptions Field made for his argument. Although generally well-reproduced, we believe there are parts of our formalization that merit closer examination.

A major feature of the Consistency Argument is the proof by induction that all theorems of the adopted theory must be true. This proof is reproduced faithfully, although it should be noted that it is the only time in Field's article (and consequently, this thesis) that induction is used. Furthermore, in this instance the individual proofs for induction start and induction step are almost trivial in that they follow directly from assumptions made about the theories of truth (i.e. that axioms are considered true, and that all rules preserve truth). For this reason, this proof is unfortunately not a particularly impressive use case of the presented induction rule. However, in writing this thesis, we had opportunity to use the induction rule successfully in more complex circumstances.

Furthermore, we believe that representing the assumptions for the Consistency Argument using defeasible rules without antecedents was well-chosen: In his article, Field focuses on how the different theories of truth contradict the Consistency argument. In this thesis, we can reproduce this by constructing arguments attacking the Consistency argument, many of which are rebutting these assumptions. Notably, these rebuttals are successful since the assumptions are represented using defeasible rules. Furthermore, the representation as rules without antecedents is authentic to Field's article, since he introduced these assumptions as unconditional statements. This representation has the added benefit that while defeasible rules cannot be used in hypotheticals (i.e. the meta-inferences of natural deduction), the conclusions of rules without antecedents can (even if the rule itself is defeasible).

One construction that merits further discussion is the use of t as an arbitrary adopted theory. This required the additional assumption that $AdoptTheo(t)$ holds (i.e. C_0) and makes the Consistency Argument an →-Introduction that is only resolved with C_{4_17}, the last step of the Consistency argument. While this construction is useful when characterizing the Incompleteness theorem in that the Consistency argument cannot be a correct argument (since a consistent adopted theory cannot show its own consistency), it is somewhat counterintuitive to the approach of structured argumentation theory. Any extension that this argument might belong to considers one specific theory, and would therefore not require an additional t (or an assumption for adopting it). Furthermore, according to the Incompleteness Theorem, any extension belonging to a consistent theory would not accept either form of the argument.

Field's interest in the Consistency Argument is ascertaining where it breaks down for the different theories of truth. However, Field did not put any significant emphasis on the Incompleteness theorem in his article. This highlights one of the disadvantages of following Field's article in representing philosophy of logic. Without him explicitly discussing the Incompleteness theorem, there was no incentive to represent it in ASPIC-END either. Consequently, the Consistency Argument (in either form) is ultimately not particularly relevant for the admissibility of different extensions.

Classical Theories

In 5.3.2, we have presented formalizations for the classical theories under consideration. In all three cases, KF, classical hyperdialetheism and classical dialetheism, we managed to formulate the theories, their restrictions and idiosyncrasies very straightforwardly.

Each of the theories presented their own proofs, some of which the other theories are not able to reproduce – for example, F_{1_1} to F_{1_20} present the Assertions of the Liar sentence that classical dialetheism accepts, which KF cannot reproduce since it rejects T-IN (and would therefore produce an undercut on this argument). This is another demonstration that the individual theories must be realized by individual extensions.

On another note, while some of the counterarguments against the Consistency argument are technically complicated, we believe that their main thrusts are easy to grasp and they correct.

Paracomplete Theories

In 5.3.3, we have presented Field's paracomplete theory of truth in ASPIC-END. The definition of the logic can be achieved almost without problem: in- and excluding the relevant building blocks is very straightforward (after all, Field lists the included building blocks explicitly). The most notable challenge is in defining pathologicality and restricting the meta-inferences to non-pathological cases. We believe the rules presented fully achieve the restriction as intended by Field.

However, some of the approaches that worked very well for classical logic could not be applied equally well to paracomplete logic: For example, the conditional $p \to p$ had to be introduced explicitly, instead of being the trivial to infer using $\to$-Introduction. While we believe the assertion of this conditional should not be controversial, another difference from classical logic might warrant closer inspection: *meta-reasoning*. Since paracomplete logic is the first logic to restrict any form of meta-inference, the normal modes of reasoning could not be used here. We believe Field also (implicitly) acknowledges this by using more careful forms of phrasing.[4] To avoid using pathological meta-inferences, we have presented Field's arguments as a form of meta-reasoning, i.e. reasoning about the theorems of paracomplete logic. This posed additional problems since no classical logic utilized this mode of reasoning, leading us to sketch some inferences instead of presenting explicit rules. We will discuss possible solutions for these workarounds in 6.2.5, but here suffice it to say that the arguments, while complicated, are rather straightforward to formulate in meta-reasoning.

6.2 Future Work

Above, we have made an account of the different modeling challenges we encountered in this thesis and whether the solutions developed are generally successfully or should be considered carefully for other contexts.

Here we would like to present challenges that we have not been able to develop a fully formed solution for. For some of these challenges, such as the definition of a syntax theory or the implementation of meta-reasoning, we will outline a possible path to a solution, though we believe in all cases that the problems are complex and interesting enough that future research should prove worthwhile.

6.2.1 Using Explanatory Argumentation Frameworks

As (briefly) mentioned in 3.2, ASPIC-END incorporates *explanatory argumentation* in its definition. The typical ASPIC-style structured argumentation is enhanced by the ability to both outline specific *explananda* (i.e. facts to be explained) and the ability to judge whether a given argument explains any of these explananda. Using the resulting *explanatory argumentation framework* (EAF), an argumentation theory can highlight extensions that explain many explananda or provide better explanations (we will refrain from the technical details here, since it is beyond the scope of this thesis).

As a potentially useful application of the idea of EAFs, let us look at paradoxes in our argumentation theory. Every semantic paradox we introduced has the ability to prove the theory of truth that contains it trivial (i.e. inferring $\bot$). Accordingly, singling out the extensions that are capable of averting triviality could be of great interest. For

[4] We have noted, for example, that Field tries to express a proof by contradiction as "Since we do not have $C\&\neg C \to 0 = 1$, we *should not expect* [emphasis in original] $True(\langle C\&\neg C\rangle) \to True(\langle 0 = 1\rangle)$.".

this purpose, it is possible to take as explananda all the arguments of the argumentation theory that derive $\bot$, and consider such a paradoxical argument explained when it is attacked.

However, while this would be an interesting application, there is currently the complication that it is not as helpful as it might seem at first glance: if there is at least one theory that is consistent (i.e. which attacks all arguments that derive $\bot$), then that theory is admissible and thus automatically preferable to all theories that do not handle all arguments for $\bot$. If no theory can handle all arguments for $\bot$, then no admissible theory exists (which makes the additional semantics provided by explanations obsolete). Unfortunately, using EAFs for working with semantic paradoxes remains a challenge that will require more consideration.

As noted in 3.2, explanatory argumentation can be useful in realizing more specific reasoning goals, while the aim of this thesis is to generally survey the modeling challenges structured argumentation of philosophy of logic has to face. Thus, while EAFs were not essential to the formulation of this thesis, we believe that there are reasoning tasks that would benefit from the additional specificity that EAFs offer.

For example, Gödel's Incompleteness theorem states that no consistent theory can prove its own consistency, which (as we noted above) has not been implemented in this thesis. Explanatory argumentation could be used to select those theories (or rather, corresponding extensions) that are capable of "explaining away" their instance of Field's Consistency Argument. Comparing explanatory power or depth could offer additional insights into the relations between these theories.

6.2.2 Defining a Syntax Theory

As mentioned earlier, an important trait of the philosophy of logic is the fact that object- and meta-language are not entirely separated. For example, it should be possible to discuss *how* an argument is formulated – this type of meta-level discussion of the underlying syntax of an argument should itself be possible on the object level. This type of reasoning can be realized using a *syntax theory*: rules that define the syntax of the argumentation theory on the object level.

We believe that it is possible to construct a syntax theory for the logical language defined in 3.3. In fact, the ability to refer to syntactical features was one of the key motivations for introducing meta-terms, which is the reason why there are meta-terms not only for fully formed formulae, but also for individual syntactical objects as well (the definition focuses on predicate and function symbols, but a syntax theory might in fact also require the logical connectives, variable symbols and the meta-brackets).

While providing a syntax theory for the logical language is beyond the scope of this thesis, we believe that this would be a worthwhile task. As noted above, a syntax theory could (for example) be used to define negation or other syntactical properties simpler than using the measures we employ here (i.e. the *neg* function) or to construct the semantic paradoxes without having to explicitly introduce them.

Furthermore, a syntax theory may allow additional reasoning about the "near-syntactical" properties $Alike_T$ and $Patho$. Both properties are by definition propagated by syntactical connectives such as $\vee$, $\neg$ or $\forall$. However, it is (at the moment) not possible to show that a formula is not pathological or that two statements are not alike according to Intersubstitutivity. A syntax theory is necessary for the syntactical arguments such reasoning would require.

6.2.3 Constructing an Automated ASPIC-END Solver

As noted in 1.1, an argumentation framework that represents the theories of truth and their properties must be infinite in size. However, this posed a problem for the analysis of this thesis: the complete assessment of all extensions and the search for admissible one may not be feasible in this context. While we were still able to provide some relevant insights into the challenges of philosophy of logic, the lack of formal analysis is still lamentable. For similar argumentation mediums such as the axiomatized theory of mathematics, while general analysis is again not possible, there exist tools for the assessment of specific reasoning tasks such as proof checking or assisted proof generation. Unfortunately, ASPIC-END is a relatively new approach to structured argumentation theory and an automated solver has not been constructed.

However, while the overall search for all admissible extensions may not be realistically possible, we believe that – similar to the case for mathematics – there are reasoning tasks where automated solving would be helpful. As mentioned in chapter 1, one motivation for this thesis is the work of Christoph Benzmüller, who used computational methods to show that a philosophical God-proof by Gödel was in fact consistent with its underlying assumptions ([BP14]), essentially a form of *automated proof checking*. We believe there is merit in producing an automated solver for ASPIC-END that is capable of preforming this and other tasks that are limited in scope.

Beyond automated proof checking, another interesting but feasible task is the search for arguments, i.e. identifying all arguments that conclude a specified formula. This task is useful for determining whether this goal can be believed by any extension, or if it is completely unfounded. Furthermore, this search is an important step for several other reasoning tasks in argumentation theory. It can (for example) be used to find attacks against a given argument (i.e. by determining if there are any arguments for negated rule applications or intermediary steps).

An important challenge for an automated theorem solver of philosophy of logic would be the rule schemes. While there have been attempts at providing a automated solver for other structured argumentation formalisms, none had to contend with the infinite instances produced by rule schemes. Any automated solver based on this thesis will have to provide some solution to this issue. While a full instantiation would be a trivial solution, this might (again) not be feasible. It may be necessary to either determine only "relevant" instances of the rules to fully instantiate or implement a "schematic" form of reasoning that can use the rule schemes in their uninstantiated form.

6.2.4 Paraconsistency as a Problematic Theory of Truth

Implementing a paraconsistent logic, such as "true" dialetheism, is one of the challenges that we have not managed to find an answer for in this thesis. As mentioned in 4.3.1, the problem with it lies in the fact that paraconsistent theories may admit some inconsistencies (i.e. some formulae φ and $\neg\varphi$) without devolving into triviality. However, any inconsistency in an extension of ASPIC-END leads to that theory being inadmissible.[5]

Furthermore, there is the question how to represent inconsistency in the argumentation theory. It is easy to see that $\bot$ represents triviality, since the rules of $\mathcal{R}_{ce}$ force it to entail every formula. On the other hand, inconsistency might be easiest to represent using the *Consistent* meta-predicate, although this does not solve the underlying problem of inadmissibility.

Another challenge of paraconsistency lies in the fact that it rejects the principle of explosion, making several of our "standard" arguments that rely on it or proof by contradiction difficult to use. This includes the argument constructed above for how an extension that adopts one theory might be able to reject any other theory that is also adoptable (i.e. I_1 to I_8), making it possible that a theory that adopts paraconsistency would have no counterargument to adopting any other theory as well. Again this leads to contradictions, such as accepting and rejecting the principle of explosion, and makes the extension inadmissible.

Most devastatingly, paraconsistency (or rather, dialetheism) clashes with the definition of rebuttals in ASPIC-END. In paraconsistent logic, the results φ and $\neg\varphi$ need not be mutually exclusive and should therefore not necessarily constitute a rebuttal. However, the only solution to this can be found in amending the definition of rebuttal. Since this solution has far-reaching and complex consequences, we have decided against attempting it as part of this thesis.

Ultimately, the problems paraconsistency pose can only be solved by changing the fundamental definitions of ASPIC-END, since the current definition of admissibility is based on the fact that inconsistencies should be avoided at all costs, which is not the case for paraconsistent logics.

6.2.5 Realizing Meta-Level Reasoning

In 5.3.3, we have found that some of Field's arguments within the paracomplete approach employ *meta-level reasoning* (i.e. reasoning about whether certain formulae can be theorems, instead of directly reasoning towards them as conclusions). Elevating reasoning to a meta-level allowed reasoning without any direct occurrences of the *True* predicate and avoiding pathological statements, making reasoning in paracomplete logic possible.

However, we have found that while implementing meta-level reasoning based on the rules of ASPIC-END (representing rules of inference or axioms of philosophy of logic) is

[5] In truth, there is a caveat for this statement: in an extension only of intuitively strict rules and without the principle of explosion, an inconsistency cannot form a rebuttal and does not lead to $\bot$ (i.e. infer everything). Although this means that there could technically be an admissible inconsistent extension, this is not relevant for our thesis, since the Consistency Argument utilizes defeasible rules.

straightforward to do, the same can not be said for the meta-rules of natural deduction (i.e. reasoning by cases, proof by contradiction and →-introduction). In 5.3.3, we have presented workaround solutions for two instances where these inferences were used within meta-level reasoning: One instance of this reasoning is the fact that Field saw $Theorem(\langle\neg C \wedge \neg\neg C\rangle)$ as equivalent to $Theorem(\langle C \wedge \neg C\rangle)$. On the object level, a proof can be constructed using double negation elimination and reasoning by cases. However, the latter (a meta-rule) has not been implemented on the meta-level. Another instance is the fact that truth-preservation must be able to replace the universal quantification of the antecedents within an implication with the individual antecedents it quantifies. This is possible on the object level using →-Introduction, which is not implemented on the meta-level.

We believe that the key to meta-level reasoning for these inferences is the implementation of a derivability predicate. A major commonality between the meta-inferences of natural deduction is the fact that they are based on an assumption and show what can be derived from it. For the following explanations we will assume the characterization of derivability with the ternary *Derivable* predicate, where $Derivable(X, y, t)$ denotes that y can be derived with the prerequisites listed in X in the theory t. Notably, the list of prerequisites X would require the implementation of lists (or sets) in ASPIC-END, which has not yet been done (neither as a syntactical element nor with rules) and is beyond the scope of this thesis. In the following we will use capitalized letters to refer to lists, ∈ as a predicate for membership and we will use $a \cdot B$ to denote adding a new first element a to a list B.

While we will not completely characterize the *Derivable* predicate, as imagined here it shares concepts with the *Theorem* predicate: truth theory axioms constitute derivability without prerequisites and rules "propagate" it. In fact, derivability can be used to express theorems as exactly those formulae that can be derived without assumptions. The following "rules"[6] can express some concepts for the *Derivable* predicate intuitively.

$$\begin{aligned}
Axiom(\widetilde{f}, \widetilde{t}) &\quad\rightsquigarrow\quad Derivable(\varnothing, \widetilde{f}, \widetilde{t}) \\
\forall \widetilde{a}.Ante(\widetilde{a}, \widetilde{r}) \rightarrow [Derivable(\widetilde{X}, \widetilde{a}, \widetilde{t}) \vee \widetilde{a} \in \widetilde{X}] &\quad\rightsquigarrow\quad Derivable(\widetilde{X}, conc(\widetilde{r}), \widetilde{t}) \\
Derivable(\varnothing, \langle\widehat{p}\rangle, \widetilde{t}) &\quad\rightsquigarrow\quad Theorem(\langle\widehat{p}\rangle, \widetilde{t}) \\
Derivable(\widetilde{X}, \widetilde{Y}, \widetilde{t}) &\quad\rightsquigarrow\quad Derivable(\langle\widehat{y}\rangle \cdot \widetilde{X}, \widetilde{Y}, \widetilde{t})
\end{aligned}$$

While the characterization of a *Derivable* predicate was beyond the scope of this thesis, we believe that it could be used to formalize meta-level natural deduction inferences. Notably, this application of the meta-inferences on the meta-level cannot inherently reproduce the assumption-attacks of ASPIC-END (and the consequent rejection of the meta-rules in the corresponding extensions). Instead, we will use a predicate $MetaRule(R, t)$ to state that the theory t accepts all instances of the meta-inference R.

[6] It is easy to see that what is provided here are not rules of our logical language. Instead, they are provided here as a sketch of what the actual rules could look like.

The following "rules" sketch how the meta-rules of natural deduction could be defined in meta-reasoning:

$$\begin{aligned}
\mathit{MetaRule}(\mathit{ImpIntro}, \widetilde{t}), \mathit{Derivable}(\langle\widehat{x_0}\rangle \cdot \widetilde{X}, \langle\widehat{y}\rangle, \widetilde{t}) &\quad\rightsquigarrow\quad \mathit{Derivable}(\widetilde{X}, \langle\widehat{x_0} \to \widehat{y}\rangle, \widetilde{t}) \\
\mathit{MetaRule}(\mathit{Contra}, \widetilde{t}), \mathit{Derivable}(\langle\widehat{x_0}\rangle \cdot \widetilde{X}, \langle\bot\rangle, \widetilde{t}) &\quad\rightsquigarrow\quad \mathit{Derivable}(\widetilde{X}, \langle\neg\widehat{x_0}\rangle, \widetilde{t}) \\
\mathit{Derivable}(\langle\widehat{x_1}\rangle \cdot \widetilde{X}, \langle\widehat{y}\rangle, \widetilde{t}), \mathit{Derivable}(\langle\widehat{x_2}\rangle \cdot \widetilde{X}, \langle\widehat{y}\rangle, \widetilde{t}), & \\
\mathsf{Assumable}_{\vee}(\widehat{x_1}), \mathsf{Assumable}_{\vee}(\widehat{x_2}), & \\
\mathit{MetaRule}(\mathit{Cases}, \widetilde{t}), \mathit{Derivable}(\widetilde{X}, \langle\widehat{x_1} \vee \widehat{x_2}\rangle, \widetilde{t}) &\quad\rightsquigarrow\quad \mathit{Derivable}(\widetilde{X}, \langle\widehat{y}\rangle, \widetilde{t})
\end{aligned}$$

While this is arguably only an insufficient sketch of how the meta-inferences may be defined, we nevertheless believe that meta-level reasoning will be an important tool in characterizing discussions in philosophy of logic, especially for non-classical theories of truth.

Chapter 7

Conclusion

Within this thesis, we have set out to model the reasoning of the philosophy of logic as an interesting use case of structured argumentation theory. We have found that philosophy of logic presents a variety of unique challenges to representation in ASPIC-END, most of which we managed to solve in a way that merits further consideration.

Within this thesis, we have outlined the most important building blocks – axioms and rules of inference – of the different theories of truth and represented them in ASPIC-END. The underlying infrastucture of our argumentation framework provided extensions for the different theories of truth, allowing these building blocks to be defined globally, but applied (or rejected) locally.

We have realized both self-reference and the truth-predicate, introducing meta-terms as a representation of Gödel codes. This construction was powerful enough for this thesis, but we believe it can also be used to define a syntax theory, offering even more opportunities for the future.

Representing the different theories of truth Field considered has been a useful test for this framework, showing that arguments that could be formulated in one theory were rejected by the others. Additionally, we have faithfully reproduced Field's arguments for how these theories rejected the Consistency Argument, further proving that this framework does represent philosophy of logic adequately.

Lastly, we have identified several issues where further research would be insightful. For example, representing paraconsistency is a challenge that requires a fundamental redefinition of argumentation theories, whereas the meta-level argumentation of paracompleteness can only be adequately represented with lists and a derivability predicate, both of which will need more modeling to be defined properly.

Bibliography

[BGGVdT18] Pietro Baroni, Dov Gabbay, Massimilino Giacomin, and Leendert Van der Torre. *Handbook of Formal Argumentation.* College Publications, 2018.

[BGH$^+$14] Philippe Besnard, Alejandro Garcia, Anthony Hunter, Sanjay Modgil, Henry Prakken, Guillermo Simari, and Francesca Toni. Introduction to Structured Argumentation. *Argument & Computation*, 5(1):1–4, 2014.

[BGR20] Jc Beall, Michael Glanzberg, and David Ripley. Liar Paradox. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy.* Metaphysics Research Lab, Stanford University, Fall 2020 edition, 2020.

[BP14] Christoph Benzmüller and Bruno Woltzenlogel Paleo. Automating Gödel's Ontological Proof of God's Existence with Higher-order Automated Theorem Provers. In *ECAI*, volume 263, pages 93–98, 2014.

[CA07] Martin Caminada and Leila Amgoud. On the Evaluation of Argumentation Formalisms. *Artificial Intelligence*, 171(5-6):286–310, 2007.

[CD20] Marcos Cramer and Jérémie Dauphin. A Structured Argumentation Framework for Modeling Debates in the Formal Sciences. *Journal for General Philosophy of Science*, 51(2):219–241, 2020.

[Dun95] 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.

[Fie06] Hartry Field. Truth and the Unprovability of Consistency. *Mind*, 115(459):567–606, 2006.

[Fie08] Hartry Field. *Saving Truth from Paradox.* OUP Oxford, 2008.

[Haa78] Susan Haack. *Philosophy of Logics.* Cambridge University Press, 1978.

[HH06] Volker Halbach and Leon Horsten. Axiomatizing Kripke's Theory of Truth. *The Journal of Symbolic Logic*, 71(2):677–712, 2006.

[Iem20] Rosalie Iemhoff. Intuitionism in the Philosophy of Mathematics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy.* Metaphysics Research Lab, Stanford University, Fall 2020 edition, 2020.

[Kri76] Saul Kripke. Outline of a Theory of Truth. *The Journal of Philosophy*, 72(19):690–716, 1976.

[MP14] Sanjay Modgil and Henry Prakken. The ASPIC+ Framework for Structured Argumentation: A Tutorial. *Argument & Computation*, 5(1):31–62, 2014.

[Pol87] John L Pollock. Defeasible Reasoning. *Cognitive science*, 11(4):481–518, 1987.

[Pra10] Henry Prakken. An Abstract Framework for Argumentation with Structured Arguments. *Argument & Computation*, 1(2):93–124, 2010.

[Raa22] Panu Raatikainen. Gödel's Incompleteness Theorems. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy.* Metaphysics Research Lab, Stanford University, Spring 2022 edition, 2022.

[SB21] Lionel Shapiro and Jc Beall. Curry's Paradox. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy.* Metaphysics Research Lab, Stanford University, Winter 2021 edition, 2021.

[ŠS13] Dunja Šešelja and Christian Straßer. Abstract Argumentation and Explanation Applied to Scientific Debates. *Synthese*, 190(12):2195–2217, 2013.

[Tar36] Alfred Tarski. Der Wahrheitsbegriff in den formalisierten Sprachen. *Studia philosophica*, 1, 1936.

[Tar44] Alfred Tarski. The Semantic Conception of Truth: and the Foundations of Femantics. *Philosophy and Phenomenological Research*, 4(3):341–376, 1944.

[Vel06] Daniel J Velleman. Variable Declarations in Natural Deduction. *Annals of Pure and Applied Logic*, 144(1-3):133–146, 2006.