Expressive Limitations and the Liar's Revenge: A Strict-Tolerant Solution and A Pragmatic Solution For Dialetheism

MSc Thesis *(Afstudeerscriptie)*

written by

Ho-Yin Lui
(born April 7th, 1989 in Hong Kong)

under the supervision of **Prof. Dr. Robert van Rooij**, and submitted to the Board of Examiners in partial fulfillment of the requirements for the degree of

MSc in Logic

at the *Universiteit van Amsterdam*.

Date of the Public Defense:	**Members of the Thesis Committee:**
July 1, 2019	Prof. Dr. Ronald de Wolf (chair)
	Prof. Dr. Frank Veltman
	Dr. Peter Hawke
	Prof. Dr. Robert van Rooij

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

For those who are perplexed by dilemmas, paradoxes and contradictions.

Abstract

In this thesis, I examine three different approaches to the Liar paradox and its revenge that respect the naive principles of truth. These approaches are the paracomplete gap approaches, the paraconsistent dialetheic approaches, and the strict-tolerant dialetheic approaches.

It is often argued that these approaches are either expressively incomplete or suffer from revenge paradoxes. Firstly, on these approaches, certain important semantical notions, if formalized in an obvious way, cannot get the desired interpretation. In paracomplete logics, the claim that the Liar sentence is neither true nor false, if formalized as $\neg(T\langle\lambda\rangle \vee T\langle\neg\lambda\rangle)$, does not come out true. In paraconsistent logics and strict-tolerant logics, the claim that a sentence A is just true, if formalized as $T\langle A\rangle \wedge \neg T\langle\neg A\rangle$, can still be a contradiction. Secondly, to fix these problems, one may suggest that we introduce some extra connectives to increase the expressive power of the theory in question. However, it is often argued that adding extra connectives gives rise to revenge paradoxes: given that semantical notions such as *truth-value gaps* and *just true* are expressible, they breed some liar-like paradoxes with which the theory cannot deal.

The first task of the thesis is that I argue that while the paracomplete gap approaches and the paraconsistent dialetheic approaches are plagued with revenge paradoxes, the strict-tolerant dialetheic approaches can resist the revenge arguments which makes use of the material biconditional or the semantic equivalence to represent self-reference.

The issue of expressive limitations is sometimes formulated as a problem of expressing disagreement. It has been argued that dialetheists have trouble in expressing what they disagree about to their opponents. Suppose that a dialetheist disagrees with A. Asserting $\neg A$ (or $\neg T\langle A\rangle$) cannot do the job, because $\neg A$ (or $\neg T\langle A\rangle$) is compatible with A. This is known as the *exclusion problem*.

It has been suggested that pragmatic implicatures accounts for how dialetheists communicate what they disagree about to their opponents: when a dialetheist asserts that it is not the case that A (or A is not true), his assertion will implicate that he does not accept A. It is often claimed that this suggestion does not work, because pragmatic implicatures cannot act on embedded sentences. In this thesis, we present some linguistic evidence that implicatures can arise at the subsentential level. The second task of the thesis is to develop an account of embedded implicatures. Our account explains how dialetheists communicate disagreement to their opponents through implicatures.

Acknowledgements

I would like to express my great appreciation and thanks to my supervisor Robert van Rooij. Robert's class was the main inspiration for this thesis. I still remember that Robert and my fellow students discussed the problem of expressive limitations and revenge paradoxes in class. Most of my fellow students argued that we have to ascend to the metalanguage to express what we cannot express in the object language. Some went for paraconsistent dialetheism. No one had any idea how to argue that we can communicate what we cannot directly express via implicatures. So out of curiosity, I planned to develop the idea in this thesis. I also highly appreciate Robert and his coworkers Pablo Cobreros, Paul Egré and David Ripley's work on semantic paradoxes. The solutions I developed in this thesis are heavily based on their work. Finally, I am very grateful to Robert for reading and commenting on every draft of this thesis.

I would like to thank Ronald de Wolf. I thank him for being the chair of the thesis committee, and for his comments on the first chapter of my thesis. I would also like to express my appreciation and thanks to Frank Veltman and Peter Hawke. I thank them for reading my thesis, as well as their critical comments and questions.

I would like to thank Shuai Wang for all his help. In particular, I am indebted to him for helping me to prepare the defense of this thesis.

Special thanks go to Polly Lam. I thank Polly for her generosity, kindness and all her help during my stay in the Netherlands.

I would like to extend my warmest thanks to Claire Tang. Her support made everything easier.

Finally, I would like to express my deepest thanks to my parents. Without their support, this thesis would not have been possible.

Contents

Chapter 1

Introduction

1.1 The Project of The Thesis

In this thesis, we will consider three different approaches to truth and the Liar paradox and the Liar's revenge. These approaches are the *paracomplete gap* approaches, the *paraconsistent dialetheic* approaches, and the *strict-tolerant dialetheic* approaches.

It is often argued that these approaches are either expressively incomplete or suffer from revenge paradoxes: if the theory in question can express certain important semantical notions, then such notions breed some new paradoxes by some liar-like reasoning. The first task of the thesis is to show that the strict-tolerant dialetheic approaches can deal with revenge paradoxes; whereas both the paracomplete gap approaches and the paraconsistent dialetheic approaches suffer from revenge paradoxes.

It is often suggested that since the dialetheic approaches suffer from revenge paradoxes, dialetheists cannot express certain seemingly meaningful notions such as *just true* and *just false*. If so, dialetheists do not have an exclusion-expressing device to indicate what they disagree about to their opponents: without the notion of just true, a dialetheist cannot communicate the fact that he accepts A but not $\neg A$. It is because if he asserts A, his assertion does not rule out $\neg A$. If so, when the dialetheist disagrees with $\neg A$, it seems that he cannot communicate what he has in mind by asserting A. This is called the *exclusion problem*. However, it is suggested that even if dialetheists do not have an exclusion-expressing device, they can still communicate what they disagree with their opponents through pragmatic implicatures. The second task of the thesis is to develop a formal account of implicatures.

In the rest of this chapter, we give a bit of background for the main discussion.

1.2 The Liar Paradox

1.2.1 The Liar paradox In Natural Languages

Typically, semantic paradoxes in natural languages have these characteristics:

- Any meaningful and declarative sentence can be exhaustively and exclusively characterized as true or false.
- The involved sentences seem meaningful and declarative.
- But somehow, they cannot be so characterized.

In the case of the Liar paradox, the involved sentences entail their own falsity. For instance, the Liar sentence explicitly says of itself that it is false (or says of itself that it is not true):

(1) (1) is false.

Our initial thought is that any sentence can be exhaustively and exclusively characterized as true or false. However, the Liar paradox arises when we consider whether sentences like (1) are true or false: it seems that no matter what we say, we are led to a contradiction. Suppose that (1) is true. Given what (1) says, it is false. Contradiction. Conversely, suppose that (1) is false. Then, '(1) is false' is false. But this means that (1) is true. Contradiction. However, (1) is either true or false. Either way, (1) is both true and false.

One may think that (1) is odd and insist that it is not meaningful. But why is it not meaningful? One typical reaction is to say that any self-referential sentence is meaningless. But this reply is an overreaction; because some self-referential sentences are meaningful:

(2) This sentence is in English.

(3) This sentence has five words.

On the other hand, banning self-reference is not enough to solve the Liar paradox. The following sentences are not self-referential:

(4) The next sentence is true.

(5) The last sentence is false.

But (4) is paradoxical. Suppose that (4) is true. According to what (4) says, (5) is true. But according to what (5) says, (4) is false. Contradiction. On the other hand, suppose that (4) is false. According to what (4) says, since (4) is false, (5) is false. But this means that (4) is true. Contradiction. Either way, we have a contradiction. Thus, banning self-reference does not help solving some variants of the Liar paradox.

Accordingly, it seems that the Liar argument shows that our initial thought cannot be held. This is puzzling – what shall we say about the status of sentences

like (1) and (4)? How can we exhaustively and exclusively characterize declarative sentences?

1.2.2 Constructing The Liar Sentence

Truth and Self-Reference. To formalize the Liar paradox, we first need to construct sentences like (1) and (4) in a formalized language. To do so, we need to express the concept of truth. We also need to be able to refer to sentences. Typically, truth is treated as a predicate of sentences, which we will write as T. To talk about sentences, we introduce names for sentences in the language: for any sentence A in the language, $\langle A \rangle$ is a name for A. Accordingly, for any sentence A in the language, $T\langle A \rangle$ is also a sentence in the language, where T is a predicate and $\langle A \rangle$ is a name for A. (We will discuss the formal details of how to provide names for sentences and mimic the (self)-referential character of sentences in §1.5 of this chapter.)

In what follows, $\mathcal{L}$ is a language that contains the usual logical connectives, and quantifiers and allows for self-reference; whereas $\mathcal{L}^+$ is a language that contains $\mathcal{L}$ and a newly introduced predicate T which is intended to express the concept of truth.

Naive Principles of Truth. We expect that the concept of truth obeys some principles which we call the *naive principles of truth*. The first one is Tarski's *T-schema*. According to Tarski (1936), any adequate theory of truth must satisfy the *Convention T*. That is, any adequate theory of truth must entail that:

- TS: $T\langle A \rangle \Leftrightarrow A$

for any sentence $A \in \mathcal{L}^+$. TS is known as the T-schema. Notice that given that we have $\models T\langle A \rangle \Leftrightarrow A$, for any sentence A, we should have:

- Release (Conditional Form): $\models T\langle A \rangle \Rightarrow A$
- Capture (Conditional Form): $\models A \Rightarrow T\langle A \rangle$

Parenthetical Remark. In this thesis, since we primarily focus on model-theoretic approaches to truth, we use $\models$ to indicate entailment relation.

How to understand $\Leftrightarrow$ depends on which semantics of conditionals one endorses. Tarski takes $\Leftrightarrow$ to be the material biconditional $\equiv$ in the classical logic.

Apart from the T-schema, the behavior of truth is also expected to obey the following principle:

- The Transparent Truth Principle (TT): A is intersubstitutable for $T\langle A \rangle$ in extensional contexts. That is, A and $T\langle A \rangle$ are inter-derivable:$T\langle A \rangle \dashv\models A$.

This amounts to the following rules:

- Release (Rule Form): $T\langle A \rangle \models A$
- Capture (Rule Form): $A \models T\langle A \rangle$

Finally, the following principle, which we will call the *intersubstitutivity principle*, seems plausible:

- Intersubstitutivity Principle (IP): For any sentence A, B and C, if B and C are alike, except where B has A, and C has $T\langle A\rangle$, then $B \eqslantless\models C$.

All of these principle are intuitively plausible, despite the fact that some theories of truth reject at least one of them. At the very least, we would not be skeptical of these principles, if there were no Liar paradox.

Parenthetical Remark. Some philosophers such as Beall (2009) endorse a similar but different formulation of intersubstitutivity principle: Let B be any sentence in which A occurs. Then the result of substituting $T\langle A\rangle$ for any occurrence of A in B has the same semantic value as B.

The Liar Sentence. In natural languages, the Liar sentence says of itself that it is not true. It is typical to formalize the Liar sentence by stipulating that:

$$\lambda \eqslantless\models \neg T\langle\lambda\rangle$$

That is, λ is a sentence such that it is intersubstitutable for the claim that it is not true, that is, $\neg T\langle\lambda\rangle$.

Instead of $\eqslantless\models$, the Liar sentence is also often formalized by the use of the material biconditional: $\lambda \equiv \neg T\langle\lambda\rangle$ (or via any $\Leftrightarrow$ you like).

Strictly speaking, the sentence $\neg T\langle\lambda\rangle$ is not self-referential. What $\neg T\langle\lambda\rangle$ refers to is another sentence λ. Heck (2007) thereby proposes the following stipulation:

$$\langle\lambda\rangle = \langle\neg T(\lambda)\rangle$$

Then, we can say that $\neg T\langle\lambda\rangle$ does refer to itself and says of itself that it is not true; because the term $\langle\lambda\rangle$ in $\neg T\langle\lambda\rangle$ is identical to the names of $\neg T\langle\lambda\rangle$, that is, $\langle\neg T(\lambda)\rangle$. That being said, to represent the self-referential character of the Liar sentence, it is customary to require that $\lambda \eqslantless\models \neg T\langle\lambda\rangle$ or $\lambda \equiv \neg T\langle\lambda\rangle$. In this thesis, we primarily follow the custom, unless otherwise stated.

1.2.3 Formal Paradox

Classical Laws. Suppose that we have a formal language $\mathcal{L}^+$ that contains a truth predicate T and allows for self-reference so that the Liar sentence can be represented via $\lambda \eqslantless\models \neg T\langle\lambda\rangle$. We also suppose that the semantics of $\mathcal{L}^+$ is closed under the T-schema or TT. Moreover, the semantics of $\mathcal{L}^+$ obeys the following laws:

- Law of Excluded Middle (LEM): $\models A \vee \neg A$
- *Ex Contradictione Quodlibet* (ECQ): $A \wedge \neg A \models B$ (Also known as *explosion*)
- Reasoning by Cases: If $A \models C$ and $B \models C$, then $A \vee B \models C$

In addition, the entailment relation $\models$ obeys the following principle:

- Validity-Preservation (VP): if $\models A$, and $A \models B$, then $\models B$

Formalizing the Liar Argument. Then we can show that the sentence λ implies a contradiction, and even triviality – everything is true.

Theorem 1 (The Liar Paradox)**.** Let **T** be a theory of truth that is closed under TT (or the T-schema). Let λ be equivalent to $\neg T\langle\lambda\rangle$ in **T**. Let $v_{\mathcal{M}^+}$ be a classical valuation for **T**. Then, for any $B \in \mathcal{L}^+$, $\mathbf{T} \models B$.

Proof. We apply $v_{\mathcal{M}^+}$ to λ:

i. Showing that given that $v_{\mathcal{M}^+}(\lambda) = 1$, it follows that $v_{\mathcal{M}^+}(\lambda \wedge \neg\lambda) = 1$:

   Suppose that $v_{\mathcal{M}^+}(\lambda) = 1$. Since λ is equivalent to $\neg T\langle\lambda\rangle$, it follows that $v_{\mathcal{M}^+}(\neg T\langle\lambda\rangle) = 1$. By Capture and contraposition, we have $v_{\mathcal{M}^+}(\neg\lambda) = 1$. Hence, we have: $v_{\mathcal{M}^+}(\lambda \wedge \neg\lambda) = 1$.

ii. Showing that given that $v_{\mathcal{M}^+}(\neg\lambda) = 1$, it follows that $v_{\mathcal{M}^+}(\lambda \wedge \neg\lambda) = 1$:

   Suppose that $v_{\mathcal{M}^+}(\neg\lambda) = 1$. By Release and contraposition, it follows that $v_{\mathcal{M}^+}(\neg T\langle\lambda\rangle) = 1$. By the definition of λ, we have $v_{\mathcal{M}^+}(\lambda) = 1$. Hence, we have: $v_{\mathcal{M}^+}(\lambda \wedge \neg\lambda) = 1$.

iii. $\lambda \vee \neg\lambda \models \lambda \wedge \lambda$. That is, if $v_{\mathcal{M}^+}(\lambda \vee \neg\lambda) = 1$, then $v_{\mathcal{M}^+}(\lambda \wedge \neg\lambda) = 1$.
(i, ii: Reasoning By Cases)

iv. $v_{\mathcal{M}^+}(\lambda \vee \neg\lambda) = 1$. (LEM)

v. $v_{\mathcal{M}^+}(\lambda \wedge \neg\lambda) = 1$. (iii, iv: VP)

vi. $\lambda \wedge \neg\lambda \models B$. That is, if $v_{\mathcal{M}^+}(\lambda \wedge \neg\lambda) = 1$, then $v_{\mathcal{M}^+}(B) = 1$ for any sentence $B \in \mathcal{L}^+$. (ECQ)

vii. For any sentence $B \in \mathcal{L}^+$, $v_{\mathcal{M}^+}(B) = 1$. (v vi: VP)

□

Now the above argument shows that $\mathcal{L}^+$ has its own paradox:

- if $\mathcal{L}^+$ is expressive enough to express truth and allows for self-reference, then $\mathcal{L}^+$ is a trivial language.

Parenthetical Remark. The thesis that everything is true is known as trivialism. While trivialism is usually assumed to be unacceptable in the literature, it has been defended by Kabay (2008). Nevertheless, in what follows, we follow the custom and hold that a formal language should not be trivial.

Two Projects. The Liar phenomenon occurs in natural languages, and in formal languages as well. The natural paradox and the formal paradox give rise to two distinct but interrelated projects:

- The Semantic Characterization Project: explain how we can exhaustively and exclusively characterize all meaningful and declarative sentences in natural languages.
- The Non-Triviality Project: show how the formal language in question is expressive enough to express truth (and some related semantical notions) and allows for self-reference, while, at the same time, is non-trivial.

1.3 Theories of Transparant Truth

In this thesis, we primarily focus on three different approaches to truth and the Liar paradox. Those are the paracomplete gap approaches, the paraconsistent dialetheic approaches, and the strict-tolerant dialetheic approaches. The main reason for focusing on such approaches is that these various approaches attempt to interpret the predicate T in the way that naive principles of truth are obeyed. The following are some brief descriptions of these approaches:

The Paracomplete Gap Approaches. Some philosophers (e.g. van Fraassen 1968; Kripke 1975) have suggested what the Liar paradox shows us is that the Liar sentence is neither true nor false. On this approach, Strong Kleene logic K_3 is the most commonly used logic. It is a three-valued logic that allows sentences to take an intermediate value $\frac{1}{2}$. The value $\frac{1}{2}$ can be construed as representing a truth-value gap. Moreover, the most important property of K_3 is that the LEM is invalid in this logic.

Kripke (1975) shows us how to extend K_3 to a logic that is closed under TT. We call this logic K_3TT. One of the major problems of K_3TT is that it cannot express its own solution to the Liar paradox. Specifically, the claim that the Liar sentence λ is neither true nor false, formalized as $\neg(T\langle\lambda\rangle \vee T\langle\neg\lambda\rangle)$, does not come out true in K_3TT. Moreover, K_3TT has another serious problem: K_3TT does not have a reasonable conditional. In K_3TT, $A \supset A$, $A \equiv A$ and the T-schema does not hold.

To fix these problems, Field (2003, 2007, 2008) introduces a new conditional, which is intended to model our ordinary understanding of conditional reasoning and validate the T-schema. Field also uses the new conditional to define a determinacy operator D to characterize paradoxical sentences as defective.

The Paraconsistent Dialetheic Approaches. Dialetheists hold that the Liar sentence is both true and false. They (e.g., Priest 2006, Beall, 2009) usually make use of the Logic of Paradox LP. Formally, LP is a three-valued logic such that the intermediate value $\frac{1}{2}$ can be construed as representing a truth value glut. The key feature of LP is that the ECQ fails, that is, $A, \neg A \models B$ does not hold in LP. The main reason for employing LP is that LP, unlike classical logic, admits contradictions without being trivialized.

Parenthetical Remark. In LP, $\supset$-modus ponens is not valid: $A, A \supset B \not\models^{LP} B$. Paraconsistent theorists usually introduce a new conditional $\rightarrow$ to play the role of conditional reasoning.

Priest (2006) extends LP to a logic such that the T-schema holds but TT does not. In particular, Priest's logic rejects $\neg A \models \neg T\langle A\rangle$ for any $A \in \mathcal{L}^+$. Notice that the T-schema, which Priest's theory accepts, gives us $\models T\langle A\rangle \leftrightarrow A$ and thus $\models T\langle A\rangle \rightarrow A$. But it is illegitimate to infer that $\neg A \models \neg T\langle A\rangle$, as Priest rejects modus tollens (i.e., $A \rightarrow B, \neg B \models\!\!\rightarrow \neg A$). Similarly, Priest (2006) also rejects $\neg A \rightarrow \neg T\langle A\rangle$. At the same time, Priest rejects contraposition, that is, $A \rightarrow B \models \neg B \rightarrow \neg A$. Hence, one cannot infer from $T\langle A\rangle \rightarrow A$ to $\neg A \rightarrow \neg T\langle A\rangle$.

Our discussion will center on a more simple logic called $LPTT$ developed by Beall (2009). $LPTT$ has a transparent truth predicate and the T-schema holds in $LPTT$.

The Strict-Tolerant Dialetheic Approaches. Paraconsistent logic is not the only option for dialetheists. Cobreros, Egré, Ripley and van Rooij (2014) recently develop a logic of truth called $STTT$. This logic is just like $LPTT$, except that it makes use of a non-transitive notion of consequence. Specifically, strict-tolerant consequence goes from a strictly true set of premises (i.e., every premise takes the value 1) to a tolerantly true conclusion (i.e., some conclusions in the conclusion set do not take the value 0). On these approaches, the principle of validity-preservation (VP) (i.e., if $\models A$, and $A \models B$, then $\models B$) does not hold. So even if every premise and every step in the Liar argument is valid, the conclusion (i.e., an arbitrary sentence) is not guaranteed to be valid.

1.4 Expressive Limitations and Revenge Paradoxes

It is often argued that any theory of truth is either expressively incomplete, or suffers from revenge paradoxes.

The first horn of the dilemma is that certain important semantical notions cannot get the desired interpretation, if such notions are formalized in terms of the truth predicate T. In paracomplete gap theories, the claim that the Liar sentence is neither true nor false cannot be true, if the claim is formalized as $\neg(T\langle\lambda\rangle \vee T\langle\neg\lambda\rangle)$. Thus, paracomplete gap theories cannot express the notion of truth-value gaps in terms of the truth predicate T. On the dialetheic approaches, while certain sentences are taken as both true and false, most sentences are taken as non-glutty. Those non-glutty sentences are either characterized as being *just true* or *just false*. However, in paraconsistent gap theories and strict-tolerant dialetheic theories, the claim that a sentence A is just true can still be a contradiction, if the claim is formalized as $T\langle A\rangle \wedge \neg T\langle\neg A\rangle$. Thus, paraconsistent dialetheic theories and strict-tolerant dialetheic theories cannot express the notion of just true in terms of the truth predicate T.

Can we increase the expressive power of such theories by adding extra connectives? The second horn of the dilemma is that if such theories are augmented with extra connectives so that certain important semantical notions can be expressed, then it appears that the augmented theories are subject to some new paradoxes – revenge paradoxes. For paraconsistent gap theories, if the notion of gaps can be expressed, then we can form a Strengthened Liar sentence like this:

(6) (6) is gappy or false.

Then, we can generate a new paradox by some familiar reasoning. Suppose that (6) is true. Then according to what it says, it is gappy or false. Thus, (6) is not true. Contradiction. On the other hand, suppose that (6) is not true. This means that either it is gappy or false. But according to what it says, it means that (6) is true. Contradiction. Either way, we have a contradiction.

For dialetheic theories, if we can express semantical notions like 'just true' and 'just false' , then we can form a Strengthened Liar sentence like this:

(7) (7) is just false.

Then we can argue as follows. Suppose that (7) is just true. Then according to what it says, it is just false. Suppose that (7) is just false. Then according to what it says, it is just true. Suppose that (7) is glutty. Then (7) is not just false. But (7) says of itself that it is just false. Hence, in any case, (7) cannot be exhaustively and exclusively characterized as just true, glutty, or just false.

Of course, whether a formal theory of truth is subject to revenge paradoxes depends on the formal details. In particular, it depends on the formal details of the theory and the formal representation of the paradoxes. But for the moment, the above informal arguments seem to show that both the gap approaches and the dialetheic approaches cannot escape the paradox. Accordingly, it seems that both the gap approaches and the dialetheic approaches must face a dilemma: such approaches are either expressively incomplete, or subject to revenge paradoxes. This issue is the main concern of this thesis.

1.5 Naming Sentences and Self-Reference

In this section, we discuss various ways to provide a name $\langle A \rangle$ for a sentence $A \in \mathcal{L}^+$. Specifically, we can introduce a name-forming device, or an auxiliary function, or represent sentences by numbers via Gödel's coding system. We will also discuss how to represent the self-referential character of sentences via these formal apparatus.

1.5.1 Name-Forming Device

In $\mathcal{L}^+$, we introduce a name-forming device $\langle\ \rangle$ such that $\langle A \rangle$ is a singular term and $\langle A \rangle$ is required to denote A in every model. This strategy was originally used by Kremer (1988).

We begin by defining $\mathcal{L}^+$-terms. There are three different kinds of $\mathcal{L}^+$-terms. They are individual variables, individual constants and $\langle A \rangle$, where A is a sentence of $\mathcal{L}^+$. In other words, if A is a sentence of $\mathcal{L}^+$, then $\langle A \rangle$ is a term of $\mathcal{L}^+$. In addition, if $\langle A \rangle$ and $\langle B \rangle$ are the same term, then A and B are the same sentence. Then, we proceed as usual. Firstly, we define the notion of atomic sentences: an expression of $\mathcal{L}^+$ is an atomic sentence iff it consists of a predicate of degree n followed by n terms. Then, we define the notion of sentences by recursion.

We must also specify how to interpret $\langle A \rangle$. We interpret a formal language by models. A model is a structure $\langle \mathcal{D}, \mathcal{I} \rangle$ such that the domain $\mathcal{D}$ is non-empty set of objects and $\mathcal{I}$ assigns to each individual constant (or names) an object of $\mathcal{D}$ and specifies the extension of the predicate symbols. Now, we require each model to include all sentence of $\mathcal{L}^+$ in its domain. Moreover, we require that for any sentence A, $\langle A \rangle$ denotes A in every model. That is, we require that for every sentence A of $\mathcal{L}^+$, $\mathcal{I}(\langle A \rangle) = A$.

Notice that while we can form sentences like $T\langle A \rangle$, we cannot directly form self-referential sentences. It is because no sentence A can contain $\langle A \rangle$. However, we can form self-referential sentences by using the identity predicate $=$. For example, the Liar sentence will be represented by a sentence $\neg Tl$ such that the term l is

identical to the term $\langle \neg Tl \rangle$ (i.e., $l = \langle \neg Tl \rangle$). Suppose that the identity predicate $=$ is interpreted classically. Then, we have:

- $v_{\mathcal{M}^+}(a = b) = 1$ if $\mathcal{I}(a) = \mathcal{I}(b)$; otherwise $v_{\mathcal{M}^+}(a = b) = 0$.

Then, if $l = \langle \neg Tl \rangle$ is true in a model, then l and $\langle \neg Tl \rangle$ denote the same object: the sentence $\neg Tl$. That is, we have $\mathcal{I}(l) = \mathcal{I}(\langle \neg Tl \rangle) = \neg Tl$ in such a model. If so, we can think of $\neg Tl$ as a Liar sentence, because $\neg Tl$ is a sentence such that the constant symbol l denotes $\neg Tl$.

1.5.2 Auxiliary Function

Another approach is to introduce an auxiliary function that maps names to sentences. This approach was originally used by Barwise and Etchemendy (1987).

On this approach, we proceed at the level of interpretation. Firstly, we require that the domain of every model includes all sentence of $\mathcal{L}^+$. Secondly, we divide our constant symbols into two denumerable sets. One is the set of ordinary names. The interpretation function $\mathcal{I}$ assigns to each ordinary name an object of $\mathcal{D}$. Another is the set of distinguished names. Thirdly, we fix a 1-1 function τ from distinguished names onto formulas of $\mathcal{L}^+$. That is, we require that a distinguished name n denote the formula $\tau(n)$ in every model.

It is very easy to form a self-referential sentence via the τ-function. For example, suppose that $\tau(n) = \neg Tn$. This means that n denotes $\neg Tn$. We can say that the sentence $\neg Tn$ mimics the behavior of the Liar sentence in natural languages; since $\neg Tn$ is a sentence whose constant symbol n refers to the very sentence $\neg Tn$.

For convenience, we can define a inverse function $\langle A \rangle$ such that if $\tau(n) = A$, then $\langle A \rangle = n$. We can do so, because τ is a 1-1 function. Thus, given that $\tau(n) = A$, then n and $\langle A \rangle$ are identical. This allows us to use $\langle A \rangle$ as a term to refer to the sentence A. Accordingly, if we stipulate that $\tau(\langle \lambda \rangle) = \neg T\langle \lambda \rangle$, we can think of $\neg T\langle \lambda \rangle$ as a Liar sentence which says of itself that it is not true; since the term $\langle \lambda \rangle$ denotes the sentence $\neg T\langle \lambda \rangle$.

1.5.3 Arithmetic

Gödel Numbering. We can use the language of arithmetic to talk about sentences in the language. To do so, we begin by arbitrarily associating a natural number to each primitive symbol of the language. For illustration, suppose that our language is the standard language of arithmetic, augmented with the truth predicate T. Call the language $\mathcal{L}^+$.

$\mathcal{L}^+$ contains the usual logical symbols. That is, it has connectives, quantifiers, identity and brackets, : $\neg$, $\wedge$, $\vee$, $\supset$, $\equiv$, $\forall$, $\exists$, $=$, (,). Moreover, $\mathcal{L}^+$ contains symbols for zero and for the successor, addition and multiplication functions: 0, $\mathtt{S}$, $+$, $\times$. In addition, $\mathcal{L}^+$ contains countably infinite variables x, y, ... and the truth predicate T . We may assign such primitive symbols a natural number as follows.

$\neg$	$\wedge$	$\vee$	$\supset$	$\equiv$	$\forall$	$\exists$	$=$	$($	$)$	$\mathtt{0}$	$\mathtt{S}$	$+$	$\times$	T	x	y	...
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	...

Secondly, we define a scheme for assigning a unique number to expressions in $\mathcal{L}^+$ in terms of the basic codes. One way to do this is based on unique-prime-factorization. This method was originally used by Gödel (1991). According to the fundamental theorem of arithmetic, any natural number greater than 1 can be represented as a unique product of prime numbers, up to ordering of the factors. Thus, in principle, any expression e can be uniquely coded as the product of prime powers.

Notice that any expression e is a finite sequence of $k+1$ primitive symbols, $\langle s_0, s_1, ..., s_k\rangle$. So we can think of e as a finite sequence of basic codes $\langle c_0, c_1, ...c_k\rangle$. Then, let π_k be the $(k+1)$-th prime number. Let c_k be the basic code number of s_k. Then, e can be uniquely coded as:

$$\pi_0^{c_0} \times \pi_1^{c_1} \times ... \times \pi_k^{c_k}$$

Call this the Gödel number of the expression e. For example, the basic code of $\forall x(x = x)$ is the sequence $\langle 6, 16, 9, 16, 8, 16, 10\rangle$. Hence, the Gödel number of $\forall x(x = x)$ is

$$2^6 \times 3^{16} \times 5^9 \times 7^{16} \times 11^8 \times 13^{16} \times 17^{10}$$

We can also decode a Gödel number into an expression. To do so, we represent the Gödel number as a product of prime powers, up to the ordering of the factors. Note that the powers can be written as a sequence of basic codes $\langle c_0, c_1, c_2, ..., c_k\rangle$. Then we can recover the expression according to the table of the basic codes.

If we want to introduce abbreviatory symbols, we take the Gödel number of such symbols to be the Gödel number of the unabbreviated version.

For convenience, it is usually stipulated that the Gödel number of an expression e is abbreviated as $\langle e\rangle$. So if A is a sentence of $\mathcal{L}^+$, $\langle A\rangle$ is a shorthand for the standard numeral for the Gödel number of A. Notice that the standard numeral is a shorthand for expressions with $\mathtt{S}$s applied to $\mathtt{0}$. So $\langle A\rangle$ is a shorthand for the term that refers to the Gödel number of A. As a heuristics, we can think of $\langle A\rangle$ as representing the sentence A . Accordingly, we can express that A is true by $T\langle A\rangle$.

The Liar sentence and Gödel's Diagonalization Lemma. We can prove that for any Ax with one free variable, there exists a sentence D such that D is equivalent to $A\langle D\rangle$. This result is known the *Diagnonalization Lemma*. If we apply the lemma to a formula $\neg Tx$, then we have a sentence λ which is equivalent to $\neg T\langle\lambda\rangle$. As previously noted, since λ and $\neg T\langle\lambda\rangle$ are different sentences, $\neg T\langle\lambda\rangle$ is not genuinely self-referential. But as a heuristics, we can think of λ as a Liar sentence.

To prove the lemma, we make use of a particular substitution operation. Let Ax be a formula with one free variable. The substitution operation, called *diagonal-*

ization, is an operation such that Ax is substituted by the Gödel number of Ax. So given that Ax is a formula with one free variable, diagonalization gives us $A\langle A\rangle$.

We can define a diagonalization function $diag$ such that when applied to a number n which is the Gödel number of Ax, it yields the Gödel number of the diagonalization of Ax, that is, $A\langle A\rangle$. Notice that the diagonalization function is recursive: we can decode n to get Ax. Then, we form a sentence $A\langle A\rangle$ and calculate its Gödel number. In addition, the diagonalization function $diag$ is representable in Peano arithmetic **PA**. That is, if a number m and a number n are such that $diag(m) = n$, then there is a formula $Diag\ mn$ such that $\mathbf{PA} \models Diag\ mn$; and if a number m and a number n are such that $diag(m) \neq n$, then $\mathbf{PA} \models \neg Diag\ mn$.

Now we are going to prove the Diagnonalization Lemma:

Lemma 2 (Diagonalization Lemma)**.** Let $\mathbf{T}$ be any theory in which the diagonalization function $diag$ is representable. Then, for any Ax with one free variable, there is a sentence D such that $\mathbf{T} \models D \equiv A\langle D\rangle$.

Proof. Firstly, we construct a sentence B such that $By =_{df} \forall y(Diag\ yz \supset Az)$.

Secondly, we form a sentence D, which is the diagonalization of B. By the definition of D, we have:

a. $\mathbf{T} \models D \equiv \forall z(Diag\ \langle B\rangle z \supset Az)$

Since D is the diagonalization of B, we have $diag(\langle B\rangle) = \langle D\rangle$. And because $Diag$ represents the diagonalization function $diag$, we have:

b. $\mathbf{T} \models \forall z((Diag\ \langle B\rangle z) \equiv (z = \langle D\rangle))$

By (b) and (a), we have:

c. $\mathbf{T} \models D \equiv \forall z(z = \langle D\rangle \supset Az)$.

That is, $\mathbf{T} \models D \equiv A\langle D\rangle$. □

The Diagonalization Lemma ensures that there is a sentence λ such that if $\mathbf{T}$ can represent the diagonalization function $diag$, then $\mathbf{T} \models \lambda \equiv \neg T\langle\lambda\rangle$.

Tarski's Undefinability Theorem. Recall that Tarski requires that any adequate theory of truth must entail all instances of the form $T\langle A\rangle \equiv A$ (if the conditional is material). However, Tarski shows that no consistent theory, closed under classical logic, has as its consequence all instances of the T-schema. This result is known as the *Undefinability Theorem*.

Theorem 3 (Undefinability Theorem)**.** Let $\mathbf{T}$ be any consistent theory that contains **PA**. Then, $\mathbf{T} \nvDash T\langle A\rangle \equiv A$ for any $A \in \mathcal{L}^+$.

Proof. Suppose for reductio that $\mathbf{T} \models T\langle A\rangle \equiv A$ for any $A \in \mathcal{L}$. By the Diagonalization Lemma, there is a sentence λ such that $\mathbf{T} \models \lambda \equiv \neg T\langle\lambda\rangle$. But by our initial supposition, we also have that $\mathbf{T} \models T\langle\lambda\rangle \equiv \lambda$.

Hence, $\mathbf{T} \models T\langle\lambda\rangle \equiv \neg T\langle\lambda\rangle$. This means that $\mathbf{T}$ is inconsistent, contradicting our initial assumption that $\mathbf{T}$ is consistent. □

1.6 Synopsis of The Thesis

The thesis is organized as follows. Chapter 2 serves as a backdrop for the main discussion. In chapter 2, we discuss Tarski's hierarchy of languages. In Tarski's theory, the truth predicates in a formal language are not allowed to apply to the sentences of the same language. So the Liar sentence cannot be constructed in the formal languages. We will discuss some common objections against Tarski's theory. Such objections indiciate that we should not solve the Liar paradox by imposing syntactic constraints.

After chapter 2, we will focus on the theories of truth that allow the truth predicate to apply to the sentences in the language and respect the naive principles of truth. In Chapter 3 - 4, we discuss the paracomplete gap approaches. Then, we will discuss the paraconsistent dialetheic approaches and the strict-tolerant dialetheic approaches in chapter 5.

In Chapter 3, we will discuss how Kripke (1975) constructs a logic of truth that is closed under the transparent truth principle. The is based on Strong Kleene three-valued logic K_3. It is customary to call Kripke's theory K_3TT (K_3 with Transparent Truth).

After introducing the formal details of K_3TT, we will discuss two major defects of it. Firstly, K_3TT lacks a reasonable conditional to carry out ordinary reasoning. Secondly, K_3TT cannot truthfully report the status of the Liar sentence. Adding extra connectives to increase the expressive power of K_3TT will give rise to revenge paradoxes, trivializing the augmented theory. In the final part of chapter 3, we will discuss some replies to the issue concerning expressive limitations and revenge paradoxes. In the literature, it is commonly held that we should not aim at constructing a formal semantics that reflect the 'real semantics' of natural languages. It is also commonly argued that revenge objections conflate model-independent notions with model-theoretic notions. We will argue that such replies cannot vindicate a theory of truth that suffers from expressive limitations and revenge paradoxes.

In chapter 4, we will discuss another paracomplete gap theory developed by Field (2003, 2007, 2008). We will first discuss how Field introduces a new conditional $\rightarrow$ to Strong Kleene logic which is closed under the transparent truth principle. Then, we will discuss how Field defines a determinacy operator D in terms of the conditional $\rightarrow$ and uses it to characterize paradoxical sentences. Field's determinacy operator D has two characteristics. Firstly, the value of DDA is not identical to the value of DA. Secondly, the LEM fails for the determinacy operator D.

In the second part of chapter 4, we will argue that although Field's theory can characterize paradoxical sentences by the determinacy operator D, it should still be counted as expressively incomplete. We will see that Field's theory disallows certain intuitively attractive semantical notions. Moreover, just like K_3TT, Field's theory cannot be extended to model those semantical notions. We will argue that since Field's theory disallows many intuitively attractive notions which are commonly used by ordinary speakers, it fails to illuminate how ordinary speakers characterize sentences.

In chapter 5, we will discuss two different dialetheic approaches to truth and the Liar paradox. One is paraconsistent dialetheism. Another is strict-tolerant dialetheism. In particular, we will focus on the transparent theory of truth on both approaches: $LPTT$ (Logic of Paradox with Transparent Truth) and $STTT$ (Strict-Tolerant Transparent Truth).

Both $LPTT$ and $STTT$ have some expressive limitations. For instance, the claim that A is just false, if formalized as $T\langle A\rangle \wedge \neg F\langle A\rangle$ can still be a contradiction in these theories. However, we argue that dialetheists can deal with such expressive limitations by adding a Just True operator $\mathbb{J}$. We call the augmented theories $LPTT^{\mathbb{J}}$ and $STTT^{\mathbb{J}}$. We will distinguish various procedures to represent self-reference in a formal apparatus. As a result, revenge paradoxes can be formulated by using different self-referential procedures. We will show that the augmented theory $LPTT^{\mathbb{J}}$ can resist the revenge paradoxes that make use of the material biconditional to represent self-reference. But it fails to deal with the revenge paradoxes that make use of the semantic equivalence to represent self-reference. On the other hand, $STTT^{\mathbb{J}}$ can resist the revenge via the material biconditional, as well as the revenge via the semantic equivalence.

In chapter 6, we investigate into the issue of expressive limitations by considering an objection against dialetheism. It is argued that dialetheists cannot express what they disagree about to their opponents. If a dialetheist disagrees with $\neg A$, he cannot express his disagreement by saying A (or $T\langle A\rangle$); because A is compatible with $\neg A$. As previously said, this is the exclusion problem.

We will first evaluate various proposals to deal with this problem: Priest's (2006) arrow falsum strategy, Beall's (2013) shriek rules, Berto's (2006) primitive exclusion and absolute contradiction, and Priest's (2006) pragmatic solutions (denials and implicatures). We will argue that the pragmatic solutions are most promising. The main challenge of the pragmatic solutions is from Shapiro (2004). According to Shapiro, denials and implicatures cannot act on embedded sentences. We will reply to Shapiro's challenge by presenting some linguistic evidence showing that denials and implicatures can arise at the sub-sentential level.

While we will not develop an account of embedded denials, we will develop an account of embedded implicatures based on an exact truthmaker semantics. We will argue that our account of embedded implicatures can explain how dialetheists communicate disagreement to their opponents through implicatures. Our account also explains a linguistic phenomenon called the *exhaustive interpretation of answers*.

In chapter 7, we conclude the thesis by taking stock of the previous chapters. The thesis provides two dialetheic solutions for the issue concerning the Liar paradox, expressive limitations and revenge paradoxes. One is the strict-tolerant solution. Chapter 3 - 5 can be construed as an argument for the strict-tolerant dialetheic approaches. On the one hand, the paracomplete gap theories and the paraconsistent dialetheic theories are either expressively incomplete, or trivialized by revenge paradoxes. On the other hand, the strict-tolerant theory $STTT$ can be augmented with a Just True operator $\mathbb{J}$ without being trivialized. We will evaluate the strict-tolerant solution concerning the semantic characterization project and the non-triviality project.

Another solution is the pragmatic implicatures solution developed in chapter 6. It will be suggested that paraconsistent dialetheists and strict-tolerant dialetheists can make use of the pragmatic solution. Paraconsistent dialetheists can argue that while it is impossible to express disagreement by saying that A is just true, they can still communicate that only A is accepted (i.e., $\neg A$ is not accepted) by saying A. It is because the assertion of A will implicate the fact that only A is accepted. The pragmatic solution is also compatible with the strict-tolerant solution.

Chapter 2

Tarski's Hierarchy of Languages

In this chapter, we look at Tarski's solution of the Liar paradox: Tarski's hierarchy of languages. We consider two common objections to Tarski's theory: the one concept objection and the objection from empirical Liars.

2.1 Tarski's Theory of Truth

2.1.1 Tarski on the Liar Paradox

According to Tarski's diagnosis, the Liar paradox is led by the following assumptions about a language $\mathcal{L}$:

i. $\mathcal{L}$ is *semantically closed*: $\mathcal{L}$ contains the resources for expressing facts about its own semantics. More precisely, $\mathcal{L}$ contains names for expressions; $\mathcal{L}$ has semantical terms such as 'true' that apply to sentences in $\mathcal{L}$.

ii. The ordinary laws of logic hold in $\mathcal{L}$.

iii. $\mathcal{L}$ allows for self-reference.

To get rid of the Liar paradox, we must reject at least one of them. Tarski suggests that we should reject (i). Firstly, Tarski thinks that rejecting (iii) is not sufficient to block the Liar paradox; because the paradox can be posed without self-reference. As for (ii), Tarski says:

> It would be superfluous to stress here the consequences of rejecting the assumption [(ii)], that is, of changing our logic (supposing this were possible) even in its more elementary and fundamental parts. (Tarski, 1944, p.349)

Accordingly, Tarski proposes using a language that does not contain its own truth predicate or any other semantic predicate (e.g., the satisfaction predicate).

Suppose that such a language is called $\mathcal{L}_0$. If we want to express facts and reason about the semantics of $\mathcal{L}_0$, we have to do so in another language, say $\mathcal{L}_1$. Similarly, $\mathcal{L}_1$ cannot express facts about its own semantics. To express such facts, we have to step up to another language, say $\mathcal{L}_2$. The process goes on and on. In this way, we arrive at a hierarchy of languages as a whole. Since each language in the

hierarchy does not contain its own truth predicate, the Liar sentence cannot be constructed at all. This is how Tarski's hierarchy of languages avoids the Liar paradox.

2.1.2 Tarski's Hierarchy of languages

Let's look at the formal details. We begin with $\mathcal{L}_0$. Recall that $\mathcal{L}_0$ does not contain its own truth predicate or any other semantic predicate. $\mathcal{L}_0$ is interpreted by a classical model $\mathcal{M}_0 = \langle \mathcal{D}, \mathcal{I} \rangle$. The domain $\mathcal{D}$ is a non-empty set of objects. In particular, $\mathcal{D}$ contains (codes of) every sentence of $\mathcal{L}_0$. All non-logical symbols (i.e., individual constants, function symbols, and predicate symbols) of $\mathcal{L}_0$ are interpreted by the interpretation function $\mathcal{I}$ classically. For simplicity, we suppose that $\mathcal{M}_0$ is a standard model of arithmetic: the domain $\mathcal{D}$ is the set of natural numbers $\mathbb{N} = \{0, 1, 2, ...\}$.

If we want to express facts about the semantics of $\mathcal{L}_0$, we cannot do so in the language $\mathcal{L}_0$. It is because $\mathcal{L}_0$ does not contain its own semantic predicates. Hence, we have to move to another language $\mathcal{L}_1$. The language $\mathcal{L}_1$ is called the *metalanguage* of $\mathcal{L}_0$; whereas $\mathcal{L}_0$ is called the *object language* of $\mathcal{L}_1$.

$\mathcal{L}_1$ contains a predicate T_0 which represents the truth predicate for $\mathcal{L}_0$. The predicate T_0 applies to each sentence A in $\mathcal{L}_0$. The semantics of $\mathcal{L}_1$ is just like $\mathcal{L}_0$, except that $\mathcal{M}_0$ is expanded to interpret the predicate T_0. Specifically, $\mathcal{L}_1$ is interpreted by $\mathcal{M}_1 = \langle \mathcal{M}_0, \mathcal{T}_0 \rangle$, where $\mathcal{T}_0$ is a set of (codes of) true sentences of $\mathcal{L}_0$. The interpretation of the predicate T_0 is restricted in such a way that $T_0\langle A \rangle \equiv A$ holds for all $A \in \mathcal{L}_0$. For convenience, if a sentence A is true in a model $\mathcal{M}$, we write $\mathcal{M} \models A$. Then, we have:

- $\mathcal{M}_1 \models T_0\langle A \rangle \quad \text{iff} \quad \langle A \rangle \in \mathcal{T}_0 \quad \text{iff} \quad \mathcal{M}_0 \models A.$

$\mathcal{L}_1$ does not contain its own semantic predicates. To express facts about the semantics of $\mathcal{L}_1$, we have to move to another language $\mathcal{L}_2$, the metalangauge of $\mathcal{L}_1$.

In general, we have a hierarchy of languages $\mathcal{L}_0, \mathcal{L}_1, \mathcal{L}_2, ...$ such that each language $\mathcal{L}_i$ in the hierarchy does not contain its own truth predicate T_i. Moreover, $\mathcal{L}_{i+1}$, the metalangauge of $\mathcal{L}_i$, is the language $\mathcal{L}_i$ augmented with a predicate T_i (and other corresponding semantic predicates) which is applicable to any sentence $A \in \mathcal{L}_i$. For each language $\mathcal{L}_{i+1}$, we have the T-schema $T_i\langle A \rangle \equiv A$ for all $A \in \mathcal{L}_i$. The hierarchy of languages can be summarized by the following table:

Language	Model	T-schema
$\vdots$	$\vdots$	$\vdots$
$\mathcal{L}_{i+1}$	$\mathcal{M}_{i+1} = \langle \mathcal{M}_i, \mathcal{T}_i \rangle$	$T_i\langle A \rangle \equiv A$ for $A \in \mathcal{L}_i$
$\vdots$	$\vdots$	$\vdots$
$\mathcal{L}_2$	$\mathcal{M}_2 = \langle \mathcal{M}_1, \mathcal{T}_1 \rangle$	$T_1\langle A \rangle \equiv A$ for $A \in \mathcal{L}_1$
$\mathcal{L}_1$	$\mathcal{M}_1 = \langle \mathcal{M}_0, \mathcal{T}_0 \rangle$	$T_0\langle A \rangle \equiv A$ for $A \in \mathcal{L}_0$
$\mathcal{L}_0$	$\mathcal{M}_0 = \langle \mathcal{D}, \mathcal{I} \rangle$	-

Parenthetical Remark. Kripke (1975) criticizes that Tarski's hierarchy of languages

had only been defined for finite levels. Mathematically, extending the Tarskian hierarchy of languages to transfinite levels is not a trivial task. See Halbach (1997) for how the transfinite Tarskian hierarchy of languages can be modeled.

2.1.3 Blocking The Liar Paradox

Let's see how the Liar paradox is blocked. Suppose that we want to construct the Liar sentence in one of the languages in the hierarchy, say $\mathcal{L}_i$. Notice that:

$$\lambda \equiv \neg T\langle\lambda\rangle$$

is not a well-formed formula of $\mathcal{L}_i$; since we have to specify the level of the truth predicate T. Moreover, we must choose a predicate T_b where the level b is lower than the level i. Accordingly, what we can construct is a sentence of this form:

$$\lambda \equiv \neg T_b\langle\lambda\rangle$$

where $b < i$.

Then, notice that $T_b\langle\lambda\rangle$ is not a well-formed formula in $\mathcal{L}_b$; because $\mathcal{L}_b$ does not contain its own truth predicate. Hence, it is not the case that $T_b\langle\lambda\rangle$. Thus, in $\mathcal{L}_{i+1}$ or any higher-level languages, we have $\neg T_b\langle\lambda\rangle$. Then, we can reason as follows:

1	$\lambda \equiv \neg T_b\langle\lambda\rangle$	Hypothesis
2	$T_i\langle\lambda\rangle \equiv \lambda$	T-schema $T_i\langle A\rangle \equiv A$
3	$T_i\langle\lambda\rangle \equiv \neg T_b\langle\lambda\rangle$	1, 2: Transitivity of $\equiv$
4	$\neg T_b\langle\lambda\rangle$	Fact
5	$T_i\langle\lambda\rangle$	3, 4: $\equiv$-Modus Ponens

But there is nothing paradoxical to say that λ is true in $\mathcal{L}_{i+1}$ or in any higher-level languages: the conjunction of $T_i\langle\lambda\rangle$ and $\neg T_b\langle\lambda\rangle$ is not a contradiction.

2.2 Common Objections

2.2.1 The One Concept Objection

In this section, we look at two common objections against Tarski's theory. These objections are the *one concept objection* and the *objection from empirical Liars*.

The most notorious problem of Tarski's hierarchy of languages is that whereas there seems to be only one single concept of truth in natural languages, Tarski's theory seems to suggest that there are many distinct concepts of truth relative to some level.

However, we must notice that Tarski does not intend the hierarchy of languages as explaining the word 'true' in natural languages. But in the past, some logicians and philosophers thought that Tarski's theory provides a solution to the Liar paradox in natural languages. What Tarski is interested in is constructing

formalized languages that are free from semantic paradoxes. Tarski also admits that natural languages are *semantically open*. That is, natural languages contain their own semantic concepts:

> A characteristic feature of colloquial language [...] is its universality [...] If we are to maintain this universality of everyday language in connexion with semantical investigations, we must [...] admit into the language [...] semantic expressions as 'true sentence', 'name', 'denote', etc. (Tarski, 1956, p.164)

Tarski thereby concludes that natural languages are indeed inconsistent:

> It is just this universality of everyday language which is the primary source of all semantical antinomies, like the antinomies of the liar or of heterological words. These antinomies seem to provide a proof that every language which is universal in the above sense, and for which the normal laws of logic, hold must be inconsistent. (ibid, p.164-165)

Parenthetical Remark. It is commonly argued that Tarski has offered usable formalized languages for science and mathematics; since the language of science and the language of mathematics are semantically closed. But this is simply not true. See McGee (1991) and Halbach (2011) for further discussion.

2.2.2 The Objection from Empircal Liars

Against the above objection, one may reply that the concept of truth has a hidden index, or that it is ambiguous. However, Kripke (1975) illustrates that whether or not a sentence is paradoxical does not merely depend on the syntax or the semantics of the sentence. Sometimes, empirical factors also have a role. What this means is that we cannot solve the Liar paradox by imposing syntactic restrictions or *a priori* semantical restrictions. In particular, we cannot solve the paradox by disallowing that the truth predicate applies to sentences in the same language.

Kripke's illustrates his point with the following example. Consider a situation that Dean says:

(8) All of Nixon's utterances about Watergate are true.

And Nixon says:

(9) Most of Dean says about Watergate is false.

These sentences are nothing but ordinary assertions made in a political debate. They are perfectly grammatical and are by no means meaningless.

Normally, no paradox arises out of (8) and (9). These sentences are paradoxical only in some unfavorable and usually unexpected empirical circumstances. To see this, consider the following situation: except (8), half of Dean's assertions about Watergate are true, and half of them are false. Also, except (9), all Nixon's assertions about Watergate are true.

In this case, the truth value of (8) depends on the truth value of (9):

i. If (9) is false, then (8) is false either.

ii. If (9) is true, then (9) is true as well.

Analogously, the truth value of (9) depends on the truth value of (8):

iii. If (8) is false, then (9) is true.

iv. If (8) is true, then (9) is false.

Chaining (iii) and (ii) (or (i) and (iv)) together, we have (8) is true iff (8) is false. Chaining (i) and (iii) (or (ii) and (iv)) together, we have (9) is true iff (9) is false.

However, Tarski's theory cannot consistently assign a level to the truth predicate of (8) and that of (9). To see this, suppose that the truth predicate in (8) is T_i and the truth predicate in (9) is T_j. According to Tarski's theory, any truth predicate T_n can only apply to sentences in $\mathcal{L}_k$ where k is at a lower level than n. Since the truth predicate T_i in (8) applies to Nixon's utterances, in particular, (9), it means that $i > j$. Similarly, since the truth predicate T_j in (9) applies to Dean's utterances, in particular, (8), it means that $j > i$. However, this is impossible – it cannot be the case that $i > j$ and $j > i$.

The lesson we learn from Kripke's example is that Tarski's theory is overkill – in some unexpected empirical situations, some perfectly grammatical and meaningful sentences cannot be formulated in the languages of the Tarskian hierarchy.

In summary, according to the one concept objection, it seems that we only have one unified concept of truth. However, Tarski's hierarchy of languages introduces infinitely many truth predicates merely for the sake of avoiding the Liar paradox. One may reply that while it appears that we only have one concept of truth, we do have infinitely many of truth predicates at the 'deep' level. Yet, the objection from empirical Liars shows that it is fruitless to block the Liar paradox by imposing syntactic constraints or *a priori* semantical constraints.

Chapter 3

Paracomplete Gap Theory: Kripke's Theory of Truth

In the previous chapter, we saw that Kripke's empirical Liars show that a theory of truth should not block the Liar paradox by imposing syntactic constraints. Accordingly, Kripke develops a theory of truth that allows the truth predicate applies to the sentences in the same language.

Kripke's treatment of the Liar paradox is motivated by Strawson's analysis of presuppositions. Strawson deems that if a sentence A presupposes a false sentence B, then A is in the truth-value gaps. Kripke draws inspiration from Strawson's analysis: Kripke suggests that sentences such as the Liar sentence and empirical Liars fail to make a statement, and thus in the truth-value gaps. Accordingly, a logic allowing for truth-value gaps and invalidating the Law of Excluded Middle is needed. Kripke proposes using Strong Kleene Logic K_3.

We begin this chapter by discussing Kripke's theory. In §3.1, we review some philosophical motivations for gap theory. Then, we discuss the formal details of Strong Kleene logic K_3. In §3.2, we discuss how Kripke extends K_3 to interpret the truth predicate T. The extended theory has two characteristics. Firstly, for any A in the language, $T\langle A\rangle$ and A are intersubstitutable in any extensional context. Secondly, $T\langle A\rangle \vee \neg T\langle A\rangle$ is not a logical truth. The extended theory is called K_3TT (K_3 with Transparent Truth).

Then, we turn to the problems of Kripke's theory. In §3.3, we discuss some problems concerning the material conditional in K_3TT. In §3.4, we discuss certain expressive limitations of K_3TT. In particular, the claim that the Liar sentence is neither true nor false, if formalized as $\neg(T\langle\lambda\rangle \vee T\langle\neg\lambda\rangle)$, cannot be true in K_3TT. If we increase the expressive power of K_3TT by adding some extra connectives, then some new paradoxes arise. Specifically, let $K_3TT^{\mathbb{D}}$ be K_3TT extended with a determinacy operator $\mathbb{D}$ such that $\mathbb{D}A$ takes the value 1, if A takes the value 1; otherwise, $\mathbb{D}A$ takes the value 0. It can be shown that $K_3TT^{\mathbb{D}}$ is trivial. Finally, we discuss some replies to revenge paradoxes qua objection to a theory of truth. Such replies concern the nature of model theories of truth.

3.1 Gap Theory and Strong Kleene Logic

3.1.1 Truth-Value Gaps: Philosophical Motivations

According to gap theorists, the Liar paradox teaches us that some sentences like the Liar sentence are neither true nor false. Van Fraassen (1968) motivates this view by Strawson's analysis of presuppositions. Strawson's analysis can be summarized by the following principle:

- If a sentence A presupposes B, and B is false, then A is neither true nor false.

A case in point is that (10) presupposes (11):

(10) The present King of France is bald.

(11) The present King of France exists.

Since there is no King of France now, (11) is false. Hence, according to Strawson's principle, (10) is neither true nor false. Van Fraassen applies Strawson's principle to argue that the Liar sentence is neither true nor false. According to van Fraassen, the Liar sentence presupposes the falsity (or untruth) of itself. As the Liar argument shows, such a presupposition cannot be true. Thus, by Strawson's principle, the Liar sentence is neither true nor false.

Kripke (1975) also motivates a gap approach to the Liar paradox by Strawson's analysis:

> Under the influence of Strawson, we can regard a sentence as an attempt to make a statement, express a proposition, or the like. The meaningfulness or well-formedness of the sentence lies in the fact that there are specifiable circumstances under which it has determinate truth conditions (expresses a propositions), not that it always does express a proposition. [Some] sentence[s] [...] are always *meaningful*, but under various circumstances [they] may not "make a statement", or "express a proposition". (ibid p.699 - 700)

Accordingly, a logic that allows for truth-value gaps and invalidates the Law of excluded middle (LEM) is needed. Kripke (1975) proposes using Kleene's strong three-valued logic K_3. In what follows, we focus on Kripke's proposal.

3.1.2 Strong Kleene Logic

Model-theoretically, we can think of a semantics by a structure:

$$S_L = \{\mathcal{L}, \mathcal{V}, \mathcal{D}^+, \mathcal{M}, v_{\mathcal{M}}\}$$

where:

- $\mathcal{L}$ is a formal language;
- $\mathcal{V}$ is a non-empty set of values;

- $\mathcal{D}^+ \subseteq \mathcal{V}$ is a set of designated values;
- $\mathcal{M}$ is a model that gives the interpretation of non-logical symbols in the language $\mathcal{L}$; and
- $v_{\mathcal{M}}$ is a valuation scheme.

We begin with a language $\mathcal{L}$ that does not contain a predicate T representing the notion of truth. We call $\mathcal{L}$ the base language. Strong Kleene Logic K_3 is a three-valued logic with the value 1 being designated. That is, in K_3, we have $\mathcal{V} = \{1, \frac{1}{2}, 0\}$ and $\mathcal{D}^+ = \{1\}$. As a heuristics, we can think of $\mathcal{D}^+$ as the set of values that a good sentence can have. We also have Strong Kleene models $\mathcal{M} = \langle \mathcal{D}, \mathcal{I} \rangle$ to interpret the language $\mathcal{L}$:

Definition 4 (Strong Kleene Model)**.** A Strong Kleene model $\mathcal{M}$ is a pair $\langle \mathcal{D}, \mathcal{I} \rangle$ such that:

(i) $\mathcal{D}$ is a non-empty set of objects. We call $\mathcal{D}$ the domain.

(ii) $\mathcal{I}$ is an interpretation function such that

- $\mathcal{I}$ assigns each individual constant (i.e., name) an object of $\mathcal{D}$. That is, for all $a \in Con$, $\mathcal{I}(a) \in \mathcal{D}$, where Con is a set of individual constant.
- $\mathcal{I}$ assigns to each n-ary function symbol an object of $\mathcal{D}$. That is, for all $f \in Func_n$, $\mathcal{I}(f) : \mathcal{D}^n \mapsto \mathcal{D}$, where $Func_n$ is a set of n-place function symbols for all n.
- $\mathcal{I}$ assigns a pair $\langle \mathcal{I}^+, \mathcal{I}^- \rangle$ to each all n-ary predicate P. $\mathcal{I}^+$ assigns to each n-ary predicate a set of n-tuples which is called the *extension* of P; whereas $\mathcal{I}^-$ assigns to each n-ary predicate a set of n-tuples which is called the *anti-extension* of P: That is, for all $P \in Pred_n$, $\mathcal{I}^+(P) \subseteq \mathcal{D}^n$ and $\mathcal{I}^-(P) \subseteq \mathcal{D}^n$, where $Pred_n$ is the set of n-place predicate symbols for all n.
- $\mathcal{I}$ is constrained by the following stipulation: $\mathcal{I}^+(P) \cap \mathcal{I}^-(P) = \emptyset$. This condition ensures the exclusiveness of the predicate P and its negation. So there is no object in the domain $\mathcal{D}$ falling into both the extension and the anti-extension of a predicate.

As for the valuation scheme $v_{\mathcal{M}}$, we have:

Definition 5 (K_3 Valuation Scheme)**.** For any K_3 model $\mathcal{M} = \langle \mathcal{D}, \mathcal{I} \rangle$, the K_3 valuation scheme $v_{\mathcal{M}}$ is a function that assigns each sentence A in the language to either the value 1, 0, or $\frac{1}{2}$. The K_3 valuation scheme $v_{\mathcal{M}}$ is such that, for any A, $B \in \mathcal{L}$, for any model $\mathcal{M}$,

(i) For an atomic formula $Pa_0, ..., a_n$, we have:

$$v_{\mathcal{M}}(Pa_0, ..., a_n) = \begin{cases} 1, & \text{if } \langle \mathcal{I}(a_0), ..., \mathcal{I}(a_n) \rangle \in \mathcal{I}^+(P). \\ 0, & \text{if } \langle \mathcal{I}(a_0), ..., \mathcal{I}(a_n) \rangle \in \mathcal{I}^-(P). \\ \frac{1}{2}, & \text{otherwise.} \end{cases}$$

(ii) $v_{\mathcal{M}}(A \wedge B) = min\{v_{\mathcal{M}}(A), v_{\mathcal{M}}(B)\}$

(iii) $v_{\mathcal{M}}(A \vee B) = max\{v_{\mathcal{M}}(A), v_{\mathcal{M}}(B)\}$

(iv) $v_{\mathcal{M}}(\neg A) = 1 - v_{\mathcal{M}}(A)$

(v) $v_{\mathcal{M}}(A \supset B) = max\{1 - v_{\mathcal{M}}(A), v_{\mathcal{M}}(B)\}$

(vi) $v_{\mathcal{M}}(\exists x Px) = max\{v_{\mathcal{M}}(Px[a/x]) : \text{ for all } a \in \mathcal{D}\}$

(vii) $v_{\mathcal{M}}(\forall x Px) = min\{v_{\mathcal{M}}(Px[a/x]) : \text{ for all } a \in \mathcal{D}\}$

For readability, the valuation scheme for the connectives can also be depicted as follows:

$\wedge$	1	$\frac{1}{2}$	0
1	1	$\frac{1}{2}$	0
$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	0
0	0	0	0

$\vee$	1	$\frac{1}{2}$	0
1	1	1	1
$\frac{1}{2}$	1	$\frac{1}{2}$	$\frac{1}{2}$
0	1	$\frac{1}{2}$	0

A	$\neg A$
1	0
$\frac{1}{2}$	$\frac{1}{2}$
0	1

$\supset$	1	$\frac{1}{2}$	0
1	1	$\frac{1}{2}$	0
$\frac{1}{2}$	1	$\frac{1}{2}$	$\frac{1}{2}$
0	1	1	1

As a heuristic, we can think of:

- sentences with the value 1 as true;
- sentences with the value 0 as false; and
- sentences with the value $\frac{1}{2}$ as neither true nor false.

We must keep in mind that there is a conceptual difference between sentences with the value 1 and truth. The former is a model-dependent notion; whereas the latter is not. (Similar points can be made in the cases of falsity and truth-value gap.) Moreover, it should be noticed that different philosophical approaches read the value $\frac{1}{2}$ differently. While gap theorists read the value $\frac{1}{2}$ as representing 'neither true nor false', dialetheists read it as representing 'both true and false'.

Parenthetical Remark. The above remark about how to read the value $\frac{1}{2}$ in the models should be taken carefully, if the models are augmented to deal with transparent truth. In particular, it might be misleading to say that the value $\frac{1}{2}$ of K_3TT represents 'neither true nor false', while the value $\frac{1}{2}$ in $LPTT$ or $STTT$ represents 'both true and false'. In such logics, the claim that A is neither true nor false, formalized as $\neg(T\langle A\rangle \vee T\langle \neg A\rangle)$, is equivalent to the claim that A is both true and false $T\langle A\rangle \wedge T\langle \neg A\rangle$. Such oddities relate to the expressive limitations of such logics, which we will discuss in §3.4 of this chapter and §5.4 of chapter 5.

3.2 Kripke's Theory of Truth

3.2.1 Kripke's Project

In the last section, the language $\mathcal{L}$ we considered is without a truth predicate. This language is interpreted by the base models $\mathcal{M} = \langle \mathcal{D}, \mathcal{I}\rangle$. Now we consider an expanded language $\mathcal{L}^+$ that allows for self-reference and contains a predicate T which is intended to play the role of the truth predicate.

To give the interpretations of $\mathcal{L}^+$, we extend the base models to do so. Specifically, the extended models $\mathcal{M}^+ = \langle \mathcal{D}, \mathcal{I}, \mathcal{T}\rangle$ are such that $\mathcal{T}$ is intended to interpret the

predicate T as truth. One important constraint on $\mathcal{T}$ is that the interpretation of the predicate T should obey the transparent truth principle (TT).

- The Transparent Truth Principle (TT): $T\langle A\rangle \eqvDash A$

for any $A \in \mathcal{L}^+$. TT ensures that for any $A \in \mathcal{L}^+$, $T\langle A\rangle$ and A are intersubstitutable in any extensional context.

Moreover, $T\langle\lambda\rangle \vee \neg T\langle\lambda\rangle$, an instance of the LEM, should fail in the logic. A logic where the LEM fails is not hard to come by. In fact, we have $\nvDash^{K_3} A\vee\neg A$. However, it is not a trivial matter to construct such a logic in which $T\langle\lambda\rangle \vee \neg T\langle\lambda\rangle$ fails, if we require the intersubstitutability of $T\langle A\rangle$ and A.

Since the target logic is based on K_3 and closed under TT, we call the logic K_3TT. In summary, we require that K_3TT has the following desiderata:

i. K_3TT interprets the predicate T as truth. Specifically, we should have: any $A \in \mathcal{L}^+$, $A \eqvDash^{K_3TT} T\langle A\rangle$.

ii. $\nvDash^{K_3TT} T\langle A\rangle \vee \neg T\langle A\rangle$

In what follows, we will discuss how Kripke constructs a logic with these desiderata.

Firstly, Kripke proposes seeing the predicate T as partially defined. Rather than a simple extension, the predicate T is assigned a pair of sets: the extension $\mathcal{T}^+$ and the nti-extension $\mathcal{T}^-$. The truth predicate is true of the things in $\mathcal{T}^+$, and false of the things in $\mathcal{T}^-$.

The extension $\mathcal{T}^+$ and the anti-extension $\mathcal{T}^-$ are mutually exclusive: $\mathcal{T}^+\cap\mathcal{T}^- = \emptyset$. So the predicate T cannot be true and false of the same thing. However, some objects (i.e., (codes of) sentences of $\mathcal{L}^+$) are allowed to fall neither in the extension $\mathcal{T}^+$ nor the anti-extension $\mathcal{T}^-$. In other words, there could be something that the predicate T is neither true of nor false of.

Secondly, to interpret the predicate T as truth, we need to ensure that

- Identity of Truth (IT): $T\langle A\rangle$ and A always have the same value.

Notice that IT and TT are different. IT only concerns how models assign value; it says nothing about entailment relation between $T\langle A\rangle$ and A. Thus, one should not conflate IT with TT. That being said, IT is a crucial step for obtaining TT. Generally speaking, a countermodel to an argument is defined in such a way that it must assign different values to the premises and the conclusions. Accordingly, if $T\langle A\rangle$ and A always have the same value, there is no countermodel to an argument from $T\langle A\rangle$ to A and vice versa. Thus, given that logical consequence is defined as the absence of countermodels, then IT ensures TT.

Imposing IT on models amounts to ensuring $\mathcal{T}^+$ to be the set of things of which it is true, and $\mathcal{T}^-$ to be the set of thing of which it is not true. Let $\mathcal{T}^+_{\mathcal{M}^+}$ be the set of (codes of) true sentences of $\mathcal{M}^+$, and let $\mathcal{T}^-_{\mathcal{M}^+}$ be the set of objects in $\mathcal{D}$ such that either the objects are not (codes of) sentences of $\mathcal{M}^+$, or are (codes of) false sentences of $\mathcal{M}^+$. To ensure IT, we need to ensure that $\mathcal{T}^+ = \mathcal{T}^+_{\mathcal{M}^+}$ and $\mathcal{T}^- = \mathcal{T}^-_{\mathcal{M}^+}$. It is because if this is the case, this means that given that A is a sentence of $\mathcal{L}^+$, the

predicate T is true of (false of) A iff A takes the value 1 (the value 0). In summary, the target model is as follows.

Definition 6 (K_3^+ model)**.** A K_3^+ model is a structure $\mathcal{M}^+ = \langle \mathcal{D}, \mathcal{I}, \mathcal{T} \rangle$ such that

(i) $\mathcal{D}$ is the domain.

(ii) $\mathcal{I}$ is the same as $\mathcal{I}$ in the K_3 models.

(iii) $\mathcal{T}$ is a pair $\langle \mathcal{T}^+, \mathcal{T}^- \rangle$, such that

- $\mathcal{T}^+$ assigns a set of objects to the extension of the truth predicate $\mathcal{T}^+ \subseteq \mathcal{D}$;
- $\mathcal{T}^-$ assigns a set of objects to the anti-extension of the truth predicate $\mathcal{T}^- \subseteq \mathcal{D}$.
- $\mathcal{T}^+ \cap \mathcal{T}^- = \emptyset$.
- For any $\mathcal{T}^+, \mathcal{T}^- \in \mathcal{D}$, $\langle \mathcal{T}^+, \mathcal{T}^- \rangle = \langle \mathcal{T}^+_{\mathcal{M}^+}, \mathcal{T}^-_{\mathcal{M}^+} \rangle$, where

  $\mathcal{T}^+_{\mathcal{M}^+}$ is the set of (codes of) true sentences of $\mathcal{M}^+$, and

  $\mathcal{T}^-_{\mathcal{M}^+}$ is the set of objects in $\mathcal{D}$ such that either the objects are not (codes of) sentences of $\mathcal{L}^+$, or are (codes of) false sentence of $\mathcal{M}^+$.

Logical consequence is construed as the absence of countermodels. In many-valued logics, it is customary to use the notion of designated value to define countermodels. Since we can think of the designated values as the values for good sentences, it is natural to define countermodels as follow:

- A countermodel to an argument from the premises Γ to the conclusions Δ is a model that assigns a designated value to every member of Γ and assigns a non-designated value to every member of Δ.

Call this the *principle of designated value preservation*. Since the value 1 is the only designated value, we have:

- A K_3^+ model is a countermodel to an argument from the premises Γ to the conclusions Δ iff it assigns the value 1 to every premise of Γ and the value $\frac{1}{2}$ or the value 0 to every conclusion of Δ.

Thus, K_3TT consequence amounts to:

Definition 7 (K_3TT Consequence)**.** $\Gamma \models^{K_3TT} \Delta$ iff if $v_{\mathcal{M}^+}(A) = 1$ for all $A \in \Gamma$, then $v_{\mathcal{M}^+}(B) = 1$ for some $B \in \Delta$.

3.2.2 A Hierarchy of Learning

Kripke shows us how to construct K_3^+ models. In particular, he shows that there is a pair of set $\langle \mathcal{T}^+, \mathcal{T}^- \rangle = \langle \mathcal{T}^+_{\mathcal{M}^+}, \mathcal{T}^-_{\mathcal{M}^+} \rangle$. To show this, he constructs a number of candidate pairs $\langle \mathcal{T}^+, \mathcal{T}^- \rangle$, and then finds out which one meets the condition.

The basic idea of Kripke's construction is motivated by the following picture. Suppose that we want to explain the word 'true' to someone who has yet to understand it. We can say to him that we are entitled to assert (or deny) that a

sentence is true just when one is entitled to assert (or deny) that sentence. The learner can thereby understand that when he is entitled to say that:

(12) Snow is white.

he is also entitled to say that:

(13) 'Snow is white' is true.

Moreover, he can apply the notion of truth to sentences which already contain the word 'true'. Consider the following sentence:

(14) Some sentence printed in the *New York Daily News* October 7, 1971 is true.

If one does not know whether he is entitled to assert (14), he does not know whether he is entitled to assert '(14) is true'. Now, we suppose that (12) is one of the sentences printed in the *New York Daily News* October 7, 1971. In this case, if one is entitled to assert (12), he is entitled to assert (14) as well. If so, he is also entitled to assert that (14) is true. The learner will be able to predicate truth to more and more sentences that already contain the word 'true'.

We can think of this picture as a hierarchy of learning how the concept of truth is used. Such a hierarchy continues into the transfinite. To see this, consider a sequence of sentences: (12), (13), (13) is true, '(13) is true' is true, ' "(13) is true" is true' is true, and so on. To grasp the meaning of truth, the learner has to learn to 'reflect upon' the sequence of sentences: he has to learn whether he is entitled to say that all sentences in the sequence are true. Since the sequence can be extended into the transfinite, the learning process is in principle transfinite. Nevertheless, the learning process does not go on indefinitely. The learner will reach a 'fixed point' eventually: further iterating the process does not improve how the learner applies the word 'true' to sentences. What this means is that if the learner goes through the learning process, he will finally grasp the correct meaning of truth.

3.2.3 Fixed Points

The Kripke Construction. We can formally represent the above picture as follows. Initially, the learner does not know what the predicate T is true of and false of. Thus, we have:

$$\langle \mathcal{T}_0^+, \mathcal{T}_0^- \rangle = \langle \emptyset, \emptyset \rangle$$

Call the resulting model $\mathcal{M}_0^+ = \langle \mathcal{D}, \mathcal{I}, \mathcal{T}_0 \rangle$. However, the model does not have the identity of truth property. For instance, suppose that the atomic sentence Pa takes the value 1 in this model. But $T\langle Pa \rangle$ takes the value $\frac{1}{2}$, because $\langle \mathcal{T}_0^+, \mathcal{T}_0^- \rangle = \langle \emptyset, \emptyset \rangle$.

Accordingly, the learner has to improve the understanding of the notion of truth. Suppose that A is true (false) at the α level. At the $\alpha + 1$ level, the learner should deem A to be something of which the predicate T is true (or false). This is how the learner improves the application of the word 'true'.

Formally, we suppose that $\mathcal{M}^+_\alpha = \langle \mathcal{D}, \mathcal{I}, \mathcal{T}_\alpha \rangle$ is a model of $\mathcal{L}^+$ at the level α. We also suppose that $\mathcal{T}_\alpha$ is well-defined. We can say that a model $\mathcal{M}^+_{\alpha+1}$ is an improvement of the model $\mathcal{M}^+_\alpha$, if the model $\mathcal{M}^+_{\alpha+1} = \langle \mathcal{D}, \mathcal{I}, \mathcal{T}_{\alpha+1} \rangle$ is such that $\mathcal{T}_{\alpha+1}$ is a pair $\langle \mathcal{T}^+_{\alpha+1}, \mathcal{T}^-_{\alpha+1} \rangle$ where:

- $\mathcal{T}^+_{\alpha+1}$ is the set of (codes of) true sentences of $\mathcal{M}^+_\alpha$, and
- $\mathcal{T}^-_{\alpha+1}$ is the set of objects in $\mathcal{D}$ such that either the objects are not (codes) of sentences of $\mathcal{L}^+$, or are (codes of) false sentences of $\mathcal{M}^+_\alpha$.

We can think of $\mathcal{T}_{\alpha+1}$ as a revaluation of $\mathcal{T}_\alpha$: $\mathcal{T}_{\alpha+1}$ improves the accuracy of $\mathcal{T}_\alpha$. For convenience, we define a revaluation operation $\mathcal{J}$:

$$\mathcal{J}(\mathcal{T}^+_\alpha) = \{\langle A \rangle : v_{\mathcal{M}^+_\alpha}(A) = 1\}$$

$$\mathcal{J}(\mathcal{T}^-_\alpha) = \{\langle A \rangle : v_{\mathcal{M}^+_\alpha}(A) = 0\} \cup \{\underline{n} : \neg sent_{\mathcal{L}^+}(n)\}$$

where $v_{\mathcal{M}^+_\alpha}(A)$ is the value assigned to A by $\mathcal{M}^+_\alpha$; $\underline{n}$ is a term for any natural number n; $sent_{\mathcal{L}^+}$ is a numerical predicate such that $sent_{\mathcal{L}^+}(n)$ iff n is the code of a sentence of $\mathcal{L}^+$. Thus, we have:

$$\langle \mathcal{T}^+_{\alpha+1}, \mathcal{T}^-_{\alpha+1} \rangle = \mathcal{J}\langle \mathcal{T}^+_\alpha, \mathcal{T}^-_\alpha \rangle$$

At 'transfinite' levels, the learner learns how to collect up the previous levels in the hierarchy of learning. Thus, at such levels, the extension (the anti-extension) is simply the set of objects in any extension (anti-extension) at the previous levels. Hence, we have: for any limit ordinal γ,

$$\langle \mathcal{T}^+_\gamma, \mathcal{T}^-_\gamma \rangle = \langle \bigcup_{\beta<\gamma} \mathcal{T}^+_\beta, \bigcup_{\beta<\gamma} \mathcal{T}^-_\beta \rangle$$

Hence, at the transfinite levels γ, the models are $\mathcal{M}^+_\gamma = \langle \mathcal{D}, \mathcal{I}, \mathcal{T}_\gamma \rangle$.

In summary, Kripke constructs a number of candidate interpretations via transfinite induction:

- Base case: $\langle \mathcal{T}^+_0, \mathcal{T}^-_0 \rangle = \langle \emptyset, \emptyset \rangle$
- Successor case: $\langle \mathcal{T}^+_{\alpha+1}, \mathcal{T}^-_{\alpha+1} \rangle = \mathcal{J}\langle \mathcal{T}^+_\alpha, \mathcal{T}^-_\alpha \rangle$
- Limit case: for any limit ordinal γ, $\langle \mathcal{T}^+_\gamma, \mathcal{T}^-_\gamma \rangle = \langle \bigcup_{\beta<\gamma} \mathcal{T}^+_\beta, \bigcup_{\beta<\gamma} \mathcal{T}^-_\beta \rangle$

The Fixed Point Theorem. In the informal picture, the learner finally reaches a 'fixed point' in the sense that he could not further improve his understanding of the concept of truth. Now the key question is: does the formal construction correspond to the informal picture? In other words, is there an ordinal σ such that $\langle \mathcal{T}^+_\sigma, \mathcal{T}^-_\sigma \rangle = \langle \mathcal{T}^+_{\sigma+1}, \mathcal{T}^-_{\sigma+1} \rangle = \mathcal{J}(\langle \mathcal{T}^+_\sigma, \mathcal{T}^-_\sigma \rangle)$?

The answer is yes. For convenience, we say that:

Definition 8. $\langle S'_1, S'_2 \rangle$ extends $\langle S_1, S_2 \rangle$ (i.e., $\langle S_1, S_2 \rangle \leq \langle S'_1, S'_2 \rangle$), iff $S_1 \subseteq S'_1$ and $S_2 \subseteq S_2$.

We can show that if $\langle \mathcal{T}_\sigma^+, \mathcal{T}_\sigma^- \rangle \leq \langle \mathcal{T}_\tau^+, \mathcal{T}_\tau^- \rangle$, then any sentence A that is true (false) in $\mathcal{M}_\sigma$ remains the same in $\mathcal{M}_\tau$. This entails that our revaluation operation $\mathcal{J}$ extends the interpretation of the predicate T: no sentence that was previously 'interpreted' (i.e., having the value 1 or having the value 0) would be reevaluated as 'uninterpreted' (i.e., having the value $\frac{1}{2}$). If more and more sentences of $\mathcal{L}^+$ get interpreted at each level, we will eventually exhaust the language $\mathcal{L}^+$ at some level σ. It is because there are only denumerably infinitely many sentences of $\mathcal{L}^+$. Hence, the interpretation of the predicate T will eventually stabilize. This is the basic idea of proving the existence of a fixed point.

Accordingly, our proof of the existence of a fixed point consists of three parts. Firstly, we show the following lemma.

Lemma 9. Let $\mathcal{M}_\sigma^+ = \langle \mathcal{D}, \mathcal{I}, \mathcal{T}_\sigma \rangle$ and $\mathcal{M}_\tau^+ = \langle \mathcal{D}, \mathcal{I}, \mathcal{T}_\tau \rangle$ be two models such that $\langle \mathcal{T}_\sigma^+, \mathcal{T}_\sigma^- \rangle \leq \langle \mathcal{T}_\tau^+, \mathcal{T}_\tau^- \rangle$. Then, for any $A \in \mathcal{L}^+$,

(i) If $v_{\mathcal{M}_\sigma}(A) = 1$, then $v_{\mathcal{M}_\tau}(A) = 1$.

(ii) If $v_{\mathcal{M}_\sigma}(A) = 0$, then $v_{\mathcal{M}_\tau}(A) = 0$.

Proof Sketch. We can prove it by induction on complexity of sentences of $\mathcal{L}^+$. □

Then, we show that the operation $\mathcal{J}$ extends the interpretation of the predicate T:

Lemma 10. Let $\mathcal{M}_\sigma^+ = \langle \mathcal{D}, \mathcal{I}, \mathcal{T}_\sigma \rangle$ and $\mathcal{M}_\tau^+ = \langle \mathcal{D}, \mathcal{I}, \mathcal{T}_\tau \rangle$ be two models such that $\langle \mathcal{T}_\sigma^+, \mathcal{T}_\sigma^- \rangle \leq \langle \mathcal{T}_\tau^+, \mathcal{T}_\tau^- \rangle$. Then, $\mathcal{J}(\langle \mathcal{T}_\sigma^+, \mathcal{T}_\sigma^- \rangle) \leq \mathcal{J}(\langle \mathcal{T}_\tau^+, \mathcal{T}_\tau^- \rangle)$.

Proof. By lemma 9, we have: for any $A \in \mathcal{L}^+$, if $v_{\mathcal{M}_\sigma}(A) = 1$, then $v_{\mathcal{M}_\tau}(A) = 1$. By the definition $\mathcal{J}$, if $\langle A \rangle \in \mathcal{J}(\mathcal{T}_\sigma^+)$, then $\langle A \rangle \in \mathcal{J}(\mathcal{T}_\tau^+)$.

Similar for the case of anti-extension. We can ignore the non-sentence case, since if an object is not a (code) of sentence of $\mathcal{L}^+$, it always is not. □

Finally, we can prove the fixed point theorem by using lemma 10:

Theorem 11 (Fixed Point Theorem)**.** There is an ordinal σ such that

$$\langle \mathcal{T}_\sigma^+, \mathcal{T}_\sigma^- \rangle = \langle \mathcal{T}_{\sigma+1}^+, \mathcal{T}_{\sigma+1}^- \rangle = \mathcal{J}(\langle \mathcal{T}_\sigma^+, \mathcal{T}_\sigma^- \rangle)$$

Proof. We show this by showing that there is an ordinal σ such that

$$\mathcal{J}(\mathcal{J}(\langle \mathcal{T}_\sigma^+, \mathcal{T}_\sigma^- \rangle)) = \mathcal{J}(\langle \mathcal{T}_\sigma^+, \mathcal{T}_\sigma^- \rangle)$$

Suppose for reductio that there is no such σ. Let θ be the first cardinal number that is larger than the cardinal number of the domain $\mathcal{D}$. By lemma 10, we have: for every $\beta < \theta$, $\mathcal{J}(\langle \mathcal{T}_\beta^+, \mathcal{T}_\beta^- \rangle) \leq \mathcal{J}(\langle \mathcal{T}_{\beta+1}^+, \mathcal{T}_{\beta+1}^- \rangle)$. By the definition of $\mathcal{J}$, we have: $\mathcal{J}(\langle \mathcal{T}_\beta^+, \mathcal{T}_\beta^- \rangle) \leq \mathcal{J}(\mathcal{J}(\langle \mathcal{T}_\beta^+, \mathcal{T}_\beta^- \rangle))$.

By our initial supposition, we have $\mathcal{J}(\langle \mathcal{T}_\beta^+, \mathcal{T}_\beta^- \rangle) < \mathcal{J}(\mathcal{J}(\langle \mathcal{T}_\beta^+, \mathcal{T}_\beta^- \rangle))$. Reapplying the definition of $\mathcal{J}$, we have: for any $\beta < \theta$, $\mathcal{J}(\langle \mathcal{T}_\beta^+, \mathcal{T}_\beta^- \rangle) < \mathcal{J}(\langle \mathcal{T}_{\beta+1}^+, \mathcal{T}_{\beta+1}^- \rangle)$.

Pick an object from $\mathcal{J}(\mathcal{T}_{\beta+1}^+) - \mathcal{J}(\mathcal{T}_\beta^+)$. We obtain a set with more objects (in particular, sentences) than one can find in $\mathcal{D}$. Yet, this is impossible. □

Thus, $\langle \mathcal{T}_\sigma^+, \mathcal{T}_\sigma^- \rangle$ is our target interpretation of the predicate T. Adding it to the base models, we obtain the target models.

Remarks on the Kripke's Construction. Firstly, we saw that Kripke shows us how to construct a fixed point in the sense that the interpretation of the truth predicate stabilizes. There is another sense in which $\langle \mathcal{T}_\sigma^+, \mathcal{T}_\sigma^- \rangle$ can be called a fixed point. Mathematically, a is a fixed point of the function $f(x)$ if $f(a) = a$. Thus, since $\mathcal{J}(\langle \mathcal{T}_\sigma^+, \mathcal{T}_\sigma^- \rangle) = \langle \mathcal{T}_\sigma^+, \mathcal{T}_\sigma^- \rangle$, we can say that $\langle \mathcal{T}_\sigma^+, \mathcal{T}_\sigma^- \rangle$ is a fixed point of $\mathcal{J}$. Such a fixed point is called the *minimal fixed point*. It is because it can be shown that if $\langle \mathcal{T}_\delta^+, \mathcal{T}_\delta^- \rangle$ is a fixed point of $\mathcal{J}$, then $\langle \mathcal{T}_\sigma^+, \mathcal{T}_\sigma^- \rangle \leq \langle \mathcal{T}_\delta^+, \mathcal{T}_\delta^- \rangle$. That is, if A has the value 1 or 0 in $\mathcal{M}_\sigma^+ = \langle \mathcal{D}, \mathcal{V}, \langle \mathcal{T}_\sigma^+, \mathcal{T}_\sigma^- \rangle \rangle$, the value of A remains the same in $\mathcal{M}_\delta^+ = \langle \mathcal{D}, \mathcal{V}, \langle \mathcal{T}_\delta^+, \mathcal{T}_\delta^- \rangle \rangle$. We call the models with the minimal fixed point is called the *minimal fixed point models*.

In fact, there are other fixed points. One can construct a fixed point using a pair other than $\langle \emptyset, \emptyset \rangle$ as the starting point. Martin and Woodruff (1975) show us the existence of the maximal fixed point in the weak Kleene valuation scheme B_3^I. Using Zorn's lemma, they also show that every fixed point can be extended to the maximal fixed point. There is an issue with which fixed points properly model our concept of truth in natural languages. Gupta and Belnap (1993) argue that only the fixed points in which there is no vicious reference in the language give the correct interpretation of truth. We leave this aside.

Thirdly, notice that to construct models that have the identity of truth property, we can begin with a valuation scheme other than K_3. In fact, a valuation scheme $v_\mathcal{M}$ is suitable, as long as $v_\mathcal{M}$ is *monotonic* in the following sense.

Definition 12 (Model Extension)**.** We say that $\mathcal{M}'$ extends $\mathcal{M}$, iff $\mathcal{M}'$ and $\mathcal{M}$ are just the same, except that $\mathcal{M}'$ interprets whatever, if anything, $\mathcal{M}$ left uninterpreted.

Definition 13 (Monotonicity)**.** A valuation scheme $v_\mathcal{M}$ is monotonic iff for any $\mathcal{M}$ and $\mathcal{M}'$, and any $A \in \mathcal{L}$, given that $\mathcal{M}'$ extends $\mathcal{M}$,

- if $v_\mathcal{M}(A) = 1$, then $v_{\mathcal{M}'}(A) = 1$; and
- if $v_\mathcal{M}(A) = 0$, then $v_{\mathcal{M}'}(A) = 0$.

We can show that the K_3 valuation scheme via induction. We leave the proof for the readers.

The Liar Sentence, Groundedness and Paradoxicality. One important question we have not discussed is: what is the status of the Liar sentence? Recall that in the intuitive picture, the learner iteratively learns how to predicate truth to sentences until he reaches a fixed point. However, not all sentence will be predicated as true or false in the learning process. In particular, the truth-teller 'This sentence is true' and the Liar sentence will not be characterized as true or false in the process. Kripke calls these sentence *ungrounded*. Specifically, Kripke's defines:

- A sentence is *grounded* if the learner is entitled to attribute truth or falsity to the sentence in the process described; otherwise, *ungrounded*.

Formally, the notion of groundedness can be defined as follows:

Definition 14 (Groundedness)**.** For any $A \in \mathcal{L}$, A is grounded iff either $\langle A \rangle \in \mathcal{T}^+$ or $\langle A \rangle \in \mathcal{T}^-$, where $\langle \mathcal{T}^+, \mathcal{T}^- \rangle$ is the interpretation of the truth predicate in the minimal fixed point models.

Kripke distinguishes ungroundedness from another notion called paradoxicality. He suggests that we can define paradoxicality in this way:

- For any $A \in \mathcal{L}$, A is paradoxical iff neither $\langle A \rangle \in \mathcal{T}^+$ nor $\langle A \rangle \in \mathcal{T}^-$, where $\langle \mathcal{T}^+, \mathcal{T}^- \rangle$ is the interpretation of the truth predicate in any fixed point model.

Note that all paradoxical sentence is ungrounded; but not vice versa. Cases in point are the truth-teller and the Liar sentence. The truth-teller can be arbitrarily assigned the value 1 or the value 0. On the other hand, the only possible value that the Liar sentence λ can have is the value $\frac{1}{2}$. To see the latter claim, we first suppose that $\lambda \models\!\!\models^{K_3TT} \neg T\langle\lambda\rangle$. Secondly, recall that fixed point models have the identity of truth property. Thus, $v_{\mathcal{M}^+}(\lambda) = v_{\mathcal{M}^+}(T\langle\lambda\rangle)$. According to the semantics of $\neg$, we have $v_{\mathcal{M}^+}(\neg T\langle\lambda\rangle) = 1 - v_{\mathcal{M}^+}(T\langle\lambda\rangle)$. Hence, it follows that $v_{\mathcal{M}^+}(\neg T\langle\lambda\rangle) = 1 - v_{\mathcal{M}^+}(\lambda)$. Suppose for reductio that λ does not take the value $\frac{1}{2}$. We have two cases:

» Suppose that Liar sentence λ takes the value 1. By the identity of truth, $T\langle\lambda\rangle$ takes the value 1. According to the semantics of $\neg$, $\neg T\langle\lambda\rangle$ takes the value 0. Yet, this means that λ cannot be equivalent to $\neg T\langle\lambda\rangle$, contradicting our assumption.

» Suppose that Liar sentence λ takes the value 0. By the identity of truth, $T\langle\lambda\rangle$ takes the value 0. According to the semantics of $\neg$, $\neg T\langle\lambda\rangle$ takes the value 1. Thus, $\neg T\langle\lambda\rangle \not\models^{K_3TT} \lambda$. This means that λ cannot be equivalent to $\neg T\langle\lambda\rangle$, contradicting our assumption.

Accordingly, given that $\lambda \models\!\!\models^{K_3TT} \neg T\langle\lambda\rangle$, the Liar sentence can only take the value $\frac{1}{2}$. Thus, the truth-teller is ungrounded but not paradoxical; whereas the Liar sentence is ungrounded and paradoxical.

Parenthetical Remark. Kripke's account of paradoxicality is *prima facie* dubious. As argued by Gupta and Belnap (1993), not all fixed points give the correct interpretation of truth in natural languages. If so, there is no reason to take all fixed point into account to determine whether or not a sentence is paradoxical. Nevertheless, we will not pursue the issue with which fixed points best model the concept of truth in natural languages. In what follows, we take the models in Kripke's theory to be K_3^+ models defined in definition 6.

3.3 Conditional

In this section, we discuss some problems concerning the material conditional $\supset$ in K_3TT. K_3TT is often criticized for not having a suitable conditional. Firstly, the material conditional $\supset$ in K_3TT is too weak to carry out ordinary reasoning (Field, 2008). Specifically,

i. $\not\models^{K_3TT} A \supset A$,

ii. $\not\models^{K_3TT} A \supset (A \vee B)$

iii. $A \supset B \not\models^{K_3TT} (C \supset A) \supset (C \supset B)$

To see these, consider the K_3^+ model in which A, B and C take the value $\frac{1}{2}$. It is easy to check that this model shows (i) - (iii).

Secondly, since $A \supset A$ is not valid in K_3TT, the T-schema $T\langle A\rangle \equiv A$ is not valid either. This is obvious, because each direction of the T-schema is an instance of $A \supset A$, given that A and $T\langle A\rangle$ are intersubstitutable. (Field, 2008).

Thirdly, K_3TT cannot represent self-reference via the material biconditional. A case in point is that the Liar sentence is sometimes represented by supposing that $\lambda \equiv \neg T\langle\lambda\rangle$. However, since λ must have the value $\frac{1}{2}$, we also have $\neg T\langle\lambda\rangle$. Hence, $\lambda \equiv \neg T\langle\lambda\rangle$ must take the value $\frac{1}{2}$. Thus, $\lambda \equiv \neg T\langle\lambda\rangle$ cannot be true in K_3TT.

3.4 Expressive Limitations and Revenge Paradoxes

3.4.1 The Problem of Expressing the Solution

In this section, we turn to another sort of problems concerning K_3TT. The issue concerns whether or not K_3TT can express certain important notions such as determinate truth, truth-value gaps and exclusion negation.

Kripke's theory K_3TT is motivated by the intuition that the Liar sentence is neither true nor false. It appears that K_3TT achieves the desired result: K_3^+ models assign the Liar sentence λ the intermediate value $\frac{1}{2}$.

However, such an appearance of success is merely a mirage. For one thing, K_3TT cannot truthfully report its own solution. In particular, the claim that the Liar sentence is neither true nor false, if formalized as $\neg(T\langle\lambda\rangle \vee T\langle\neg\lambda\rangle)$, cannot have the value 1.

To see this, recall that the Liar sentence λ takes the value $\frac{1}{2}$ in any K_3^+ models $\mathcal{M}^+$. According to the valuation scheme, $\neg\lambda$ must have the value $\frac{1}{2}$ as well. By the identity of truth, it follows that both $T\langle\lambda\rangle$ and $T\langle\neg\lambda\rangle$ must take the value $\frac{1}{2}$. Thus, according to the valuation scheme again, $T\langle\lambda\rangle \vee T\langle\neg\lambda\rangle$ takes the value $\frac{1}{2}$ and thus $\neg(T\langle\lambda\rangle \vee T\langle\neg\lambda\rangle)$ takes the value $\frac{1}{2}$ in any K_3^+ models $\mathcal{M}^+$. Thus, it is doubtful whether or not K_3TT succeeds in explaining how we can exhaustively and exclusively characterize (paradoxical) sentences.

For another thing, the intermediate value $\frac{1}{2}$ should not be construed as meaning 'neither true nor false', if 'neither true nor false' is formalized via the truth predicate. In K_3TT, the claim that the Liar sentence is neither true nor false amounts to the claim the Liar sentence is both true and false (if formalized as $T\langle\lambda\rangle \wedge \neg T\langle\lambda\rangle$). To see this, notice that, for any model $\mathcal{M}^+$, both $T\langle\lambda\rangle$ and $\neg T\langle\lambda\rangle$ take the value $\frac{1}{2}$. Hence, $T\langle\lambda\rangle \wedge \neg T\langle\lambda\rangle$ must take the value $\frac{1}{2}$. Accordingly, $T\langle\lambda\rangle \wedge \neg T\langle\lambda\rangle$ and $\neg(T\langle\lambda\rangle \vee T\langle\neg\lambda\rangle)$ are equivalent and must fall together. Hence, given that 'neither true nor false' is formalized in terms of the truth predicate, the intermediate value $\frac{1}{2}$ cannot be construed as meaning 'neither true nor false' as intended.

Parenthetical Remark. $T\langle\lambda\rangle \wedge \neg T\langle\lambda\rangle$ and $\neg(T\langle\lambda\rangle \vee T\langle\neg\lambda\rangle)$ stand together in the dialetheic logics $LPTT$ and $STTT$. In such logics, there is no countermodel to the sentences which always take the intermediate value $\frac{1}{2}$. Accordingly, it is misleading to say that the intermediate value $\frac{1}{2}$ as meaning 'both true and false', but not 'neither true nor false' in $LPTT$ and $STTT$.

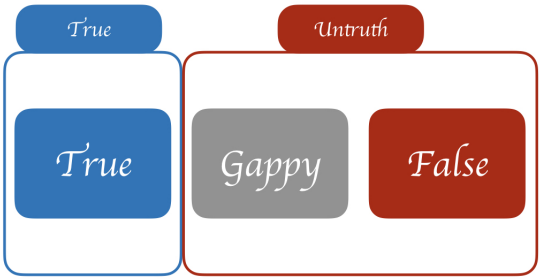

Figure 3.1: Semantic Characterization in Kripke's Theory

On a related note, it appears that there is a close connection between being neither true nor false and being not true. Specifically, it appears that if a sentence A is neither true nor false, then A is not true. Unfortunately, $\neg T\langle\lambda\rangle$ cannot have the value 1. Recall that $T\langle\lambda\rangle$ must have the value $\frac{1}{2}$. Then, according to the usual valuation scheme, it follows that $\neg T\langle\lambda\rangle$ must have the value $\frac{1}{2}$. As a matter of fact, Kripke is well-aware of this expressive limitation of his logic:

> [T]he present approach certainly does not claim to give a universal language, and I doubt that such a goal can be achieved... [T]here are assertions we can make about the object language which we cannot make in the object language. For example, Liar sentences are not true in the object language, in the sense that the inductive process never makes them true; but we are precluded from saying this in the object language by our interpretation of negation and the truth predicate. If we think of the minimal fixed point, say under the Kleene valuation as giving a model of natural language, then the sense in which we can say, in natural language, that a Liar sentence is not true must be thought of as associated with some later stage in the development of natural language, one in which speakers reflect on the generation process leading to the minimal fixed point. It is not itself a part of that process. The necessity to ascend to a metalanguage may be one of the weaknesses of the present theory. The ghost of the Tarski hierarchy is still with us. (Kripke, 1975, p.714)

3.4.2 Revenge Paradoxes

Increasing the Expressive Power. Facing the above problems, one may suggest that we increase the expressive power of a language by adding new connectives:

A	$\sim A$	$T\langle A\rangle$	$\mathbb{D}A$	$\mathbb{G}A$
1	0	1	1	0
$\frac{1}{2}$	1	$\frac{1}{2}$	0	1
0	1	0	0	0

The proposal continues: we can truthfully express that the Liar sentence is not true by saying that $\sim T\langle\lambda\rangle$. The new negation is usually called the *exclusion negation*. Unlike the old negation $\neg$ (which is known as the *choice negation*), the exclusion negation $\sim$ always determines a classical value: $\sim A$ must either take the value 1 or 0. Since the Liar sentence λ and $T\langle\lambda\rangle$ must take the intermediate value $\frac{1}{2}$, $\sim T\langle\lambda\rangle$ must take the value 1.

Alternatively, we can express that the Liar sentence is not determinately true by saying $\neg\mathbb{D}\lambda$; since the intuitive meaning of $\mathbb{D}A$ is that A is determinately true. Since λ must take the value $\frac{1}{2}$, $\mathbb{D}\lambda$ must take the value 0. Thus, $\neg\mathbb{D}\lambda$ must take the value 1.

Similarly, we can express that the Liar sentence is neither true nor false by saying that $\mathbb{G}\lambda$. Since λ must take the value $\frac{1}{2}$, $\mathbb{G}A$ must take the value 1.

For simplicity, we take the $\mathbb{D}$ operator as primitive. Other operators can be defined in terms of $\mathbb{D}$ and usual connectives:

- $\mathbb{G}A =_{df} \neg\mathbb{D}A \wedge \neg\mathbb{D}\neg A$
- $\sim A =_{df} \mathbb{G}A \vee \neg\mathbb{D}A$

Let $K_3TT^{\mathbb{D}}$ be K_3TT augmented with the operator $\mathbb{D}$ defined by the above truth table.

Revenge in Natural Languages. However, if we can express notions, such as determinate truth, truth-value gaps and exclusion negation, we can express sentences like:

(15) a. (15a) is not true.
b. (15b) is not determinately true.
c. (15c) is gappy or false.

These sentences can be called the Strengthened Liar sentences. As expected, if we could express the Strengthened Liar sentences, some unpalatable consequences would follow: by some familiar reasoning, we could generate some new paradoxes, which are just as virulent as the Liar paradox. Take (15b) as an example. Either (15b) is true, false or gappy. Then:

» Suppose that (15b) is true. Then it is determinately true. But this contradicts to what (15b) says. Hence, (15b) is false.

» Suppose that (15b) is false. Then it is not determinately true. But this is precisely what (15b) says. Hence, (15b) is true.

» Suppose that (15b) is gappy. Then it is not determinately true. But this is precisely what (15b) says. Hence, (15b) is true.

Thus, in any case, there is a contradiction. Thus, (15b) cannot be exhaustively and exclusively characterized as true or false or gappy.

Parenthetical Remark. No paradox arises out of (15a), if the 'not' in (15a) is understood as the choice negation; the paradox arises only if the 'not' is construed as the exclusion negation. In particular, if (15a) is neither true nor false, then the choice negation of it is also neither true nor false. On the other hand, if (15a) is neither true nor false, then the exclusion negation of it is true.

Formalizing the Revenge Paradoxes. The usual way to represent paradoxical sentences like (15b) is to suppose that ξ to be equivalent to $\neg\mathbb{D}\xi$. While there are two notions of equivalence – the semantic equivalence $=\!|\!\models$ and the (material) biconditional $\equiv$, K_3TT cannot represent the self-referential character of paradoxical sentences via the material biconditional $\equiv$. Thus, we formalize the Strengthened Liar sentence by the semantic equivalence $=\!|\!\models$. Let $K_3TT^{\mathbb{D}}_{=\!|\!\models}$ be K_3TT augmented with the operator $\mathbb{D}$ and uses the semantic equivalence $=\!|\!\models$ to represent self-reference. By some familiar liar-like reasoning, we can show that $K_3TT^{\mathbb{D}}_{=\!|\!\models}$ is trivial.

Fact 15. $K_3TT^{\mathbb{D}}_{=\!|\!\models}$ is trivial.

Proof. Suppose that $\xi =\!|\!\models^{K_3TT^{\mathbb{D}}} \neg\mathbb{D}\xi$.

- Suppose that $v_{\mathcal{M}^{+\mathbb{D}}_{=\!|\!\models}}(\xi) = 1$. By the semantics of $\mathbb{D}$, we have $v_{\mathcal{M}^{+\mathbb{D}}_{=\!|\!\models}}(\mathbb{D}\xi) = 1$. Then, by the semantics of $\neg$, we have $v_{\mathcal{M}^{+\mathbb{D}}_{=\!|\!\models}}(\neg\mathbb{D}\xi) = 0$. Hence, $\xi \not=\!|\!\models^{K_3TT^{\mathbb{D}}} \neg\mathbb{D}\xi$, violating the initial assumption.
- Suppose that $v_{\mathcal{M}^{+\mathbb{D}}_{=\!|\!\models}}(\xi) = 0$. By the semantics of $\mathbb{D}$, we have $v_{\mathcal{M}^{+\mathbb{D}}_{=\!|\!\models}}(\mathbb{D}\xi) = 0$. Then, by the semantics of $\neg$, we have $v_{\mathcal{M}^{+\mathbb{D}}_{=\!|\!\models}}(\neg\mathbb{D}\xi) = 1$. Hence, $\neg\mathbb{D}\xi \not=\!|\!\models^{K_3TT^{\mathbb{D}}} \xi$, violating the initial assumption.
- Suppose that $v_{\mathcal{M}^{+\mathbb{D}}_{=\!|\!\models}}(\xi) = \frac{1}{2}$. By the semantics of $\mathbb{D}$, we have $v_{\mathcal{M}^{+\mathbb{D}}_{=\!|\!\models}}(\mathbb{D}\xi) = 0$. Then, by the semantics of $\neg$, we have $v_{\mathcal{M}^{+\mathbb{D}}_{=\!|\!\models}}(\neg\mathbb{D}\xi) = 1$. Hence, $\neg\mathbb{D}\xi \not=\!|\!\models^{K_3TT^{\mathbb{D}}} \xi$, violating the initial assumption.

Thus, in any case, ξ cannot be intersubstitutable for $\neg\mathbb{D}\xi$. Given that $\xi =\!|\!\models^{K_3TT^{\mathbb{D}}} \neg\mathbb{D}\xi$, there is no value available for ξ to take. But there are no such models. Thus, there are no countermodel to any argument, and every argument is valid. □

3.4.3 Metalanguage, Instrumentalism and Model Theory

Kripke suggests that if we wish to report the status of the Liar sentence, we have to ascend to the metalanguage for $\mathcal{L}^+$:

> Note that the metalanguage in which we write this paper can be regarded as containing no truth gaps. A sentence either does or does not have a truth value in a given fixed point.
>
> Such semantical notions as "grounded", "paradoxical", etc. belong to the metalanguage. This situation seems to me to be intuitively accept-

> able; in contrast to the notion of truth, none of these notions is to be found in natural language in its pristine purity, before philosophers reflect on its semantics (in particular, the semantic paradoxes). If we give up the goal of a universal language, models of the type presented in this paper are plausible as models of natural language at a stage before we reflect on the generation process associated with the concept of truth, the stage which continues in the daily life of nonphilosophical speakers. (Kripke, 1975, p.714)

Kripke's suggestion can be distinguished into three different claims.

i. We can express semantical notions such as "grounded", "paradoxical" in a classical metatheory for $\mathcal{L}^+$.

ii. We should not aim at constructing a formal semantics that reflects the 'real semantics' of our natural languages.

iii. Semantical notions such as "grounded", "paradoxical" belong to the metalanguage.

Firstly, it has been noted that if the metatheory for $\mathcal{L}^+$ is classical, then it provides the resources to generate the revenge paradoxes (Leitgeb, 2007). The reason is that if we can form a Strengthened Liar sentence, and the metalanguage obeys the familiar classical laws, we can easily generate the Liar's revenge by familiar reasoning. Accordingly, Kripke's claim (i) must be qualified: the metatheory for $\mathcal{L}^+$ should be not a classical one. This raises the question of whether or not moving to a non-classical metatheory can avoid revenge paradoxes. We leave this as an open problem. (For more discussion about this issue, see Leitgeb 2007; Bacon 2012.)

As for the claim (ii), many theorists adopt the same attitude. For instance, Beall (2009) says:

> [T]he [revenge] argument itself relies on various assumptions that involve quite complex issues. One conspicuous assumption is that the 'semantics' of [the formal lannguage] is intended to reflect the semantics of our real language. This needn't be the case [...] The formal account of [truth] is given to elucidate the logical behavior of [truth]; it isn't intended to reflect the 'real semantics' of our real language, whatever such 'real semantics' might come to. (ibid, p.56 - 57)

However, the suggestion that we should give up the goal of giving a universal language is irrelevant. Like truth and falsity, semantical notions such as determinate truth, truth-value gaps and exclusion negation are essential in explaining how we exhaustively and exclusively characterize declarative sentences in natural languages. On the gap approaches, the reason for introducing truth-value gaps is to allow some sentences to have a semantic status other than truth and falsity. Unfortunately, new paradoxical sentences can be formed by making use of the notion of gaps. If a formal theory of truth avoids triviality by disallowing such semantical notions to be expressible, then we are still left with puzzlement: it is unclear how we can characterize the Strengthened Liar sentences in natural languages. For one thing, these semantical notions seem to be intelligible. (We

will come back to this issue.) For another thing, we can form the Strengthened Liar sentences in natural languages by making use of these semantical notions. Thus, it is one thing not to give a formal semantics that reflects the 'real semantics' of our natural languages. As far as the Liar paradox is concerned, it is another thing to turn our heads and not to allow important semantical notions to be expressible.

Like the claim (ii), the claim (iii) is shared by many theorists. For instance, Brady (1989) points out that revenge arguments involve model-theoretic notions:

> It seems to me that [...] the [relevant notion] used in generating the [alleged problem] involves reference to the details of the model [used in our 'formal account']. That is, 'p takes the value t', or some equivalent, makes reference to the specific values of a model [...] and thus goes beyond natural language expressions which just refer to truth and falsity. (ibid, p. 467)

Beall (2009) argues that revenger conflates model-dependent notions with model-independent notions:

> Some standardly object that Kripke's account doesn't answer [the non-triviality (consistency) project], the reason being that certain notions used in the metalanguage are not expressible in the object- or 'model language'[...] The standard sort of 'revenge' arguments at best seem to conflate model-dependent notions with the model-independent (or absolute) notions that the former purport to illuminate. So long as one takes a sufficiently 'instrumentalist' view of the given formal construction, the given sort of revenge objection – popular as it is – need not undermine the Kripkean proposal. (ibid, p. 75 - 76)

Field (2007) gives a similar reply to revenge paradoxes. He argues that the extension of model-independent truth and the extension of 'having the value 1' are not the same:

> Consider a classical model for the 'true'-free part [$\mathcal{L}$] of the language [$\mathcal{L}^+$]. [$\mathcal{L}$] includes standard set theory. Suppose we take a highly natural model for [$\mathcal{L}$] [...] whose domain consists of all non-sets together with all sets of rank less than the first inaccessible cardinal; call this [...] model [$\mathcal{M}$]. This assumes of course that there are inaccessible cardinals [...] But now consider the sentence 'There are inaccessible cardinals': it's true, but false in [$\mathcal{M}$], i.e. has semantic value 0 in [$\mathcal{M}$]; its negation is false, but has value 1 in [$\mathcal{M}$]. Having semantic value 1 in [$\mathcal{M}$] doesn't correspond to truth, or to determinate truth, or anything like that ... The point made here for [$\mathcal{M}$] applies to any other model that can be defined within set theory, by Tarski's Theorem, and this includes all models of set theory that are at all "natural". (ibid, p. 104)

Since model-theoretic notions are defined in terms of (natural) models, the extension of such notions can only have a subset of the domain of models. On the other hand, since model-independent notions are not defined by models, their extensions are not restricted to any particular set. So model-theoretic notions and

model-independent notions must diverge. Because of this, Field warns us:

> [T]he notion of semantic value is inevitably a somewhat artificial construction that can only be understood as model-relative, and drawing conclusions about how sentences are to be evaluated with respect to properties that are not model-relative (for instance, truth, determinate truth, and so forth) is highly problematic. (ibid, p. 105).

However, even if there are some fundamental differences between model-theoretic notions and model-independent notions, revenge paradoxes can be posed in a model-independent fashion. The operator $\mathbb{D}$ we previously introduced is not intended to be construed as model-relative. Field is well-aware of this point:

> Obviously the proponent of the simple revenge problem doesn't intend 'designated' to be understood as model-relative. The question then arises, how is it to be understood. I do not deny that it is possible to introduce into the language an operator with many of the features that the proponent of revenge wants, and which is not model-relative... But such predicates only breed paradox if they satisfy all the assumptions used in the derivations above. It turns out that one can get predicates that satisfy most of the assumptions used in the derivations above; the one place they fail is that excluded middle cannot be assumed for them.
>
> There is a revenge problem [...] only if there is reason to think that we can understand a notion of "designatedness" that obeys those other assumptions plus excluded middle. And why assume that? I think what underlies the simple revenge problem is the thought that the model-relative designatedness predicates all obey excluded middle, so there must be an absolute designatedness predicate that does too. But this assumption seems to me completely unwarranted: one just can' t assume that one can extrapolate in this way from the case of model-relative predicates, which make sense only by virtue of "misinterpreting" the quantifiers as having restricted range, to the unrelativized case where no such "misinterpretation" is in force. (Field, 2007, p. 108–109)

Field claims that we cannot assume that there is a model-independent notion of determinateness $\mathbb{D}$ such that the LEM holds for it (i.e., $\mathbb{D}A \vee \neg\mathbb{D}A$). At this point, Field seems to think that the notions of bivalent determinateness can only be motivated by the notion of designated value. According to Field, since the extension of model-independent truth and the extension of 'having the value 1' (designatedness) diverge, one cannot read a bivalent determinateness off the model-theoretic semantics.

However, bivalent determinateness needs not to be motivated by bivalent designatedness. Bivalent determinateness is in close connection with Strawson's analysis of presuppositions, which is one of the main motivation for the gap approaches to the Liar paradox. According to Strawson's analysis, sentences can be characterized as true, or false, or neither. If this is so, sentences can also be

re-characterized as true or not.

Presumably, Field might want to reject Strawson's analysis. However, it seems that where bivalent determinateness stems from is the thought that all declarative sentence can be characterized as bona fide true or not. It seems that it is part of the meaning of 'untruth' that sentences whose status is gappy or false or anything other than truth should be characterized as not true. Thus, if one admits semantic status other than truth, one should also admit that there is a notion of bivalent truth or bivalent determinateness. In the next chapter, we will see that Field rejects such a thought and proposes using a determinacy operator D that does not obey the LEM to characterize sentences. (We will come back to this issue in the next chapter.)

In any case, it is misleading to claim that revenge arguments conflate model-theoretic notions with model-independent notions. For one thing, revenge arguments can be formulated in terms of model-independent notions. For another thing, it seems that such model-independent notions need not to be motivated by the characteristics of the model-theoretic semantics.

3.5 Conclusion

In this chapter, we saw that K_3TT has two major defects. Firstly, the material conditional $\supset$ in K_3TT cannot play the role of conditional reasoning. Secondly, K_3TT cannot report the status of the Liar sentence; increasing the expressive power of K_3TT will trivialize the theory.

In the next chapter, we will see that Field attempts to introduce a reasonable conditional. Field also defines a determinacy operator D in terms of the new conditional, and uses the operator D to characterize paradoxical sentences.

Chapter 4

Paracomplete Gap Theory: Field's Theory of Truth

In this chapter, we discuss Field's theory of truth. In §4.1, we discuss how Field introduces a new conditional $\rightarrow$ to play the role of conditional reasoning. In §4.2, we discuss how Field defines a determinacy operator D in terms of the new conditional $\rightarrow$. We also discuss how Field uses the determinacy operator D to characterize paradoxical sentences. Finally, in §4.3, we evaluate whether or not Field's theory is revenge-immune. We see that Field's theory does not allow for and cannot be augmented with an intuitive notion of determinate truth, which Field calls *super-determinateness*, a general notion of truth-value gaps and exclusion negation. We argue that since Field's theory cannot express such intuitively appealing notions which are frequently used by ordinary speakers, it fails to shed light on how we can characterize sentences in natural languages.

4.1 Field's Conditional

4.1.1 Field's Construction

Field shows us how to introduce a suitable conditional $\rightarrow$ into Strong Kleene logic such that the truth predicate T obeys the identity of truth.

The basic idea is that we start from an arbitrary transparent valuation for sentences of the form $A \rightarrow B$. We say that:

- A valuation $v_{\mathcal{M}^+}$ for a set of sentences Γ is *transparent* iff for any $B \in \Gamma$, and any $\mathcal{M}^+$, if $B^{@}$ is the result of replacing some subsentence A of B with $T\langle A\rangle$, then $v_{\mathcal{M}^+}(B) = v_{\mathcal{M}^+}(B^{@})$.

Let $\mathcal{L}^+$ be the base language $\mathcal{L}$ extended with the truth predicate T and the new conditional $\rightarrow$. Based on a transparent valuation for $A \rightarrow B$, we construct a minimal Kripkean (Strong Kleene) fixed point. We can show by induction that the minimal fixed point over a transparent valuation for $A \rightarrow B$ is a transparent valuation for the whole language $\mathcal{L}^+$. Such a valuation has the identity of truth property. However, since the initial transparent valuation for $A \rightarrow B$ is arbitrary, it should be revised.

Field proposes the following revision procedure:

- Base: for any A and B in $\mathcal{L}^+$,

  $v_{\mathcal{M}_0^+}(A \to B) = \frac{1}{2}$

- Successor: suppose $v^*_{\mathcal{M}_\alpha^+}$ is transparent valuation for the whole language $\mathcal{L}^+$ at the level α. Then, the valuation for conditional sentences $v_{\mathcal{M}_{\alpha+1}^+}$ at the level $\alpha + 1$ is:

$$v_{\mathcal{M}_{\alpha+1}^+}(A \to B) = \begin{cases} 1, & \text{if } v^*_{\mathcal{M}_\alpha^+}(A) \preceq v^*_{\mathcal{M}_\alpha^+}(B). \\ 0, & \text{if } v^*_{\mathcal{M}_\alpha^+}(A) \succ v^*_{\mathcal{M}_\alpha^+}(B). \end{cases}$$

- If γ is a limit ordinal, then:

$$v_{\mathcal{M}_\gamma^+}(A \to B) = \begin{cases} 1, & \text{if } v_{\mathcal{M}_\alpha^+}(A) \preceq v_{\mathcal{M}_\alpha^+}(B) \\ & \text{for some } \beta < \gamma \text{ and for all } \alpha \text{ such that } \beta \leq \alpha < \gamma. \\ 0, & \text{if } v_{\mathcal{M}_\alpha^+}(A) \succ v_{\mathcal{M}_\alpha^+}(B) \\ & \text{for some } \beta < \gamma \text{ and for all } \alpha \text{ such that } \beta \leq \alpha < \gamma. \\ \frac{1}{2}, & \text{otherwise.} \end{cases}$$

At the base level, sentences of the form $A \to B$ are assigned the value $\frac{1}{2}$. Call the valuation $v_{\mathcal{M}_0^+}$. Trivially, $v_{\mathcal{M}_0^+}$ is a transparent valuation for $A \to B$. Then, using Kripke's construction, we can extend $v_{\mathcal{M}_0^+}$ to a transparent valuation for the whole language $\mathcal{L}^+$. Let $v^*_{\mathcal{M}_0^+}$ be the minimal Kripkean fixed point over $v_{\mathcal{M}_0^+}$. However, $v^*_{\mathcal{M}_0^+}$ is not satisfactory. To see this, let A be $1 = 1$ and B be $1 = 0$. In this case, $v^*_{\mathcal{M}_0^+}(1 = 1) = 1$, and $v^*_{\mathcal{M}_0^+}(1 = 0) = 0$. This should render $(1 = 1) \to (1 = 0)$ take the value 0. However, $v^*_{\mathcal{M}_0^+}$ does not respect this, because sentences of the form $A \to B$ must have the value $\frac{1}{2}$ in $v^*_{\mathcal{M}_0^+}$.

Nevertheless, we can successively revise the transparent valuations for the whole language $\mathcal{L}^+$. Suppose that $v^*_{\mathcal{M}_\alpha^+}$ is a transparent valuation for the whole language. Based on the revised valuation $v_{\mathcal{M}_{\alpha+1}^+}$, we can construct another Kripkean fixed point $v^*_{\mathcal{M}_{\alpha+1}^+}$. Again, $v^*_{\mathcal{M}_{\alpha+1}^+}$ has the identity of truth property and is a transparent valuation for the whole language $\mathcal{L}^+$. The revision process continues transfinitely. At the limit levels γ, we collect up all prior Kripkean fixed points over transparent valuation for $A \to B$. Then, we construct a Kripkean fixed point over $v_{\mathcal{M}_\gamma^+}$.

There are three possible outcomes of the revision process. Such possible outcomes can be construed as the *ultimate value* of a given sentence. The ultimate value of A, written as ${v^u}_{\mathcal{M}^+}(A)$, can be defined as follows:

- ${v^u}_{\mathcal{M}^+}(A) = \begin{cases} 1, \text{if there is a } \beta \text{ such that for all } \alpha \leq \beta, v_{\mathcal{M}_\alpha^+}(A) = 1. \\ 0, \text{if there is a } \beta \text{ such that for all } \alpha \leq \beta, v_{\mathcal{M}_\alpha^+}(A) = 0. \\ \frac{1}{2}, \text{otherwise.} \end{cases}$

For simplicity, we stipulate that ${v^u}_{\mathcal{M}^+}(A)$ takes the value $\frac{1}{2}$ even if $v_{\mathcal{M}_\alpha^+}(A)$ keeps oscillating in the revision process.

Field (2003) shows that the ultimate value ${v^u}_{\mathcal{M}^+}(A)$ obeys the Strong Kleene valuation for the negations, disjunctions, conjunctions, and quantifiers. Field also provides a consistency proof for his theory.

As expected, the consequence relation $\models^F$ in Field's theory is defined as the preservation of the value 1 in all models:

Definition 16 (F-Consequence)**.** $\Gamma \models^F \Delta$ iff if $v^u_{\mathcal{M}^+}(A) = 1$ for all $A \in \Gamma$, then $v^u_{\mathcal{M}^+}(B) = 1$ for some $B \in \Delta$.

Parenthetical Remark. Field's final theory is continuum-valued, although his construction focuses on the three-valued case. But for simplicity, we will not present the details of his continuum-valued semantics.

4.1.2 Some Features of Field's Conditional

Field's conditional does have many desirable features. For instance, we have:

- $\models^F A \to A$
- $\models^F T\langle A\rangle \leftrightarrow A$
- $A, A \to B \models^F B$
- $\models^F A \to \neg\neg A$
- $\models^F (A \wedge B) \to A$
- $A, \neg B \models^F \neg(A \to B)$

One important issue concerning conditionals is Curry's Paradox. Curry's paradox involves a sentence like this:

(16) If this is true, then anything is true.

Field represents the Curry sentence (16) as:

$$\kappa \leftrightarrow (T\langle\kappa\rangle \to \bot)$$

From this, we can argue that everything would follow:

1	$\kappa \leftrightarrow (T\langle\kappa\rangle \to \bot)$	The Def of κ
2	$T\langle\kappa\rangle \leftrightarrow (T\langle\kappa\rangle \to \bot)$	1: Intersubstitutivity of $T\langle A\rangle$ and A
3	$T\langle\kappa\rangle \to (T\langle\kappa\rangle \to \bot)$	2: Left to right of $\leftrightarrow$
4	$(T\langle\kappa\rangle \wedge T\langle\kappa\rangle) \to \bot$	3: Importation Principle
5	$T\langle\kappa\rangle \to \bot$	4: Intersubstitutivity of $A \wedge A$ with A
6	$(T\langle\kappa\rangle \to \bot) \to T\langle\kappa\rangle$	2: Right to left
7	$T\langle\kappa\rangle$	5, 6: Modus Ponens
8	$\bot$	5, 7: Modus Ponens

Field deems the Importation Principle $A \to (B \to C) \models (A \wedge B) \to C$) to be the main culprit for giving rise Curry's paradox. Accordingly, Field's theory invalidates the Importation Principle. Notice that the Importation Principle amounts

to the Contraction Principle $A \rightarrow (A \rightarrow C)\ A \rightarrow C$, if B is A. Hence, in Field's theory, the Importation Principle and the Contraction Principle are invalid:

- The Importation Principle is invalid: $A \rightarrow (B \rightarrow C) \not\models^F (A \wedge B) \rightarrow C$
- The Contraction Principle is invalid: $A \rightarrow (A \rightarrow C) \not\models^F A \rightarrow C$

4.2 Field's Determinacy Operator

4.2.1 Some Desiderata of The Determinacy Operator

Field introduces a determinacy operator D and uses it to characterize paradoxical sentences. He wants to characterize the Liar sentence by saying that it is not determinately true. So, if λ_0 is the Liar sentence, then $\neg D\lambda_0$. Yet, we can form a Strengthened Liar sentence 'This is not determinately true'.

Formally, the Strengthened Liar sentence can be represented by stipulating that

$$\lambda_1 =\!|\!\models^F \neg DT\langle\lambda_1\rangle$$

So what is the status of λ_1? Clearly, we cannot say λ_1 is not determinately true, that is, $\neg DT\langle\lambda_1\rangle$. Otherwise, we would have λ_1, entailing that λ_1 is true: $T\langle\lambda_1\rangle$. On the other hand, we cannot say that λ_1 is determinately true: $DT\langle\lambda_1\rangle$. Because we would have $\neg\lambda_1$ for saying so. This implies that λ_1 is not true: $\neg T\langle\lambda_1\rangle$.

However, according to Field, there is nothing wrong to say that λ_1 is not determinately determinately true: $\neg DDT\langle\lambda\rangle$ (or abbreviated as $\neg D^2T\langle\lambda_1\rangle$). Of course, we can always form a more strengthened paradoxical sentence; but the same strategy can be applied to characterize new paradoxical sentences.

The iteration of determinacy operators can be construed as a hierarchy of determinacy operators. The hierarchy of determinacy operators can be defined via transfinite induction. The base case and the successor case are straightforward. As for the limit case, we suppose that for all $\alpha < \lambda$ is well-defined. Then we define λ^{th} iteration of D as the operator which applies to the sentence A such that for all $\alpha < \lambda$, the result of prefixing the α^{th} iteration of D to the sentence A is true.

Field notices that we cannot have $v^u_{\mathcal{M}^+}(DDA) = v^u_{\mathcal{M}^+}(DA)$; otherwise we would have $v^u_{\mathcal{M}^+}(DDT\langle\lambda\rangle) = v^u_{\mathcal{M}^+}(DT\langle\lambda\rangle)$. This means that we cannot characterize λ_n by saying that $\neg D^2T\langle\lambda_n\rangle$, where $\lambda_n =\!|\!\models^F \neg DT\langle\lambda_n\rangle$. It is because $\neg D^2T\langle\lambda_n\rangle$ would entail λ_n. Call an operator O *idempotent* if OO is as same as O. To avoid revenge paradoxes, Field requires that:

i. The determinacy operator D is not idempotent: for any $A \in \mathcal{L}^+$, $v^u_{\mathcal{M}^+}(D^2A) \neq v^u_{\mathcal{M}^+}(DA)$.

Moreover, Field requires that the LEM fails for the determinacy operator D.

ii. For any $A \in \mathcal{L}^+$ and any β, $D^\beta A \vee \neg D^\beta A$ is invalid.

This is Field's paracomplete strategy for avoiding revenge paradoxes.

4.2.2 Defining the Determinacy Operator

Field defines the determinacy operator D as follow:

$$DA =_{df} A \wedge \neg(A \rightarrow \neg A)$$

Then, the operator D has the following properties:

(ia) if $v^u_{\mathcal{M}^+}(A) = 1$, then $v_{\mathcal{M}^+}(DA) = 1$.

(ib) if $v^u_{\mathcal{M}^+}(A) = 0$, then $v^u_{\mathcal{M}^+}(DA) = 0$.

(ib-s) if $v^u_{\mathcal{M}^+}(A) \preceq v^u_{\mathcal{M}^+}(\neg A)$, then $v^u_{\mathcal{M}^+}(DA) = 0$.

(ic) if $0 \prec v^u_{\mathcal{M}^+}(A) \prec 1$, then $v^u_{\mathcal{M}^+}(DA) \preceq v(A)$.

(ii) if $v^u_{\mathcal{M}^+}(A) \preceq v^u_{\mathcal{M}^+}(B)$, then $v^u_{\mathcal{M}^+}(DA) \preceq v^u_{\mathcal{M}^+}(DB)$.

These properties of the operator D ensures the following laws:

1. $\models^F DA \rightarrow A$
2. $A \models^F DA$
3. $\models^F A \rightarrow \neg A$, then $\models \neg DA$
   (Thus, (i) if $\models^F \neg A$, then $\models^F \neg DA$; and (ii) $\models^F \neg D\lambda$, where λ is the Liar sentence)
4. if $\models^F A \rightarrow B$, then $\models^F DA \rightarrow DB$

(1) is guaranteed by (ia), (ib) and (ic). (2) is guaranteed by (ia). (3) is guaranteed by (ib-s). (4) is guaranteed by (ii).

4.2.3 Showing the Determinacy Operator Satisfies The Desiderata

We can show that the determinacy operator D is non-idempotent and the LEM fails for D. We begin by considering the following sequence of paradoxical sentences:

$$\lambda_0 \eqslantless\models^F \neg T\langle\lambda_0\rangle$$
$$\lambda_1 \eqslantless\models^F \neg DT\langle\lambda_1\rangle$$
$$\lambda_2 \eqslantless\models^F \neg DDT\langle\lambda_2\rangle$$

and so on. Then, for any model $\mathcal{M}^+$, we have:

$$v^u_{\mathcal{M}^+}(\lambda_0) = v^u_{\mathcal{M}^+}(\neg T\langle\lambda_0\rangle)$$
$$v^u_{\mathcal{M}^+}(\lambda_1) = v^u_{\mathcal{M}^+}(\neg DT\langle\lambda_1\rangle)$$
$$v^u_{\mathcal{M}^+}(\lambda_2) = v^u_{\mathcal{M}^+}(\neg DDT\langle\lambda_2\rangle)$$

and so on.

Recall that the operator D has the following properties:

(ia) if $v^u_{\mathcal{M}^+}(A) = 1$, then $v_{\mathcal{M}^+}(DA) = 1$.

(ib) if $v^u_{\mathcal{M}^+}(A) = 0$, then $v^u_{\mathcal{M}^+}(DA) = 0$.

(ib-s) if $v^u_{\mathcal{M}^+}(A) \preceq v^u_{\mathcal{M}^+}(\neg A)$, then $v^u_{\mathcal{M}^+}(DA) = 0$.

(ic) if $0 \prec v^u_{\mathcal{M}^+}(A) \prec 1$, then $v^u_{\mathcal{M}^+}(DA) \preceq v(A)$.

(ii) if $v^u_{\mathcal{M}^+}(A) \preceq v^u_{\mathcal{M}^+}(B)$, then $v^u_{\mathcal{M}^+}(DA) \preceq v^u_{\mathcal{M}^+}(DB)$.

Making use of these properties, we can prove that the operator D is non-idempotent and does not obey the LEM. (The following proofs are tedious. Free free to skip them, as you will not miss anything important.)

Theorem 17 (Hierarchy Theorem)**.** Suppose that D^β is well-defined. Then for any $\alpha < \beta$,

(I) For any A, $v^u_{\mathcal{M}^+}(D^\beta A) \preceq v^u_{\mathcal{M}^+}(D^\alpha A)$.

(II) For some A, $v^u_{\mathcal{M}^+}(D^\beta A) \prec v^u_{\mathcal{M}^+}(D^\alpha A)$. (Specifically, $v^u_{\mathcal{M}^+}(D^\beta T\langle\lambda_\alpha\rangle) = 0$ but $v^u_{\mathcal{M}^+}(D^\alpha T\langle\lambda_\alpha\rangle) \succ 0$.)

Proof. We can show (I) by transfinite induction on β. The base case is trivial. The successor case and the limit case are proved by applying (ia), (ib) and (ic).

To show (II), we first show $v^u_{\mathcal{M}^+}(D^\alpha T\langle\lambda_\alpha\rangle) \succ 0$. By (ia) and the identity of truth, if $v^u_{\mathcal{M}^+}(\lambda_\alpha) = 1$, then for every α, $v^u_{\mathcal{M}^+}(D^\alpha T\langle\lambda_\alpha\rangle) = 1$. According to the semantics of $\neg$, $v^u_{\mathcal{M}^+}(\neg D^\alpha T\langle\lambda_\alpha\rangle) = 0$. Hence, if $v^u_{\mathcal{M}^+}(\lambda_\alpha) = 1$, then $v^u_{\mathcal{M}^+}(\lambda_\alpha) \neq v^u_{\mathcal{M}^+}(\neg D^\alpha T\langle\lambda_\alpha\rangle)$. But we have $v^u_{\mathcal{M}^+}(\lambda_\alpha) = v^u_{\mathcal{M}^+}(\neg D^\alpha T\langle\lambda_\alpha\rangle)$. Thus, $v^u_{\mathcal{M}^+}(\lambda_\alpha) \neq 1$. Since $v^u_{\mathcal{M}^+}(\lambda_\alpha) = v^u_{\mathcal{M}^+}(\neg D^\alpha T\langle\lambda_\alpha\rangle)$, it means that $v^u_{\mathcal{M}^+}(\neg D^\alpha T\langle\lambda_\alpha\rangle) \neq 1$. Then, according to the semantics of $\neg$, $v^u_{\mathcal{M}^+}(D^\alpha T\langle\lambda_\alpha\rangle) \neq 0$. That is, $v^u_{\mathcal{M}^+}(D^\alpha T\langle\lambda_\alpha\rangle) \succ 0$.

Then, we show that if $\alpha < \beta$, then $v^u_{\mathcal{M}^+}(D^\beta T\langle\lambda_\alpha\rangle) = 0$. We can show this by transfinite induction on β. The successor case: $\beta = \delta + 1$. There are two sub-cases.

Case 1: $\alpha < \delta$. Our induction hypothesis is that $v^u_{\mathcal{M}^+}(D^\delta T\langle\lambda_\alpha\rangle) = 0$. By (ib), we have $v^u_{\mathcal{M}^+}(D^{\delta+1} T\langle\lambda_\alpha\rangle) = 0$.

Case 2: $\alpha = \delta$. By the result (I), we have $v^u_{\mathcal{M}^+}(D^\delta T\langle\lambda_\delta\rangle) \preceq v^u_{\mathcal{M}^+}(T\langle\lambda_\delta\rangle)$. By the equation of λ_δ, $v^u_{\mathcal{M}^+}(D^\delta T\langle\lambda_\delta\rangle) \preceq v^u_{\mathcal{M}^+}(\neg D^\delta T\langle\lambda_\delta\rangle)$. Then by (ib-s), we have $v^u_{\mathcal{M}^+}(D^{\delta+1} T\langle\lambda_\delta\rangle) = 0$.

The limit case: given that β is a limit, then $\beta > \alpha + 1$. Our induction hypothesis is that $v^u_{\mathcal{M}^+}(D^{\alpha+1} T\langle\lambda_\alpha\rangle) = 0$. Thus, by the result (I), it must be $v^u_{\mathcal{M}^+}(D^\beta T\langle\lambda_\alpha\rangle) = 0$. □

We have shown that the determinacy operator D is not idempotent as desired. We call this result the Hierarchy Theorem. This Theorem entails that the LEM does not hold for the determinacy operator D.

Corollary 18. For any β for which D^β is well-defined, there are some sentences A for which $v^u_{\mathcal{M}^+}(D^\beta A \vee \neg D^\beta A) \prec 1$ for any model $\mathcal{M}^+$.

Proof. As we saw in the above proof, for any model $\mathcal{M}^+$, there are some A such that $v^u_{\mathcal{M}^+}(D^{\beta+1} A) = 0$ but $v^u_{\mathcal{M}^+}(D^\beta A) \succ 0$. (For instance, $A = T\langle\lambda_\beta\rangle$). According to the semantics of $\neg$, since $v^u_{\mathcal{M}^+}(D^\beta A) \succ 0$, it means that $v^u_{\mathcal{M}^+}(\neg D^\beta A) \neq 1$; since,

clearly, we have $v^u_{\mathcal{M}^+}(D^\beta A) \neq 1$ as well. Thus, according to the semantics of $\vee$, it follows that $v^u_{\mathcal{M}^+}(D^\beta A \vee \neg D^\beta A) \prec 1$. □

4.3 Expressive Limitations and Natural Languages

4.3.1 Bivalent Determinataness

Field claims that his theory is revenge-immune; because he has offered a consistency proof for his theory. However, Field's theory avoids paradoxes, only because certain intuitively appealing notions are not allowed to express in the theory. One particular notion is bivalent determinateness.

Let $F^{\mathbb{S}}$ be Field's theory augmented with an operator $\mathbb{S}$ intended to represent bivalent determinateness. The semantics of the operator $\mathbb{S}$ can be defined as follows:

- $v^u_{\mathcal{M}^{+\mathbb{S}}}(\mathbb{S}A) = \begin{cases} 1, \text{ if } v^u_{\mathcal{M}^{+\mathbb{S}}}(A) = 1. \\ 0, \text{ if } v^u_{\mathcal{M}^{+\mathbb{S}}}(A) \neq 1. \end{cases}$

Field calls this *super-determinate truth*. Hence, we use $\mathbb{S}$ to represent the notion.

Clearly, the operator $\mathbb{S}$ is bivalent. For any sentence $A \in \mathcal{L}^+$, either A takes the ultimate value 1 or not. In the former case, since $\mathbb{S}A$ takes the ultimate value 1, we have $\mathbb{S}A \vee \neg\mathbb{S}A$. In the latter case, since $\neg\mathbb{S}A$ takes the ultimate value 1, we have $\mathbb{S}A \vee \neg\mathbb{S}A$ as well.

Also, the operator $\mathbb{S}$ is idempotent. Suppose that $\mathbb{S}A$ takes the value ultimate 1. Then, $\mathbb{S}\mathbb{S}A$ takes the ultimate value 1 as well. Conversely, suppose that $\mathbb{S}\mathbb{S}A$ takes the ultimate value 1. Then, $\mathbb{S}A$ takes the ultimate value 1.

Notice that we have the following laws:

(1) $\models^{F^{\mathbb{S}}} \mathbb{S}A \to A$

(2) $A \models^{F^{\mathbb{S}}} \mathbb{S}A$

(2_w) $A\,, \neg\mathbb{S}A \models^{F^{\mathbb{S}}} \bot$

(3) $\models^{F^{\mathbb{S}}} \mathbb{S}A \to DA$

(3_w) $A \to \neg A \models^{F^{\mathbb{S}}} \neg\mathbb{S}A$

(4) $\models^{F^{\mathbb{S}}} \mathbb{S}(A) \to \mathbb{S}\mathbb{S}(A)$

(4_w) $\neg\mathbb{S}\mathbb{S}A \models^{F^{\mathbb{S}}} \neg\mathbb{S}A$

In the previous chapter, we saw that Field argues that the extension of model-independent truth and the extension of having the value 1 are not the same. For this reason, Field concludes that one cannot read off the model-theoretic semantics that there is a notion of bivalent determinateness. But we argue that bivalent determinateness needs not be motivated by the properties of the model-theoretic semantics. Nevertheless, Field has another reason against bivalent determinateness. He argues that the operator $\mathbb{S}$ is unintelligible, because the operator $\mathbb{S}$ leads to inconsistency (and triviality):

> Well, we can postulate these things all we like, just as we can postulate the naive truth theory together with classical logic; postulation is no guarantee against inconsistency. And it is easy to see that these postulates are inconsistent. (Field, 2008, p. 344)

Let ξ be equivalent to $\neg\mathbb{S}T\langle\xi\rangle$. Since we have $\models^{F^{\mathbb{S}}} T\langle\xi\rangle \rightarrow \xi$, it follows that $\models^{F^{\mathbb{S}}} T\langle\xi\rangle \rightarrow \neg\mathbb{S}T\langle\xi\rangle$. By (1), we have $\models^{F^{\mathbb{S}}} \mathbb{S}T\langle\xi\rangle \rightarrow T\langle\xi\rangle$. Thus, $\models^{F^{\mathbb{S}}} \mathbb{S}T\langle\xi\rangle \rightarrow \neg\mathbb{S}T\langle\xi\rangle$. By (3_w), we have $\models^{F^{\mathbb{S}}} \neg\mathbb{S}\mathbb{S}T\langle\xi\rangle$. By (4_w), we have $\models^{F^{\mathbb{S}}} \neg\mathbb{S}T\langle\xi\rangle$. By the initial supposition, it follows that $\models^{F^{\mathbb{S}}} \xi$. By (2_w), $\models^{F^{\mathbb{S}}} \xi$ and $\models^{F^{\mathbb{S}}} \neg\mathbb{S}T\langle\xi\rangle$ gives us $\models^{F^{\mathbb{S}}} \bot$.

Field insists that we do not need the operator $\mathbb{S}$ to characterize paradoxical sentences as defective. Rather, his determinacy operator D can be used to characterize paradoxical sentences.

> [T]he claim that I dispute is that the model theory ought to allow for a super-determinateness operator meeting intuitive preconceptions. I've argued that this is not so, that in fact such an operator doesn't really make sense (though I grant that this is initially quite surprising); but Priest's view seems to be that this position is one I can't coherently hold, because (he thinks) the notion is presupposed by my own theory, or by the motivation for it.
>
> But why? The motivation of the theory does require that the ordinary Liar sentence $[\lambda_0]$ is in some sense defective, but I can say that: it is defective in the most straightforward sense of being neither determinately true nor determinately false. (It is defective$_1$). The motivation also requires that the "next level Liar" $[\lambda_1]$ is in some sense defective. It needn't be defective in quite the same sense: perhaps it is "defective" only in the extended sense that the claim that it isn't defective (in the first sense) is defective (in the first sense). That's what the theory yields: $[\neg DDT\langle\lambda_1\rangle]$ so [defective$_1$ (defective$_1 T\langle\lambda_1\rangle$]. But we can introduce a broader sense of defectiveness, defectiveness$_2$, that includes both this "second order defectiveness" and defectiveness$_1$; in this broader sense, $[\lambda_0]$ and $[\lambda_1]$ are both defective. (Field, 2008, p. 357)

However, it seems that Field's argument is just that there is no way to define super-determinateness in his theory. Simmons comments on this point:

> [I]t is a distinct drawback of any solution to paradox if it is forced to deny the intelligibility of notions that appear quite intelligible to us, especially if the main motivation for the denial is to protect one's theory from the threat of paradox. (Simmons, 2008, P. 160)

4.3.2 Truth-Value Gaps and Exclusion Negation

Truth-Value Gaps. Another important notion for which Field's theory does not allow is a general notion of truth-value gaps. In Field's theory, a hierarchy of gappiness can be so defined:

$$G^0A =_{df} \neg T\langle A\rangle \wedge \neg F\langle A\rangle$$
$$G^1A =_{df} \neg DT\langle A\rangle \wedge \neg DF\langle A\rangle$$
$$G^2A =_{df} \neg D^2T\langle A\rangle \wedge \neg D^2F\langle A\rangle$$

and so on, where the falsity predicate F is defined as the truth of negation (i.e., $F\langle A\rangle$ iff $T\langle\neg A\rangle$). However, Field's theory cannot be augmented with a general notion of truth-value gaps defined as follows:

- $v^u_{\mathcal{M}^{+\mathbb{S}}}(\mathbb{G}A) = \begin{cases} 1, \text{ if } v^u_{\mathcal{M}^{+\mathbb{S}}}(A) \neq 1 \text{ and } v^u_{\mathcal{M}^{+\mathbb{S}}}(A) \neq 0. \\ 0, \text{ otherwise.} \end{cases}$

Let ξ be intersubstitutable for $\mathbb{G}\xi \vee F\xi$. If ξ takes the ultimate value 1, then $\mathbb{G}\xi \vee F\xi$ takes the ultimate value 0. Because both $\mathbb{G}\xi$ and $F\xi$ takes the ultimate value 0. If ξ takes the ultimate value 0, then $\mathbb{G}\xi \vee F\xi$ takes the ultimate value 1. Because $F\xi$ takes the ultimate value 1. If ξ takes some intermediate value , then $\mathbb{G}\xi \vee F\xi$ takes the ultimate value 1. Because $\mathbb{G}\xi$ takes the ultimate value 1. In any case, ξ cannot be intersubstitutable for $\mathbb{G}\xi \vee F\xi$, violating the initial assumption.

As Simmons (2018) points out, the notion of truth-value gaps has certain interactions with many linguistic phenomena:

> Field cannot allow exclusion negation or the full notion of a gap while maintaining naïve truth and the Intersubstitutivity Principle, and he denies the intelligibility of these notions. This is counter to our ordinary semantic usage. The notion of a truth-value gap drives Kripke's own presentation of his theory, and it naturally arises not just in connection with the Liar, but also in connection with, for example, vagueness, presupposition failures and category mistakes. (Simmons, 2018, p. 164)

Field does not agree with some linguistic analyses that make use of truth-value gaps.

> In my opinion there is little reason to believe [that many or all sentences with non-denoting singular terms are neither true nor false.]: the intuitions that seem at first to support it are best explained pragmatically (Stalnaker 1974), and by far the simplest systematic theory governing all sentences with non-denoting terms (including, e.g. negative existentials) is one that takes those sentences to have exactly one of the two standard truth values (putting aside the special issues about certain sentences containing predicates like 'true'). (Field, 2008, p.206)

But Field still admits that some (non-declarative) sentences are neither true nor false.

> Are there sentences that are neither true nor false? Sure: there are sentences in nonindicative moods (questions, commands, and so forth), and there are sentences that are by ordinary criteria meaningless. (Also orthographic sentence-types that are ambiguous or require a context to fix what they say on a given occasion.) (ibid)

In any case, disallowing the general notion of truth-value gaps seems to be undesirable. For one thing, such a notion cannot be dismissed as unintelligible. For another thing, it is often used in linguistic analyses. Hence, Field's theory is not available to those linguists who make use of truth-value gaps to characterize (declarative) sentences. (For more details on the conceptual issues concerning truth-value gaps, see Shaw 2014.)

Exclusion Negation. The notion of truth-value gaps also has a close connection with the use of negation. Given the presence of truth-value gaps, it is natural to infer from the claim that if A is neither true nor false to the claim that A is not true. Even if we admit that there is a hierarchy of gappiness, it is natural to infer from the claim that λ_n is gappy$_n$ to the claim λ_n is not true simpliciter. Formally, the above uses of 'not' can be defined as follows:

- $v^u_{\mathcal{M}^{+\mathbb{S}}}(\sim A) = \begin{cases} 1, \text{ if } v^u_{\mathcal{M}^{+\mathbb{S}}}(A) \neq 1 \\ 0, \text{ if } v^u_{\mathcal{M}^{+\mathbb{S}}}(A) = 1 \end{cases}$

The operator $\sim$ is known as exclusion negation. However, Field's theory cannot be augmented with $\sim$. To see this, let ξ be equivalent to $\sim\xi$. If ξ takes the ultimate value 1, then, according to the semantics of $\sim$, $\sim\xi$ takes the ultimate value 0. If ξ does not take the ultimate value 1, then, according to the semantics of $\sim$, $\sim\xi$ takes the ultimate value 1. Either way, ξ cannot be equivalent to $\sim\xi$, violating the initial supposition.

Field seems to think that his opponents have the burden of proof to show that exclusion negation is intelligible. Moreover, Field thinks such arguments do not work. For instance, he charges Hahn's argument with circular reasoning.

> One old argument (Hahn 1933) starts from the idea that we can stipulate that an operator NEG obeys the following truth rule: for any sentence x, NEG(x) is true if and only if x is not true. From this stipulation (the argument goes) we can logically derive that exactly one of x and NEG(x) is true (as well as that NEG(NEG(x)) is true if and only if x is true, and so forth); and this (when combined with similar stipulations for the other connectives) will make the usual Boolean laws such as excluded middle come out true. Here the proper response is not to deny the legitimacy of the stipulation; rather, the proper response is that from the stipulation, one can derive such claims as that exactly one of x and NEG(x) is true *only if we assume Boolean laws for the 'not' used in making the stipulation*. If, for instance, one doesn't assume excluded middle for 'not', then there is no way to derive from the stipulation that either x or NEG(x) is true. (Field, 2008, p. 310)

However, it seems unfair to put the burden of proof on the party who advocates that exclusion negation is coherent, and accuse their arguments of circular. It is because it seems that there is no way to argue for the intelligibility of some fundamental logical notions without assuming the notion in question.

Field also remarks that his theory obeys the law for exclusion negation:

(E) $\neg_E A$ is true iff A is not true,

but not the law for choice negation:

(C) $\neg_C A$ is true iff A is not true and not gappy.

It is just that the 'not' on the right hand side of (E) does not obey excluded middle in Field's theory:

> The arguments that exclusion negation as defined via (E) leads to inconsistency with naive truth all turn [...] on assuming that the 'not' used on the right hand side of (E) obeys excluded middle. (Field, 2008, p. 311)

Despite Field's remark, we saw that Field's theory cannot be extended with the negation $\sim$, where $\sim A$ always takes a classical value.

In any case, it seems that the ways we use 'not' should be interpreted as exclusion negation in most cases. Scharp (2013) cites some linguistic evidence for that 'not' in English should be as construed as exclusion negation:

> [L]inguists claim that 'not' in English (at least sometimes) is properly interpreted as exclusion negation, and linguists use exclusion negation in their theories. Here are two examples.
>
> Jay Atlas in Philosophy without Ambiguity (1989) argues that 'not' has a general sense and on particular occasions of use it can express either choice negation or exclusion negation. There is linguistic evidence that 'not' is univocal and invariant because it fails ambiguity tests and context-dependence tests; thus, it is neither ambiguous nor context-dependent. Nevertheless, on many occasions, it makes the most sense to interpret English speakers as meaning exclusion negation when they use 'not'. A second example is that Laurence Horn in A Natural History of Negation (2001) surveys views on negation from Aristotle to present, the evidence for choice negation readings of 'not' vs. exclusion negation readings of 'not', and how these readings interact with other linguistic phenomena (presupposition, conversational implicature, scope, etc.). He too argues that 'not' is not ambiguous or context-dependent. Rather, exclusion negation provides the semantics for natural language descriptive (non-metalinguistic) negation or predicate denial (in Aristotle's sense), and what seems like choice negation is an artifact of pragmatic tendencies like that of reading topical/definite subjects as taking wide scope with respect to ordinary predicate denial. (ibid, p. 110)

Scharp notices that Field might reply that negation fails to obey excluded middle, because of paradoxical sentences:

> I suppose Field could say that the evidence cited by Atlas and by Horn is compatible with English having some kind of expression that behaves like exclusion negation in their examples, but which fails to obey excluded middle in paradoxical settings. Of course, to be convincing, he would have to find some kind of independent evidence to support this claim. However, instead of developing this line of

> thought, it might be better to just stop banging on a square peg and recognize that the hole is round. (ibid)

In any case, Field's hierarchy of determinateness, as he confesses, is merely motivated by the intuition that paradoxical sentences are in some sense defective:

> [A]lthough it would be incorrect to say that the Liar sentence is not true, in the sense of 'not' used in the Liar sentence, it seems like there ought to be some "weaker form of negation" 'wnot' in which we can correctly say that it is "wnot true". (This seems plausible even if one thinks that 'not' in English is unambiguous, so that the "weaker sense of negation" is not a legitimate reading of the English 'not'.) A natural way to express this in English would be to say that the Liar sentence is not determinately true. (Field, 2008, p.73)

Apart from that, the hierarchy itself seems to be artificial and has little connection to natural languages. On the other hand, we have plenty of evidence for interpreting 'not' as exclusion negation.

In summary, Field's theory leaves us an impression that it is too remote from natural languages. It attributes ordinary speakers artificial notions of determinateness and defectiveness, and it is saved from paradoxes by disallowing some commonly used notions.

4.4 Conclusion

In this chapter, we argued that Field's theory has little connection with natural languages. Field's solution to the Liar's revenge attributes to ordinary speakers technically and carefully constructed notions that are defined by the iteration of the determinacy operator. It is one thing to say that we can avoid paradoxes by adopting the hierarchy of determinacy operators. It is another thing to say that this is how ordinary speakers can avoid semantic paradoxes.

Chapter 5

Paraconsistent Dialetheism and Strict-Tolerant Dialetheism

In this chapter, we discuss two different dialetheic approaches to truth. These approaches are the paraconsistent approaches and the strict-tolerant approaches, both of which take the Liar sentence as both true and false. Specifically, we will discuss the transparent theory of truth of both approaches: $LPTT$ (Logic of Paradox with Transparent Truth) and $STTT$ (Strict-Tolerant Transparent Truth).

We will extend both $LPTT$ and $STTT$ with a *Just True* operator: the resulting theories are called $LPTT^{\mathbb{J}}$ and $STTT^{\mathbb{J}}$ respectively. We will show that $LPTT^{\mathbb{J}}$ can resist the revenge argument which uses the material biconditional to represent the self-referential character of the Strengthened Liar sentence. Yet, $LPTT^{\mathbb{J}}$ cannot deal with the revenge argument which uses the semantic equivalence to mimic self-reference. On the other hand, $STTT^{\mathbb{J}}$ can resist the revenge via the material biconditional, as well as the revenge via the semantic equivalence.

5.1 Dialetheism

Dialetheism, the view that there are true contradictions, has been defended as a solution of the Liar paradox. In what follows, we review the arguments for the dialetheic solutions given by Priest and Beall.

Revenge Immunity and Semantic Closure. According to Priest, the recurring revenge paradoxes show that the original Liar paradox is just an instance of a more general phenomenon. In general, all sentence can be divided into two different sets: the bona fide truths and its complement. The Strengthened Liar sentence is essentially a sentence that says of itself that it is the Rest. Using some familar reasoning, we can generate a paradox: if the Strengthened Liar sentence is in the bona fide truths, then according to what it says, it is in the Rest; if the Strengthened Liar sentence is in the Rest, then what it says is exactly true. So the Strengthened Liar sentence is in the bona fide truths.

To get rid of the original paradox, theorists usually posit a new category: the Deflective (e.g. truth-value gaps, the unassertable, the unstable, etc.) Such a strategy for solving the original paradox works, only because the original problem is not

properly formulated. Once the problem is formulated in the most general form, it is fruitless to posit the Deflective; since the Deflective is a proper part of the Rest.

Facing the new paradoxes, there are three options. The first option is to admit that the notion in the proposed theory gives rise to a contradiction again, just as the notion of truth gives rise to a contradiction in the first place. The second option is to admit that the notion is inexpressible in the language $\mathcal{L}$ – the formal language that is intended to serve as a heuristic model of natural languages; but insist that the notion is expressible in another formal language $\mathcal{L}'$. The third option is to claim that the notion in question is not a real notion.

The strategy theorists usually employ is a mixture of the second and the third option. Such a strategy consists in confessing that the notion in question is not expressible in the object language; but insisting that the proposed theory is formulated in the metalanguage: the notion is not a real notion, but a model-dependent notion.

Priest argues against this strategy. He agrees with Tarski that our natural languages are semantically closed: semantic terms can be applied to sentences in natural languages. Semantical notions such as truth-value gaps are expressible in natural languages. If the theorists deny that those semantical notions are meaningful notions, this simply amounts to a self-refutation. Accordingly, Priest thinks that, to deal with the Liar paradox and its related semantic paradoxes, we should dispense with the distinction between object-language and metalanguage, and embrace inconsistency.

Methdological Deflationism. Beall (2009) holds that the truth predicate is a constructed device that allows us to express generalizations we could not express otherwise. In particular, the truth predicate is introduced so that it obeys the transparent truth principle (TT) and the following formulation of intersubstitutivity principle:

- Let B be any sentence in which A occurs. Then the result of substituting $T\langle A\rangle$ for any occurrence of A in B has the same semantic value as B.

The only reason for introducing the transparent truth predicate is to allow us to express infinite conjunctions in a finitary language. Since we have only finite amount of time and capacity, we cannot assert infinitely many sentences. But the transparent truth predicate allows us to express claims like this:

(17) All of the infinitely many axioms of Peano arithmetic are true.

However, our transparent truth predicate has some 'spandrels' – inevitable, and unintended, by-products. In particular, given the presence of transparent truth predicate, we must have the consequence that the Liar sentence is both true and false. Beall accepts the classical laws used in the Liar argument. In particular, he accepts that negation is exhaustive: he claims that the essential role of negation is to divide our sentences into the true and the false (Beall, 2009, p. 4). And according to Beall, to accept the exhaustive nature of negation is to accept the LEM. Given the LEM and reasoning by cases, the transparent truth principle ensures that the Liar sentence is both true and false.

Dialetheic Logics of Truth. Currently, there are two different kinds of dialetheic logics of truth: the paraconsistent theories and the strict-tolerant theories.

The distinctive feature of the paraconsistent theories is that the ECQ (i.e., $A, \neg A \models B$) is invalid in such logics. Specifically, Priest and Beall make use of the Logic of Paradox LP to develop their logics of truth. Because of the failure of the ECQ, the dialetheists who adopts the paraconsistent approaches are not committed to accepting triviality. Priest's logic validates the T-schema but rejects the transparency principle (TT); whereas Beall's logic validates both the T-schema and TT. For simplicity, we will focus on Beall's logic $LPTT$ in what follows.

As for the strict-tolerant theories, the most crucial characteristic of these theories is their non-transitive consequence relation: a strict-tolerant ST valid argument is an argument such that every premise is strictly true (i.e., every premise takes the value 1) but some conclusions are at least tolerantly true (i.e., some conclusions take a value that is greater than 0). Such a consequence relation has a significant implication that helps dealing with the Liar paradox. Specifically, the following principle does not hold in the ST theories:

- Validity-Preservation (VP): if $\models A$ and $A \models B$, then $\models B$

Thus, according to the ST analysis, even if the Liar argument begins with valid premises, and every step in the argument is valid, it does not ensure that the conclusion – an arbitrary sentence – is valid.

This chapter is organized as follows. In §5.2, we will discuss the formal details of the paraconsistent approaches. In particular, we will focus on Beall's theory of transparent truth $LPTT$ and his recent defence against the charge that $LPTT$ does not have a detachable conditional. In §5.3, we will look at the strict-tolerant theory of transparent truth $STTT$. Our main discussion will be in §5.4. Firstly, we will see that $LPTT$ and $STTT$ has some expressive limitations. In particular, if the claim that A is just true is formalized as $T\langle A\rangle \wedge \neg T\langle \neg A\rangle$, it cannot get the desired interpretation. Secondly, we will extend both $LPTT$ and $STTT$ with a *Just True* operator $\mathbb{J}$ and discuss whether the extended theories $LPTT^{\mathbb{J}}$ and $STTT^{\mathbb{J}}$ are plagued by revenge paradoxes.

In the discussion of revenge paradoxes, we distinguish two kinds of revenge arguments, corresponding to two different ways to represent the self-referential character of sentences. According to the strong self-referential procedure, we can represent self-reference by requiring that the constant symbol l denotes Al. Alternatively, we can also require the constant symbol l to be identical to the term $\langle Al\rangle$. Then, in both cases, Al is a self-referential sentence. According to the weak self-referential procedure, we can represent self-reference by requiring that a sentence D to be equivalent to a sentence $A\langle D\rangle$ which talks about D. Accordingly, the revenge arguments against our extended theories $LPTT^{\mathbb{J}}$ and $STTT^{\mathbb{J}}$ can be distinguished into two kinds: the revenge via strong procedures and the revenge via weak procedures. We show that $LPTT^{\mathbb{J}}$ and $STTT^{\mathbb{J}}$ are plagued by the revenge via strong procedures.

Then, we explore the option of giving up the strong self-referential procedure. Specifically, we further distinguish the revenge via weak procedures into the one that uses the material biconditional $\equiv$ and the one that uses the semantic equiv-

alence $=\!|\!=$. It is because the question of whether A is semantically equivalent to B cannot always be reduced to the question that A is materially equivalent to B. We will show that the semantic equivalence is a stronger notion than the material biconditional in $LPTT^{\mathbb{J}}$; whereas the semantic equivalence and the material biconditional are equally strong in $STTT^{\mathbb{J}}$. Most importantly, we will show that $LPTT^{\mathbb{J}}$ and $STTT^{\mathbb{J}}$ can resist the revenge via the material biconditional; but only $STTT^{\mathbb{J}}$ can resist the revenge via the semantic equivalence.

5.2 The Paraconsistent Approaches

5.2.1 The Logic of Paradox

The Logic of Paradox LP is similar to Strong Kleene Logic K_3. It is a three-valued logic, with the value 1 and $\frac{1}{2}$ being designated:

- the set of designated value $\mathcal{D}^+ = \{1, \frac{1}{2}\}$.

In LP, the value $\frac{1}{2}$ is construed as representing a truth-value glut. As usual, the value 1 and 0 are construed as truth and falsity respectively.

Suppose that we have a base language $\mathcal{L}$ without any truth predicate. Then, the semantics of $\mathcal{L}$ is interpreted by LP models.

Definition 19 (LP model)**.** LP interpretations are models $\langle \mathcal{D}, \mathcal{I} \rangle$, where:

(i) $\mathcal{D} \neq \emptyset$;

(ii) $\mathcal{I}$ is an interpretation function such that

- $\mathcal{I}$ assigns each individual constant (i.e., name) an object of $\mathcal{D}$. That is, for all $a \in Con$, $\mathcal{I}(a) \in \mathcal{D}$, where Con is a set of individual constant.
- $\mathcal{I}$ assigns to each n-ary function symbol an object of $\mathcal{D}$. That is, for all $f \in Func_n$, $\mathcal{I}(f) : \mathcal{D}^n \mapsto \mathcal{D}$, where $Func_n$ is a set of n-place function symbols for all n.
- $\mathcal{I}$ assigns a pair $\langle \mathcal{I}^+, \mathcal{I}^- \rangle$ to each all n-ary predicate P. $\mathcal{I}^+$ assigns to each n-ary predicate a set of n-tuples which is the extension of P; whereas $\mathcal{I}^-$ assigns to each n-ary predicate a set of n-tuples which is the anti-extension of P: That is, for all $P \in Pred_n$, $\mathcal{I}^+(P) \subseteq \mathcal{D}^n$ and $\mathcal{I}^-(P) \subseteq \mathcal{D}^n$, where $Pred_n$ is the set of n-place predicate symbols for all n.
- LP models are just like K_3 models, except that $\mathcal{I}$ is constrained by this: $\mathcal{I}^+(P) \cup \mathcal{I}^-(P) = \mathcal{D}$, instead of $\mathcal{I}^+(P) \cap \mathcal{I}^-(P) = \emptyset$.
  (This condition makes sure that $Pa_0, ..., a_n$ and its negation exhausts the domain.)

Parenthetical Remark. Beall (2009) insists that there is no glut in the base language. To follow Beall's philosophical position, one has to impose the constraint $\mathcal{I}^+(P) \cap \mathcal{I}^-(P) = \emptyset$.

As for the valuation scheme, we have:

Definition 20 (LP Valuation Scheme)**.** The LP valuation scheme is as same as the K_3 valuation scheme, except that:

- For an atomic formula $Pa_0, ..., a_n$, we have:

$$v_{\mathcal{M}}(Pa_0, ..., a_n) = \begin{cases} 1, & \text{if } \langle \mathcal{I}(a_0), ..., \mathcal{I}(a_n) \rangle \in \mathcal{I}^+(P) \text{ and} \\ & \langle \mathcal{I}(a_0), ..., \mathcal{I}(a_n) \rangle \notin \mathcal{I}^-(P) \\ 0, & \text{if } \langle \mathcal{I}(a_0), ..., \mathcal{I}(a_n) \rangle \notin \mathcal{I}^+(P) \text{ and} \\ & \langle \mathcal{I}(a_0), ..., \mathcal{I}(a_n) \rangle \in \mathcal{I}^-(P) \\ \frac{1}{2}, & \text{if } \langle \mathcal{I}(a_0), ..., \mathcal{I}(a_n) \rangle \in \mathcal{I}^+(P) \cap \mathcal{I}^-(P) \end{cases}$$

5.2.2 Transparent Truth

The Kripke Construction. By using the Kripke construction, we can expand LP models to interpret the predicate T in the language $\mathcal{L}^+$, where $\mathcal{L}^+$ is the base language $\mathcal{L}$ extended with a truth predicate T and allows for self-reference. The new models $\mathcal{M}^+ = \langle \mathcal{D}, \mathcal{I}, \mathcal{T} \rangle$ generated by the Kripke construction is that:

Definition 21 (LP^+ model)**.** A LP^+ model is a structure $\mathcal{M}^+ = \langle \mathcal{D}, \mathcal{I}, \mathcal{T} \rangle$ such that

(i) $\mathcal{D}$ is the domain.

(ii) $\mathcal{I}$ is the same as $\mathcal{I}$ in the LP models.

(iii) $\mathcal{T}$ is is a pair $\langle \mathcal{T}^+, \mathcal{T}^- \rangle$, such that

- $\mathcal{T}^+$ assigns a set of objects to the extension of the truth predicate $\mathcal{T}^+ \subseteq \mathcal{D}$;
- $\mathcal{T}^-$ assigns a set of objects to the anti-extension of the truth predicate $\mathcal{T}^- \subseteq \mathcal{D}$.
- $\mathcal{T}^+ \cup \mathcal{T}^- = \mathcal{D}$.
- For any $\mathcal{T}^+, \mathcal{T}^- \in \mathcal{D}$, $\langle \mathcal{T}^+, \mathcal{T}^- \rangle = \langle \mathcal{T}^+_{\mathcal{M}^+}, \mathcal{T}^-_{\mathcal{M}^+} \rangle$, where

  $\mathcal{T}^+_{\mathcal{M}^+}$ is the set of (codes of) true sentences of $\mathcal{M}^+$, and

  $\mathcal{T}^-_{\mathcal{M}^+}$ is the set of objects in $\mathcal{D}$ such that either the objects are not sentences of $\mathcal{L}^+$, or are (codes of) false sentence of $\mathcal{M}^+$.

LP^+ models have the following properties:

- For any $\mathcal{M}^+$, $v_{\mathcal{M}^+}(T\langle A \rangle) = v_{\mathcal{M}^+}(A)$. (Identity of Truth)
- For any model $\mathcal{M}$ of $\mathcal{L}$, the new model $\mathcal{M}^+$ agrees with $\mathcal{M}$ in its interpretations on the language $\mathcal{L}$. (Model Extension)

LPTT. Recall that consequence is defined in terms of countermodel. Like K_3TT, paraconsistent approaches also rely on the principle of designated value preservation to define countermodels: a countermodel to an argument from Γ to Δ is a model which assigns a designated value to every member of Γ, but assign a non-designated value to every member of Δ. Accordingly, for the paraconsistent approaches, we have:

- A LP^+ model is a countermodel to an argument from the premises Γ to conclusions Δ iff it assigns 1 or $\frac{1}{2}$ to every premise of Γ and 0 to every conclusion of Δ.

Hence, consequence is defined as follow:

Definition 22 ($LPTT$ Consequence)**.** $\Gamma \models^{LPTT} \Delta$ iff if $v_{\mathcal{M}^+}(A) \geq \frac{1}{2}$ for all $A \in \Gamma$, then $v_{\mathcal{M}^+}(B) \geq \frac{1}{2}$ for some $B \in \Delta$.

We thereby obtain a logic that is closed under TT: A and $T\langle A\rangle$ are intersubstitutable in any extensional context. It is because A and $T\langle A\rangle$ are assigned the value in any LP^+ models $\mathcal{M}^+$. We call the resulting logic $LPTT$. The key feature of $LPTT$ is that the ECQ fails. That is, $A, \neg A \not\models^{LPTT} B$ and $A \wedge \neg A \not\models^{LPTT} B$.

The Liar Paradox. Now let's see where the Liar argument goes wrong. The Liar argument can be represented as follows:

1	$T\langle\lambda\rangle \vee \neg T\langle\lambda\rangle$	LEM
2	$T\langle\lambda\rangle$	Assumption
3	λ	2: Release
4	$\neg T\langle\lambda\rangle$	3: The Def of λ
5	$T\langle\lambda\rangle \wedge \neg T\langle\lambda\rangle$	2, 4: $\wedge$-Intro
6	$\neg T\langle\lambda\rangle$	Assumption
7	λ	6: The Def of λ
8	$T\langle\lambda\rangle$	7: Capture
9	$T\langle\lambda\rangle \wedge \neg T\langle\lambda\rangle$	6, 8: $\wedge$-Intro
10	$T\langle\lambda\rangle \wedge \neg T\langle\lambda\rangle$	1, 5, 9: Reasoning by Cases
11	B	10: ECQ

To begin with, notice that the LEM is valid in $LPTT$. Suppose for reductio that $\not\models^{LPTT} A \vee \neg A$. It means that $A \vee \neg A$ takes the value 0. By the semantics of $\vee$, it follows that both A and $\neg A$ take the value 0. By the semantics of $\neg$, since $\neg A$ takes the value 0, it follows that A takes the value 1. So A takes the value 1 and the value 0. Yet, this is impossible. Hence, the value of $A \vee \neg A$ cannot be 0. Thus, $\models^{LPTT} A \vee \neg A$. Accordingly, since we have $\models^{LPTT} A \vee \neg A$, the first step of the argument is legitimate. However, the validity of the LEM does not guarantee that either $T\langle\lambda\rangle$ or $\neg T\langle\lambda\rangle$ takes the value 1. What we have is that either $v_{\mathcal{M}^+}(T\langle\lambda\rangle) \geq \frac{1}{2}$ or $v_{\mathcal{M}^+}(\neg T\langle\lambda\rangle) \geq \frac{1}{2}$.

In fact, the Liar sentence λ takes the value $\frac{1}{2}$ any LP^+ models $\mathcal{M}^+$. Since $v_{\mathcal{M}^+}(\lambda) = \frac{1}{2}$, it follows that $v_{\mathcal{M}^+}(T\langle\lambda\rangle) = \frac{1}{2}$; because LP^+ models have the identity of truth property. By the valuation scheme, we also have $v_{\mathcal{M}^+}(\neg T\langle\lambda\rangle) = \frac{1}{2}$. Applying the valuation scheme again, it follows that $v_{\mathcal{M}^+}(T\langle\lambda\rangle \wedge \neg T\langle\lambda\rangle) = \frac{1}{2}$. Thus, we have: $\models^{LPTT} T\langle\lambda\rangle \wedge \neg T\langle\lambda\rangle$. Yet, this is the farthest we can go: the argument goes wrong at the last step. Since the ECQ is not valid in $LPTT$, we cannot infer from $T\langle\lambda\rangle \wedge \neg T\langle\lambda\rangle$ to an arbitrary B.

It is easy to construct a countermodel to the ECQ. It suffices to construct a model such that an atomic sentence Pa takes the value 0, that is, $v_{\mathcal{M}^+}(Pa) = 0$. Recall that $v_{\mathcal{M}^+}(T\langle\lambda\rangle \wedge \neg T\langle\lambda\rangle) = \frac{1}{2}$. Accordingly, $T\langle\lambda\rangle \wedge \neg T\langle\lambda\rangle \not\models^{LPTT} Pa$. Thus, the ECQ

is invalid. Notice that such a countermodel to the ECQ also shows that $LPTT$ is non-trivial: there is a LP^+ model $\mathcal{M}^+$ such that $v_{\mathcal{M}^+}(A) = 0$ for some $A \in \mathcal{L}^+$.

Parenthetical Remark. Notice that although the ECQ is invalid in $LPTT$, it does not mean that any instance of the argument from $A, \neg A$ to B is invalid. The argument from $T\langle\lambda\rangle \wedge \neg T\langle\lambda\rangle$ to $T\langle\lambda\rangle$ is clearly valid. We say that an inference rule is valid, iff all instance of the rule is valid. This means that an inference rule is invalid, iff there exists an instance of the rule is invalid.

5.2.3 Conditional

Just like K_3TT, $LPTT$ is also charged with not having a suitable conditional. Specifically, $\supset$ is not detachable, that is, $\supset$-modus ponens does not hold in $LPTT$:

- $A, A \supset B \not\models^{LPTT} B$.

To see this, consider a LP^+ model $\mathcal{M}^+$ such that $v_{\mathcal{M}^+}(A) = \frac{1}{2}$ and $v_{\mathcal{M}^+}(B) = 0$. In this model, $v_{\mathcal{M}^+}(A) = v_{\mathcal{M}^+}(A \supset B) = \frac{1}{2}$. That is, the model assigns a designated value to the premises but assigns a non-designated value to the conclusion.

Detachable Conditional. There are two possible moves. The first one is to extend $LPTT$ with a detachable conditional $\rightarrow$. The dominant approach is to take the detachable conditional to be a relevant conditional. To give the semantics of $\rightarrow$, Beall, in his *Spandrels of Truth*, makes uses of a simplified version of Routley-Meyer ternary semantics, the system B given by Priest and Sylvan (1992).

The semantics B is a modal semantics with abnormal worlds. Specifically, the semantics for $\rightarrow$ is specified by two different clauses: the normal clause and the abnormal clause. Let W be a set of worlds, $P \subseteq W$ is the set of possible worlds (i.e., normal worlds) and $I = W - P$ is the set of impossible worlds (i.e., abnormal worlds). Then, the semantics for the conditional $\rightarrow$ is defined as follows:

- Possible worlds, for $w \in P$:
  $w \models A \rightarrow B$ iff for all $w' \in W$, if $w' \models A$, then $w' \models B$.
- Impossible worlds, for $w \in I$:
  $w \models A \rightarrow B$ iff for all $w', w'' \in W$, such that $Rww'w''$, if $w' \models A$, then $w'' \models B$, where R is a tenary accessibility relation.

Beall's conditional has the following characteristics:

- $\rightarrow$-modus ponens is valid: $A, A \rightarrow B \models B$
  Also, the conditional form is valid: $\models (A \wedge (A \rightarrow B)) \Rightarrow B)$
- $\rightarrow$-identity is valid: $\models A \rightarrow A$
- All instances of $T\langle A\rangle \leftrightarrow A$ are valid: $\models T\langle A\rangle \leftrightarrow A$
- The following Curry-related Contraction principles are invalid:
  (For more on the details of Beall's treatment of Curry's paradox, see Beall 2009, chapter 2.)
  - $(A \rightarrow (A \rightarrow B)) \rightarrow (A \rightarrow B)$
  - $(A \wedge (A \rightarrow B)) \rightarrow B$

– $A \to (A \to B) \models A \to B$

However, Beall (2015) later confesses that the quest of finding a non-material detachment conditional may not be a good way to go, as the route is full of technical challenges and philosophically suspicious:

> [A]s Dunn, Meyer and Routley noted early on, Curry's paradox riddles the quest for detachable conditionals with severe problems; and overcoming the problems makes for very complicated, philosophically awkward semantics, and indeed often engenders the need to find yet other detachable conditionals to serve other pressing needs ... The quest for detachable conditionals that are suitable for glut theories can – and often does – appear to the informed observer as the wrong direction of reply to the problem of material non-detachment. (ibid, p. 410)

Material Conditional. Accordingly, Beall (2011, 2015) explores the option of not having a detachable conditional. The challenge is to explain that "we can (and do) successfully carry on rational inquiry despite the invalidity of modus ponens" (Beall, 2015, p. 418).

Beall's explanation involves the distinction between logic and the theory of reasoning (or the theory of rational change in view). Logic, Beall thinks, primarily concerns what sets of sentences follow from what sets of sentences. It does not say anything about what one ought to accept or reject. On the other hand, the theory of reasoning concerns how agents ought to reason: how one should form, retain, or abandon one's beliefs.

Nevertheless, although logic and the theory of reasoning are distinct, they can be linked by the following extralogical principles:

- Logical Implication Principle (IMP): $\Gamma \models \Delta$, then it is irrational for the agent S to accept Γ and reject Δ.
- Logical Consistency Principle (LCP): The agent S ought to reject contradictions (i.e., sentences of the form $A \wedge \neg A$).

According to IMP, if you accept everything in Γ, you ought not reject everything in Δ, given that logic tells us that Γ entails Δ. Another principle, LCP, says that you should reject inconsistencies. Yet, these principles are not indefeasible. In some circumstances, theoretical considerations, like conservativeness, simplicity, explanatory power, might pressure us disobey such principles.

With these considerations in mind, we can explain why we usually reason in accordance with modus ponens. Recall that detachment is invalid in $LPTT$:

- $\{A, A \supset B\} \not\models^{LPTT} \{B\}$

But we have a close cousin of detachment is valid:

- $\{A, A \supset B\} \models^{LPTT} \{B, A \wedge \neg A\}$

To see this, suppose that $v_{\mathcal{M}^+}(A) \geq \frac{1}{2}$ and $v_{\mathcal{M}^+}(A \supset B) \geq \frac{1}{2}$. There are two cases:

» Suppose that $v_{\mathcal{M}^+}(A) = 1$. Then $v_{\mathcal{M}^+}(B) \geq \frac{1}{2}$. To see this, suppose for reductio that $v_{\mathcal{M}^+}(B) = 0$. By the valuation scheme, $v_{\mathcal{M}^+}(A \supset B) = 0$, contradicting our initial supposition. Hence, there is some sentence in the conclusion set $\{B, A \wedge \neg A\}$ whose value is 1 or $\frac{1}{2}$. This means that $\{A, A \supset B\} \models^{LPTT} \{B, A \wedge \neg A\}$.

» Suppose that $v_{\mathcal{M}^+}(A) = \frac{1}{2}$. By the semantics of $\neg$, it follows that $v_{\mathcal{M}^+}(\neg A) = \frac{1}{2}$. Hence, by the semantics of $\wedge$, we have $v_{\mathcal{M}^+}(A \wedge \neg A) = \frac{1}{2}$. Hence, there is some sentence in $\{B, A \wedge \neg A\}$ whose value is 1 or $\frac{1}{2}$, which means that $\{A, A \supset B\} \models^{LPTT} \{B, A \wedge \neg A\}$.

Now logic tells us that $\{B, A \wedge \neg A\}$ follows from our theory $\{A, A \supset B)$. Our theory $\{A, A \supset B)$ does not entail any proper subtheory of $\{B, A \wedge \neg A\}$: $\{B\}$ and $\{A, \wedge \neg A\}$ does not follow from $\{A, A \supset B)$. IMP tells you that if you accept $\{A, A \supset B\}$, you ought not reject everything in $\{B, A \wedge \neg A\}$. Nevertheless, you still have a choice: the choice between a consistent and an inconsistent option. According to Beall, if logic says that $\Gamma \models \Delta$ but Γ fails to imply any proper subtheory of Δ (i.e., proper subset of Δ), then logic leaves us with choices – we can choose which elements in Δ to accept.

Beall suggests that this is where LCP comes into play. The reason why we accept B and reject $A \wedge \neg A$ is due to the extralogical principle LCP. And according to Beall, LCP is compatible with dialetheism:

> [T]he best balance of conservativeness and coherence has us accepting certain contradictions – the bizarre and, fortunately, rare ones like liar-paradoxical sentences. This isn't a hard knock against [LCP]; it continues in full force for the vast array of normal cases. And such force is sufficient, in the vast array of normal cases, to get us to accept B from $\{A, A \supset B\}$ via a rejection of $A \wedge \neg A$. (Beall, 2015, p. 417)

In what follows, our investigation focuses on Beall's 'non-detachable dialetheism', dialetheism without a detachable conditional. But let us first turn to another kind of dialetheic theories: the strict-tolerant approaches.

Parenthetical Remark. Dialetheists can also explain why we usually reason in accordance to modus ponens by appealing to minimally inconsistent models. Specifically, Preist (1991) develops a logic LP_m, the logic with the minimal inconsistent LP models. In LP_m, if Γ is consistent, its LP_m models are its classical models. As a consequence, if $\{A, A \supset B\}$ is consistent, then $\{A, A \supset B\} \models^{LP_m} \{B\}$. On a related note, Priest (2006) defends the cost of losing disjunctive syllogism is not as great as one might expect. For more details, see Priest 2006, chapter 8.

5.3 The Strict-Tolerant Approaches

STTT. Making use of LP^+ models, one can obtain another logic called the strict-tolerant logic $STTT$, which is recently developed by Cobreros, Egré, Ripley and van Rooij (2013). For convenience, we call the models in this logic the ST^+ models, though ST^+ models are just LP^+ models.

The key feature of $STTT$ is its consequence relation: $STTT$ is a logic that results

from a non-transitive relation of consequence over ST^+ models. Strict-tolerant consequence is defined in terms of countermodels as usual:

- A ST^+ is a countermodel to an argument from the premises Γ to conclusion Δ iff it assigns 1 to every premises of Γ and 0 to every conclusion of Δ.

Hence, strict-tolerant consequence goes from a strictly true set of premises (i.e., every premise takes the value 1) to a tolerantly true conclusions (i.e., there is a sentence in the conclusion set such that it does not take the value 0). Thus, we have:

Definition 23 ($STTT$ Consequence)**.** $\Gamma \models^{STTT} \Delta$ iff if $v_{\mathcal{M}^+}(A) = 1$ for all $A \in \Gamma$, then $v_{\mathcal{M}^+}(B) > 0$ for some $B \in \Delta$.

Notice that $STTT$ does not define countermodel in terms of designated values. That is, it does not rely on the principle of designated value preservation. As Cobreros et al. (2013) remark, $STTT$ does not divide sentences into the designated and the non-designated:

> It is natural to see the values in a model theory as intimately tied to (idealized) assertibility; this is so whether one thinks that assertibility is prior to semantic value or vice versa (or neither). More familiar approaches to three-valued models invoke a notion of 'designated value'; this amounts to imposing a two-way division over the top: either value-1 sentences are assertible and others are not, or else value-0 sentences are not assertible and others are. But there is no way to understand an $STTT$-based approach in terms of designated values, and we do not impose this two-way division. (ibid, p.14)

Classical Validities and Conditional. One important feature of $STTT$ is that it preserves all classically-valid inferences in the language $\mathcal{L}^+$. Suppose that $\models^{CL}$ is a consequence relation in classical logic. Then, we have:

Fact 24. if $\Gamma \models^{CL} \Delta$, then $\Gamma^* \models^{STTT} \Delta^*$, for any uniform substitution $*$ (*of open formulas for predicates, avoiding bound-variable conflict in the usual ways*) on the full language $\mathcal{L}^+$. (See Ripley, 2012)

Accordingly, unlike K_3TT, $STTT$ has $A \supset A$, $A \equiv A$, $T\langle A\rangle \equiv A$ as validities. Unlike $LPTT$, modus ponens is valid in $STTT$: $A, A \supset B \models^{STTT} B$.

Metainferences. However, $STTT$ is not as classical as it appears. Specifically, it loses some classical properties concerning *metainferences* – the principles under which a consequence relation might be closed. Here are some familiar examples of metainferences:

- Transitivity: If $A \models B$, and $B \models C$, then $A \models C$.
- Meta-Modus Ponens: If $\Gamma \models A$, and $\Gamma \models A \supset B$, then $\Gamma \models B$
- Reasoning by Cases: If $A \models C$ and $B \models C$, then $A \vee B \models C$

Some metainferences that hold in classical logic do not hold in $STTT$. Specifically, The following familiar metainferences fail in $STTT$:

- If $A \models B$, and $B \models C$, then $A \models C$ (Transitivity)
- If $\Gamma \models A$, and $\Gamma \models A \supset B$, then $\Gamma \models B$ (Meta-Modus Ponens)
- If $\Gamma \models \neg B$, and $\Gamma \models A \supset B$, then $\Gamma \models \neg A$ (Meta-Modus Pollens)
- If $\Gamma, A \models B \wedge \neg B, \Delta$, then $\Gamma \models \neg A, \Delta$ (A version of Meta-Reductio)
- If $\Gamma \models A \wedge \neg A, \Delta$, then $\Gamma \models \bot, \Delta$ (Meta-Explosion)
- If $\Gamma, A \vee B \models \Delta$ and $\Gamma, \neg A \models \Delta$, then $\Gamma \models B, \Delta$ (Meta- $\vee$ syllogism)

Nevertheless, it is precisely the loss of some metainferences that helps dealing with the Liar paradox.

Parenthetical Remark. Notice that while this version of Meta-Reductio, $\Gamma, A \models B \wedge \neg B, \Delta$, then $\Gamma \models \neg A, \Delta$, is not valid, other forms of Meta-Reductio are valid in $STTT$. In particular, we have:

- If $\Gamma, A \models^{STTT} \neg A, \Delta$, then $\Gamma \models^{STTT} \neg A, \Delta$
- If $\Gamma, A \models^{STTT} \bot, \Delta$, then $\Gamma \models^{STTT} \neg A, \Delta$

The Liar Paradox. Consider the Liar argument again:

1	$T\langle\lambda\rangle \vee \neg T\langle\lambda\rangle$	LEM
2	$T\langle\lambda\rangle$	Assumption
3	λ	2: Release
4	$\neg T\langle\lambda\rangle$	3: The Def of λ
5	$T\langle\lambda\rangle \wedge \neg T\langle\lambda\rangle$	2, 4: $\wedge$-Intro
6	$\neg T\langle\lambda\rangle$	Assumption
7	λ	6: The Def of λ
8	$T\langle\lambda\rangle$	7: Capture
9	$T\langle\lambda\rangle \wedge \neg T\langle\lambda\rangle$	6, 8: $\wedge$-Intro
10	$T\langle\lambda\rangle \wedge \neg T\langle\lambda\rangle$	1, 5, 9: Reasoning by Cases
11	B	10: ECQ

Since all classical laws and naive principles of truth (i.e., Capture and Release) are valid in $STTT$, every step in the above argument is valid. Nevertheless, this does not mean that the argument chaining these steps together would be valid as well. Notice that the principle of validity-preservation (VP) is invalid in $STTT$:

- Validity-Preservation (VP): if $\models A$ and $A \models B$, then $\models B$

To see this, consider a ST^+ model such that $v_{\mathcal{M}^+}(Pa) = 0$. Since the Liar sentence λ must take the value $\frac{1}{2}$, we have $v_{\mathcal{M}^+}(\lambda) = \frac{1}{2}$. By our familiar valuation scheme, we also have $v_{\mathcal{M}^+}(T\langle\lambda\rangle \wedge \neg T\langle\lambda\rangle) = \frac{1}{2}$. Thus, $\models^{STTT} T\langle\lambda\rangle \wedge \neg T\langle\lambda\rangle$ and $T\langle\lambda\rangle \wedge \neg T\langle\lambda\rangle \models^{STTT} Pa$ but not $\not\models^{STTT} Pa$. Thus, VP is invalid. This countermodel to VP shows that STTT is non-trivial: there is a ST^+ model $\mathcal{M}^+$ such that $v_{\mathcal{M}^+}(A) = 0$ for some $A \in \mathcal{L}^+$.

5.4 Expressive Limitations and Revenge Paradoxes

5.4.1 Expressive Limitations

Recall that the paracomplete approaches have problems concerning expressive limitations and revenge paradoxes.

i. The claim that the Liar sentence is neither true nor false, if formalized as $\neg(T\langle\lambda\rangle \vee T\langle\neg\lambda\rangle)$, does not come out true in paracomplete logics; $\neg(T\langle\lambda\rangle \vee T\langle\neg\lambda\rangle)$ must take the value $\frac{1}{2}$.

ii. Adding some extra connectives to increase expressive power gives rise to revenge paradoxes.

It is often argued that the dialetheic approaches a dual problem (e.g., Littmann & Simmons 2004 and Shapiro 2004). Specifically, dialetheic theories have problems of using the notion of 'just true' ('just false') to charactize non-dialetheic true (false) sentences:

i*. The claim that a sentence A is just true, if formalized as $T\langle A\rangle \wedge \neg T\langle\neg A\rangle$, is a contradiction in dialetheic logics; $T\langle A\rangle \wedge \neg T\langle\neg A\rangle$ can still have the value $\frac{1}{2}$. (Similar problems arise out of the notion of just false.)

ii*. Increasing expressive power gives rise to revenge paradoxes, trivializing the logics (i.e., the extension of $LPTT$ and $STTT$).

In what follows, we discuss whether or not (i*) and (ii*) are tenable.

Falsity. For convenience, let us introduce a falsity predicate. Falsity is defined as truth of negation:

- $F\langle A\rangle$ iff $T\langle\neg A\rangle$

Taking 'iff' to be the material biconditional, we have the following truth-table:

A	$\neg A$	$T\langle A\rangle$	$F\langle A\rangle$
1	0	1	0
$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$
0	1	0	1

Just True. It is often argued that dialetheic approaches fail to characterize sentences as being just true. An obvious way to say a sentence A is just true would be to say that A is true but not false:

$$T\langle A\rangle \wedge \neg T\langle\neg A\rangle$$

Alternatively, we can write the above as:

$$T\langle A\rangle \wedge \neg F\langle A\rangle$$

However, both are merely equivalent to $T\langle A\rangle$. Notice that $\neg F\langle A\rangle$ (or $\neg T\langle\neg A\rangle$) is equivalent to $T\langle A\rangle$. To see that, we can show that $\neg F\langle A\rangle$ and $T\langle A\rangle$ always have the same value:

» Suppose that A takes the value 1. $T\langle A\rangle$ takes the value 1. According to the semantics of F, $F\langle A\rangle$ takes the value 0. Then, the semantics of $\neg$ tells us that $\neg F\langle A\rangle$ takes the value 1.

» Suppose that A takes the value $\frac{1}{2}$. $T\langle A\rangle$ takes the value $\frac{1}{2}$. According to the semantics of F, $F\langle A\rangle$ takes the value $\frac{1}{2}$. Then, the semantics of $\neg$ tells us that $\neg F\langle A\rangle$ takes the value $\frac{1}{2}$.

» Suppose that A takes the value 0. $T\langle A\rangle$ takes the value 0. According to the semantics of F, $F\langle A\rangle$ takes the value 1. Then, the semantics of $\neg$ tells us that $\neg F\langle A\rangle$ takes the value 0.

Hence, $T\langle A\rangle \wedge \neg F\langle A\rangle$ amounts to $T\langle A\rangle$. This is undesirable; because, intuitively, there is a difference between truth and just truth: if a sentence is just true, it is not a dialetheia.

If there were no difference between truth and just truth, there would be some sentences which are dialetheic but just true. Consider the Liar sentence λ. Recall that λ must take the value $\frac{1}{2}$. By the identity of truth, $T\langle\lambda\rangle$ takes the value $\frac{1}{2}$ as well. But since $T\langle A\rangle \wedge \neg F\langle A\rangle$ is equivalent to $T\langle A\rangle$, $T\langle\lambda\rangle \wedge \neg F\langle\lambda\rangle$ takes the value $\frac{1}{2}$ as well.

Just False. Analogously, dialetheic theories has the problem of using the notion of just false to characterize non-contradictory false sentences. An apparent way to say that a sentence A is just false would be to say that A is false but not true:

$$F\langle A\rangle \wedge \neg T\langle A\rangle$$

This simply amounts to $F\langle A\rangle$. Recall that $\neg F\langle A\rangle$ and $T\langle A\rangle$ always have the same value. It can easily be checked that $F\langle A\rangle$ and $\neg T\langle A\rangle$ always have the same value as well.

Moreover, consider a sentence which says of itself that it is just false:

(18) (18) is just false.

Using our current expressive resource, we can formalize (18) as:

$$\kappa \rightleftharpoons F\langle\kappa\rangle \wedge \neg T\langle\kappa\rangle$$

Using familiar reasoning, we can show that κ is both true and just false, that is, $T\langle\kappa\rangle \wedge F\langle\kappa\rangle \wedge \neg T\langle\kappa\rangle$.

1	$T\langle\kappa\rangle \vee F\langle\kappa\rangle$	LEM
2	$T\langle\kappa\rangle$	Assumption
3	κ	2: Release
4	$F\langle\kappa\rangle \wedge \neg T\langle\kappa\rangle$	3: Def of κ
5	$T\langle\kappa\rangle \wedge F\langle\kappa\rangle \wedge \neg T\langle\kappa\rangle$	2, 4: $\wedge$-Intro
6	$F\langle\kappa\rangle$	Assumption
7	$\neg T\langle\kappa\rangle$	6: $F\langle A\rangle \equiv \neg T\langle A\rangle$
8	$F\langle\kappa\rangle \wedge \neg T\langle\kappa\rangle$	6, 7: $\wedge$-Intro
9	κ	9: Def of κ
10	$T\langle\kappa\rangle$	10: Capture
11	$T\langle\kappa\rangle \wedge F\langle\kappa\rangle \wedge \neg T\langle\kappa\rangle$	11, 9: $\wedge$-Intro
12	$T\langle\kappa\rangle \wedge F\langle\kappa\rangle \wedge \neg T\langle\kappa\rangle$	1, 5, 12: Reasoning By Cases

Priest's Reply. Priest suggests that the problem of 'just false' poses no threat to dialetheic theories. Specifically, $T\langle\kappa\rangle \wedge F\langle\kappa\rangle \wedge \neg T\langle\kappa\rangle$ is compatible with dialetheism.

> This is a contradiction of the kind that will sink any consistent solution, but it obviously does not sink a dialetheic solution. The contradiction is exactly what one should expect to get in the context. (Priest, 2006, p. 287)

Priest rightly points out that inconsistency is compatible with the dialetheic approaches. However, it is one thing to say that a contradiction is compatible with dialetheism; it is another thing to accept that the notion of just false (just true) is inconsistent. After all, it seems that it is part of the meaning of 'just false' ('just true') that it behaves consistently (Young, 2015b). In addition, on our current approach, the notion of just false (just true) cannot get the desired interpretation, since it is no different from the notion of falsity (truth).

5.4.2 Revenge Paradoxes?

To fix the above problems, one may suggest that we increase the expressive power of a language by adding more connectives. Let $LPTT^{\mathbb{J}}$ and $STTT^{\mathbb{J}}$ be the extension of $LPTT$ and $STTT$ respectively, with an unary operator $\mathbb{J}$ equipped with the following semantics.

A	$\neg A$	$T\langle A\rangle$	$F\langle A\rangle$	$\mathbb{J}A$	$\mathbb{F}A$	$\bullet A$	$\circ A$
1	0	1	0	1	0	0	1
$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	0	0	1	0
0	1	0	1	0	1	0	1

We take the $\mathbb{J}$ operator as primitive. Other newly introduced operators can be defined in terms of $\mathbb{J}$ and usual connectives:

- $\mathbb{F}A =_{df} \mathbb{J}\neg A$
- $\bullet A =_{df} \neg\mathbb{J}A \wedge \neg\mathbb{J}\neg A$
- $\circ A =_{df} \mathbb{J}A \vee \mathbb{J}\neg A$

$\mathbb{J}$, $\mathbb{F}$, $\bullet$ and $\circ$ are intended to model the notion of just true, just false, dialetheia, non-dialetheia respectively.

Now, one important question should be asked: are our new theories $LPTT^{\mathbb{J}}$ and $STTT^{\mathbb{J}}$ plagued by revenge paradoxes? The revenge paradox for the dialetheic approaches is generated by the Strengthened Liar sentence (18): '(18) is just false'. In what follows, we discuss how to formalize (18) with our current expressive resources. Specifically, we distinguish two kinds of revenge arguments, corresponding to two different kinds of procedures to formalize (18). According to Barrio et al. (2018), we can distinguish two kinds of procedures to form self-referential sentences:

> To formally mimic the self-referential character that some sentences like "This sentence is in English" or more prominently "This sentence is false" have, there seems to be two main technical options: through a strong or through a weak procedure. The latter option achieves this goal by requiring a self-referential sentence to be equivalent to a sentence that "talks about" the first one. The former involves an essential use of identities. This strong alternative, either (i) requires a term to be identical to the name of a sentence that "talks about" the first term, or (ii) involves a meta-linguistic denotation function from names to sentences of the language that have occurrences of that name in it. (ibid, p. 9- 10)

Accordingly, we can form a Strengthened Liar sentence such as (18) through a strong or a weak procedure. To form a Strengthened Liar sentence via a strong procedure, there are two ways to do so. We can represent (18) via the identity $=$:

- (18) can be represented by a sentence $\mathbb{J}\neg Tl$ such that $l = \langle \mathbb{J}\neg Tl \rangle$, where l is a constant symbol.

Given the classical interpretation of $=$, l and $\langle \mathbb{J}\neg Tl \rangle$ denote the sentence $\mathbb{J}\neg Tl$. Alternatively, (18) can be formed by a denotation function, such as the auxiliary function τ. The auxiliary function τ is a 1-1 function from distinguished names l onto sentence. Making use of τ, we can represent (18) as follows:

- If $\tau(l) = \mathbb{J}\neg Tl$, then $\mathbb{J}\neg Tl$ is a Strengthened Liar sentence (18) which says of itself that it is just false.

Because the constant symbol l in $\mathbb{J}\neg Tl$ refers to $\mathbb{J}\neg Tl$.

The Strengthened Liar sentence (18) can also be represented through a weak procedure. The Strengthened Liar sentence can be represented by a sentence $\mathbb{J}\neg T\langle L \rangle$ which is equivalent to L. Notice that L is a sentence. Care should be taken that there are two different notions of equivalence: the semantic equivalence $=\!|\!=$ and the (material) biconditional $\equiv$. One may wonder whether or not the semantic equivalence amounts to the material biconditional. In what follows, we will show that the question of whether A is semantically equivalent to B cannot always be reduced to the question whether A is materially equivalent to B. In any case, there are two ways to represent (18) via a weak procedure:

- $\xi \equiv \mathbb{J}\neg T\langle \xi \rangle$

- $\xi \dashv\models \mathbb{J}\neg T\langle\xi\rangle$

The availability of these procedures depends on our technical apparatus. If our theory is an extension of arithmetical theories which allow us to talk about sentences by Gödel's coding system, then both strong and weak procedures are available. Specifically, self-referential paradoxical sentences, such as (18) , can be formed by the strong Diagonalization Lemma or the weak Diagonalization Lemma. To do so, we first define a numerical property J corresponding to the operator $\mathbb{J}$ as follows:

- $v_{\mathcal{M}^{+\mathbb{J}}}(\mathsf{J}(n)) = \begin{cases} 1, & \text{if } n \text{ is the Gödel number of a sentence } A \text{ whose value is } 1. \\ 0, & \text{Otherwise.} \end{cases}$

The semantic version of the strong Diagonalization Lemma says that, for any formula Ax, there is a term t such that $\mathbf{T} \models t = \langle At\rangle$, where $\mathbf{T}$ is a theory that contains Peano arithmetic **PA**. The strong Diagonalization Lemma ensures that there is a sentence $\mathsf{J}\langle\neg Tl\rangle$ such that the term l is identical to the Gödel number $\langle\mathsf{J}(\neg Tl)\rangle$, that is, $\mathbf{T} \models l = \langle\mathsf{J}(\neg Tl)\rangle$. The weak Diagnalization Lemma says that, for any formula Ax, there is a sentence D such that $\mathbf{T} \models D \equiv A\langle D\rangle$. The weak Diagnalization Lemma ensures that there is a sentence ξ such that $\mathsf{T} \models \xi \equiv \mathsf{J}\langle\neg T(\xi)\rangle$.

On the other hand, if our theory does not contain arithmetical theories which are expressively rich enough, then the means of representing self-reference is not available in the theory. Nevertheless, we can still represent self-reference by imposing constraints upon the valuations. To achieve self-reference through a strong procedure, we can require a term that denotes a paradoxical sentence to be identical to the name that represents the paradoxical sentence. Alternatively, we can directly impose constraints upon a denotation function, ensuring the existence of paradoxical sentences whose constant symbol denotes the sentence in question.

To achieve self-reference through a weak procedure, we can restrict the valuations of the target theory of truth in the way that paradoxical sentences are equivalent to sentences that talk about themselves.

For convenience, we call $LPTT^{\mathbb{J}}$ and $STTT^{\mathbb{J}}$ with a strong self-referential procedure:

- $LPTT^{\mathbb{J}}_{s}$ and $STTT^{\mathbb{J}}_{s}$

respectively. As for the weak procedures, self-reference can be represented either by the semantic equivalence, or by the material biconditional. Corresponding to two different ways of representing self-reference, we have these pairs of theories:

- $LPTT^{\mathbb{J}}_{\equiv}$ and $STTT^{\mathbb{J}}_{\equiv}$
- $LPTT^{\mathbb{J}}_{\dashv\models}$ and $STTT^{\mathbb{J}}_{\dashv\models}$

In what follows, we will show that whether or not a theory is plagued by revenge paradoxes depends on which form of self-referential procedures is adopted.

Revenge via Strong Procedures. Let us first examine the revenge arguments

formed by a strong self-referential procedure. We will show that $LPTT_s^{\mathbb{J}}$ and $STTT_s^{\mathbb{J}}$ are trivial.

Fact 25. $LPTT_s^{\mathbb{J}}$ is trivial.

Proof. Let l denotes $\mathbb{J}\neg Tl$. We will show that there is no value available for Tl to take.

- Suppose that $v_{\mathcal{M}_s^{+\mathbb{J}}}(Tl) = 1$. Then, by the semantics of $\neg$, $v_{\mathcal{M}_s^{+\mathbb{J}}}(\neg Tl) = 0$. By the semantics of $\mathbb{J}$, $v_{\mathcal{M}_s^{+\mathbb{J}}}(\mathbb{J}\neg Tl) = 0$. This means that $\mathbb{J}\neg Tl$ is not in the extension $\mathcal{T}^+$ but is in the anti-extension $\mathcal{T}^-$. Since l denotes $\mathbb{J}\neg Tl$, $\langle\mathbb{J}\neg Tl\rangle$ and l denote the same object: $\mathbb{J}\neg Tl$. Hence, $v_{\mathcal{M}_s^{+\mathbb{J}}}(Tl) = 0$. This is impossible, since no sentence can receive more than one value in a model.
- Suppose that $v_{\mathcal{M}_s^{+\mathbb{J}}}(Tl) = 0$. Then, by the semantics of $\neg$, $v_{\mathcal{M}_s^{+\mathbb{J}}}(\neg Tl) = 1$. By the semantics of $\mathbb{J}$, $v_{\mathcal{M}_s^{+\mathbb{J}}}(\mathbb{J}\neg Tl) = 1$. This means that $\mathbb{J}\neg Tl$ is in the extension $\mathcal{T}^+$ but not in the anti-extension $\mathcal{T}^-$. Recall that $\langle\mathbb{J}\neg Tl\rangle$ and l denote the same object: $\mathbb{J}\neg Tl$. Hence, $v_{\mathcal{M}_s^{+\mathbb{J}}}(Tl) = 1$. Again, this is impossible.
- Suppose that $v_{\mathcal{M}_s^{+\mathbb{J}}}(Tl) = \frac{1}{2}$. Then, by the semantics of $\neg$, $v_{\mathcal{M}_s^{+\mathbb{J}}}(\neg Tl) = \frac{1}{2}$. By the semantics of $\mathbb{J}$, $v_{\mathcal{M}_s^{+\mathbb{J}}}(\mathbb{J}\neg Tl) = 0$. Hence, $\mathbb{J}\neg Tl \notin \mathcal{T}^+$ and $\mathbb{J}\neg Tl \in \mathcal{T}^-$. Since $\langle\mathbb{J}\neg Tl\rangle$ and l denote $\mathbb{J}\neg Tl$, it follows that $v_{\mathcal{M}_s^{+\mathbb{J}}}(Tl) = 0$. Once again, this is impossible.

In any case, Tl receives more than one value. But there cannot be such a model. So there cannot be any countermodel to any argument. Hence, every argument is valid. □

Fact 26. $STTT_s^{\mathbb{J}}$ is trivial.

Proof. It is just the same proof as for the fact 25. □

The upshot of the above discussion is that we can either give up some expressive resources (such as $\mathbb{J}$), or give up the the strong self-referential procedure. Yet, one may feel the latter option is as undesirable as the former one. As Heck (2012) emphasizes, it is only through the use of the identity that we can truly capture the self-referential character of paradoxical sentences:

> [T]he strong form, although less well-known, is what we need if we want to capture the structure of the informal reasoning that leads to the Liar paradox. One typically begins with the assumption that there is a self-referential sentence, the Liar, that says of itself that it is not true. The weaker form of the diagonal lemma does not give us such a sentence. It only gives us a formula Λ that is *provably equivalent* to a sentence that says of Λ that it is not true. Neither Λ nor $\neg T\langle\Lambda\rangle$ refers to itself, and neither *says of itself* that it is not true. The strong form, on the other hand, does deliver a truly self-referential liar sentence. Since $\lambda = \langle T(\lambda)\rangle$, $\neg T(\lambda)$ is a sentence that really does refer to itself and really does say of itself that it is not true. (ibid, p. 37)

In the footnote 15 of Barrio et al. (2018), it is suggested that a non-classical account of identity might help, if we wish to express self-reference through a strong procedure:

> An anonymous referee wonders whether or not this counts as an expressive limitation of this approach. In a way, our results imply that the use of a strong self-referential procedure to express self-reference – e.g., through identities – should be avoided, if one wishes to recover the classically valid meta-inferences without falling into triviality. This expressive limitation is a price to pay, if identity is a classical notion. If one wishes to express self-reference through a strong procedure, though, one available option would be to explore nonclassical accounts of identity, like the one discussed by Graham Priest (2014). (ibid, p .15)

The suggestion is that we make use of Priest's notion of non-transitive identity, which is developed in a second-order version of LP. On a related note, Cobreros et al.(2012) also define a non-transitive notion of identity in second-order ST based on Priest's (2010b) work. We leave the exploration of the suggestion in future work; since that would take us too far away from the main discussion. That being said, the suggestion is certainly worth pursuing: in fact, we will show that it is the non-transitive consequence that help surviving from revenge via the semantic equivalence.

Revenge via Weak Procedures: Material Biconditional. Let us turn to the revenge arguments formed by the material biconditional. We will show that the non-triviality of $LPTT^{\mathbb{J}}_{\equiv}$ and $STTT^{\mathbb{J}}_{\equiv}$.

Our proofs are similar to the one offered by Barrio et al.'s (2018) proof for the non-triviality of $STTT$ extended with the operator $\circ$. The basic idea of the proofs for the non-triviality of $LPTT^{\mathbb{J}}_{\equiv}$ and $STTT^{\mathbb{J}}_{\equiv}$ is as follows. We first construct a set that contains all and only sentences of the form

$$p^* \equiv Ap^*$$

where p^* is a distinguished propositional variable and Ap^* is a sentence that has one instance of $T\langle p^*\rangle$ as a subformula. Call the set PseudoDL (i.e., Pseudo Diagonalization Lemma). Notice that our language is not propositional. But let us abuse of notation for readability.

PseudoDL contains traditional pathological sentences:

- $t \equiv T\langle t\rangle$ (The Truth-Teller sentence)
- $c \equiv T\langle c\rangle \supset B$ (The Curry sentence)

It also contains the Strengthened Liar sentence such as (18):

- $\xi \equiv \mathbb{J}\neg T\langle \xi\rangle$ (The Strengthened Liar sentence)

where ξ is any atomic sentence.

Then, we restrict the set of models to the ones such that $v_{\mathcal{M}^{+\mathbb{J}}_{\equiv}}(p^* \equiv Ap^*) \neq 0$ for any model $\mathcal{M}^{+\mathbb{J}}_{\equiv}$. Equipped with LP consequence and ST consequence, the resulting theories are $LPTT^{\mathbb{J}}_{\equiv}$ and $STTT^{\mathbb{J}}_{\equiv}$ respectively. In other words, in such theories, there is no countermodel for all members of PseudoDL. In what follows, we will show that the new theories are non-trivial.

Fact 27. $LPTT^{\mathbb{J}}_{\equiv}$ is non-trivial.

Proof. By assumption, $v_{\mathcal{M}^{+\mathbb{J}}_{\equiv}}(p^* \equiv Ap^*) \neq 0$ for any model $\mathcal{M}^{+\mathbb{J}}_{\equiv}$. In particular, consider $\xi \equiv \mathbb{J}\neg T\langle\xi\rangle$. By assumption, we have $v_{\mathcal{M}^{+\mathbb{J}}_{\equiv}}(\xi \equiv \mathbb{J}\neg T\langle\xi\rangle) \neq 0$. There are three cases to be considered:

- Suppose that $v_{\mathcal{M}^{+\mathbb{J}}_{\equiv}}(\xi) = 1$. According to the semantics of $\neg$ and T, it follows that $v_{\mathcal{M}^{+\mathbb{J}}_{\equiv}}(\neg T\langle\xi\rangle) = 0$. According to the semantics of $\mathbb{J}$, we have $v_{\mathcal{M}^{+\mathbb{J}}_{\equiv}}(\mathbb{J}\neg T\langle\xi\rangle) = 0$. But this means that $v_{\mathcal{M}^{+\mathbb{J}}_{\equiv}}(\xi \equiv \mathbb{J}\neg T\langle\xi\rangle) = 0$. This is impossible. Hence. $v_{\mathcal{M}^{+\mathbb{J}}_{\equiv}}(\xi) \neq 1$.
- Suppose that $v_{\mathcal{M}^{+\mathbb{J}}_{\equiv}}(\xi) = 0$. According to the semantics of $\neg$ and T, it follows that $v_{\mathcal{M}^{+\mathbb{J}}_{\equiv}}(\neg T\langle\xi\rangle) = 1$. According to the semantics of $\mathbb{J}$, we have $v_{\mathcal{M}^{+\mathbb{J}}_{\equiv}}(\mathbb{J}\neg T\langle\xi\rangle) = 1$. But this means that $v_{\mathcal{M}^{+\mathbb{J}}_{\equiv}}(\xi \equiv \mathbb{J}\neg T\langle\xi\rangle) = 0$. This is impossible. Hence, $v_{\mathcal{M}^{+\mathbb{J}}_{\equiv}}(\xi) \neq 0$.
- Suppose that $v_{\mathcal{M}^{+\mathbb{J}}_{\equiv}}(\xi) = \frac{1}{2}$. According to the semantics of $\neg$ and T, it follows that $v_{\mathcal{M}^{+\mathbb{J}}_{\equiv}}(\neg T\langle\xi\rangle) = 0$. According to the semantics of $\mathbb{J}$, we have $v_{\mathcal{M}^{+\mathbb{J}}_{\equiv}}(\mathbb{J}\neg T\langle\xi\rangle) = 0$. In this case, $v_{\mathcal{M}^{+\mathbb{J}}_{\equiv}}(\xi \equiv \mathbb{J}\neg T\langle\xi\rangle) = \frac{1}{2}$. Thus, the only possible value for ξ to take is $\frac{1}{2}$.

Given that ξ takes the value $\frac{1}{2}$, $\mathbb{J}\neg T\langle\xi\rangle$ takes the value 0. This means that our theory is non-trivial. That is, there exists a sentence whose value is 0. □

Fact 28. $STTT^{\mathbb{J}}_{\equiv}$ is non-trivial.

Proof. It is just the same proof as for the fact 27. □

Semantic Equivalence and Material Biconditional. Strengthened Liar sentences are not only formed by the use of the biconditional $\equiv$; they are often formed by the use of the semantic equivalence $=\!|\!\models$ as well. Before we discuss revenge via the semantic equivalence, we will prove some facts concerning the relationship between the semantic equivalence and the material biconditional in $LPTT^{\mathbb{J}}$ and $STTT^{\mathbb{J}}$. Specifically, in $LPTT^{\mathbb{J}}$, the semantic equivalence is a stronger notion than the material biconditional; whereas, in $STTT^{\mathbb{J}}$, the semantic equivalence and the material biconditional are equally strong. Let's consider $LPTT^{\mathbb{J}}$ first.

Fact 29. If $A =\!|\!\models^{LPTT^{\mathbb{J}}} B$, then $\models^{LPTT^{\mathbb{J}}} A \equiv B$.

Proof. Suppose that $A =\!|\!\models^{LPTT^{\mathbb{J}}} B$. Then, there is no $LP^{+\mathbb{J}}$ model $\mathcal{M}^{+\mathbb{J}}$ such that:

a. $v_{\mathcal{M}^{+\mathbb{J}}}(A) = 1$ and $v_{\mathcal{M}^{+\mathbb{J}}}(B) = 0$

b. $v_{\mathcal{M}^{+\mathbb{J}}}(A) = \frac{1}{2}$ and $v_{\mathcal{M}^{+\mathbb{J}}}(B) = 0$

c. $v_{\mathcal{M}^{+\mathbb{J}}}(B) = 1$ and $v_{\mathcal{M}^{+\mathbb{J}}}(A) = 0$

d. $v_{\mathcal{M}^{+\mathbb{J}}}(B) = \frac{1}{2}$ and $v_{\mathcal{M}^{+\mathbb{J}}}(A) = 0$

Suppose for reductio that $\not\models^{LPTT^{\mathbb{J}}} A \equiv B$. Then, there is a $LP^{+\mathbb{J}}$ model $\mathcal{M}^{+\mathbb{J}}$ such that $v_{\mathcal{M}^{+\mathbb{J}}}(A \equiv B) = 0$. This is the case, only if the $LP^{+\mathbb{J}}$ model $\mathcal{M}^{+\mathbb{J}}$ is such that:

e. $v_{\mathcal{M}^{+\mathbb{J}}}(A) = 1$ and $v_{\mathcal{M}^{+\mathbb{J}}}(B) = 0$; or

f. $v_{\mathcal{M}^{+\mathbb{J}}}(B) = 1$ and $v_{\mathcal{M}^{+\mathbb{J}}}(A) = 0$.

But either way, there cannot be such a model: (e) contradicts (a) and (f) contradicts (c). Thus, $\models^{LPTT^{\mathbb{J}}} A \equiv B$. □

Fact 30. 'If $\models^{LPTT^{\mathbb{J}}} A \equiv B$, then $A \dashv\models^{LPTT^{\mathbb{J}}} B$' does not hold.

Proof. Consider $\lambda \equiv Pa$. Recall that the Liar sentence λ must take the value $\frac{1}{2}$ in any $LP^{+\mathbb{J}}$ model $\mathcal{M}^{+\mathbb{J}}$. According to the LP valuation scheme, $v_{\mathcal{M}^{+\mathbb{J}}}(\lambda \equiv Pa) = \frac{1}{2}$. Thus, $\models^{LPTT^{\mathbb{J}}} \lambda \equiv Pa$.

On the other hand, we can show that $\lambda \not\models^{LPTT^{\mathbb{J}}} Pa$. Consider a $LP^{+\mathbb{J}}$ model such that $v_{\mathcal{M}^{+\mathbb{J}}}(Pa) = 0$. Since $v_{\mathcal{M}^{+\mathbb{J}}}(\lambda) = \frac{1}{2}$ in this model, this model is a countermodel to the inference from λ to Pa. Hence, $\lambda \not\models^{LPTT^{\mathbb{J}}} Pa$.

Thus, $\models^{LPTT^{\mathbb{J}}} A \equiv B$ does not ensure $A \dashv\models^{LPTT^{\mathbb{J}}} B$. □

Fact 31. If $A \dashv\models^{STTT^{\mathbb{J}}} B$, then $\models^{STTT^{\mathbb{J}}} A \equiv B$.

Proof. Suppose that $A \dashv\models^{STTT^{\mathbb{J}}} B$. This entails that there is no $ST^{+\mathbb{J}}$ model $\mathcal{M}^{+\mathbb{J}}$ such that:

a. $v_{\mathcal{M}^{+\mathbb{J}}}(A) = 1$ and $v_{\mathcal{M}^{+\mathbb{J}}}(B) = 0$

b. $v_{\mathcal{M}^{+\mathbb{J}}}(B) = 1$ and $v_{\mathcal{M}^{+\mathbb{J}}}(A) = 0$

Suppose for reductio that $\not\models^{STTT^{\mathbb{J}}} A \equiv B$. Then there is a ST^{+} model $\mathcal{M}^{+\mathbb{J}}$ such that $v_{\mathcal{M}^{+\mathbb{J}}}(A \equiv B) = 0$. This is the case, only if the $ST^{+\mathbb{J}}$ model $\mathcal{M}^{+\mathbb{J}}$ is such that:

c. $v_{\mathcal{M}^{+\mathbb{J}}}(A) = 1$ and $v_{\mathcal{M}^{+\mathbb{J}}}(B) = 0$; or

d. $v_{\mathcal{M}^{+\mathbb{J}}}(B) = 1$ and $v_{\mathcal{M}^{+\mathbb{J}}}(A) = 0$.

But either way, there cannot be such a model: (c) contradicts (a) and (d) contradicts (b). Thus, $\models^{STTT^{\mathbb{J}}} A \equiv B$. □

Fact 32. If $\models^{STTT^{\mathbb{J}}} A \equiv B$, then $A \dashv\models^{STTT^{\mathbb{J}}} B$.

Proof. Suppose that $\models^{STTT^{\mathbb{J}}} A \equiv B$. This means that there is no $ST^{+\mathbb{J}}$ model $\mathcal{M}^{+\mathbb{J}}$ such that $v_{\mathcal{M}^{+\mathbb{J}}}(A \equiv B) = 0$. Thus, Accordingly, there is no $ST^{+\mathbb{J}}$ model $\mathcal{M}^{+\mathbb{J}}$ such that:

a. $v_{\mathcal{M}^{+\mathbb{J}}}(A) = 1$ and $v_{\mathcal{M}^{+\mathbb{J}}}(B) = 0$

b. $v_{\mathcal{M}^{+\mathbb{J}}}(B) = 1$ and $v_{\mathcal{M}^{+\mathbb{J}}}(A) = 0$

(a) ensures $A \models^{STTT^{\mathbb{J}}} B$; whereas (b) ensures that $B \models^{STTT^{\mathbb{J}}} A$. Hence,we have $A \eqslantless\models^{STTT^{\mathbb{J}}} B$. □

Revenge via Weak Procedures: Semantic Equivalence. Finally, let us consider the revenge argument formed by the semantic equivalence. Recall that the semantic equivalence is a stronger notion than the material biconditional in $LPTT^{\mathbb{J}}$. So while $LPTT^{\mathbb{J}}_{\equiv}$ is non-trivial, it does not follow that $LPTT^{\mathbb{J}}_{\eqslantless\models}$ is so. As a matter of fact, we can show that $LPTT^{\mathbb{J}}_{\eqslantless\models}$ is trivial.

Fact 33. $LPTT^{\mathbb{J}}_{\eqslantless\models}$ is trivial.

Proof. Suppose that ξ is intersubstitutable for $\mathbb{J}\neg T\langle\xi\rangle$. That is, suppose that:

$$\xi \eqslantless\models^{LPTT^{\mathbb{J}}_{\eqslantless\models}} \mathbb{J}\neg T\langle\xi\rangle$$

- Suppose that $v_{\mathcal{M}^{+\mathbb{J}}_{\eqslantless\models}}(\xi) = 1$. By the identity of truth, $v_{\mathcal{M}^{+\mathbb{J}}_{\eqslantless\models}}(T\langle\xi\rangle) = 1$. According to the semantics of $\neg$, $v_{\mathcal{M}^{+\mathbb{J}}_{\eqslantless\models}}(\neg T\langle\xi\rangle) = 0$. According to the semantics of $\mathbb{J}$, $v_{\mathcal{M}^{+\mathbb{J}}_{\eqslantless\models}}(\mathbb{J}\neg T\langle\xi\rangle) = 0$. Hence, $\xi \not\models^{LPTT^{\mathbb{J}}} \mathbb{J}\neg T\langle\xi\rangle$, violating the initial assumption.
- Suppose that $v_{\mathcal{M}^{+\mathbb{J}}_{\eqslantless\models}}(\xi) = 0$. By the identity of truth, $v_{\mathcal{M}^{+\mathbb{J}}_{\eqslantless\models}}(T\langle\xi\rangle) = 0$. According to the semantics of $\neg$, $v_{\mathcal{M}^{+\mathbb{J}}_{\eqslantless\models}}(\neg T\langle\xi\rangle) = 1$. According to the semantics of $\mathbb{J}$, $v_{\mathcal{M}^{+\mathbb{J}}_{\eqslantless\models}}(\mathbb{J}\neg T\langle\xi\rangle) = 1$. Hence, $\mathbb{J}\neg T\langle\xi\rangle \not\models^{LPTT^{\mathbb{J}}} \xi$, violating the initial assumption.
- Suppose that $v_{\mathcal{M}^{+\mathbb{J}}_{\eqslantless\models}}(\xi) = \frac{1}{2}$. By the identity of truth, $v_{\mathcal{M}^{+\mathbb{J}}_{\eqslantless\models}}(T\langle\xi\rangle) = \frac{1}{2}$. According to the semantics of $\neg$, $v_{\mathcal{M}^{+\mathbb{J}}_{\eqslantless\models}}(\neg T\langle\xi\rangle) = \frac{1}{2}$. According to the semantics of $\mathbb{J}$, $v_{\mathcal{M}^{+\mathbb{J}}_{\eqslantless\models}}(\mathbb{J}\neg T\langle\xi\rangle) = 0$. Hence, $\xi \not\models^{LPTT^{\mathbb{J}}} \mathbb{J}\neg T\langle\xi\rangle$, violating the initial assumption.

In any case, ξ cannot be intersubstitutable for $\mathbb{J}\neg T\langle\xi\rangle$. □

On the other hand, the semantic equivalence and the material biconditional are equally strong in $STTT^{\mathbb{J}}$. Thus, given that $STTT^{\mathbb{J}}_{\equiv}$ is non-trivial, it seems reasonable to expect that we can offer a non-triviality proof for $STTT^{\mathbb{J}}_{\eqslantless\models}$. And in fact, we can.

Fact 34. $STTT^{\mathbb{J}}_{\eqslantless\models}$ is non-trivial.

Proof. We show that the existence of self-referential sentences ensures that there is a sentence whose value is 0.

Suppose that $A \eqslantless\models^{STTT^{\mathbb{J}}_{\eqslantless\models}} B$. By fact 31, $\models^{STTT^{\mathbb{J}}_{\eqslantless\models}} A \equiv B$. This entails that there is no $ST^{+\mathbb{J}}$ model $\mathcal{M}^{+\mathbb{J}}$ such that:

a. $v_{\mathcal{M}^{+\mathbb{J}}_{\eqslantless\models}}(A) = 1$ and $v_{\mathcal{M}^{+\mathbb{J}}_{\eqslantless\models}}(B) = 0$

b. $v_{\mathcal{M}^{+\mathbb{J}}_{\eqslantless\models}}(B) = 1$ and $v_{\mathcal{M}^{+\mathbb{J}}_{\eqslantless\models}}(A) = 0$

So either $v_{\mathcal{M}^{+\mathbb{J}}_{\dashv\models}}(A) = \frac{1}{2}$ or $v_{\mathcal{M}^{+\mathbb{J}}_{\dashv\models}}(B) = \frac{1}{2}$. Either way, using our just false operator $\mathbb{F}$, we can form a sentence whose value is 0.

We can check that the Strengthened Liar sentence (18) fails to trivialize $STTT^{\mathbb{J}}_{\dashv\models}$. Suppose that $\xi =\!|\!\models^{STTT^{\mathbb{J}}_{\dashv\models}} \mathbb{J}\neg T\langle\xi\rangle$. Consider a model $\mathcal{M}^{+\mathbb{J}}_{\dashv\models}$ such that $v_{\mathcal{M}^{+\mathbb{J}}_{\dashv\models}}(\xi) = \frac{1}{2}$. If so, we have $v_{\mathcal{M}^{+\mathbb{J}}_{\dashv\models}}(\mathbb{J}\neg T\langle\xi\rangle) = 0$. Yet, this does not violate the assumption that $\xi =\!|\!\models^{STTT^{\mathbb{J}}_{\dashv\models}} \mathbb{J}\neg T\langle\xi\rangle$. □

The upshot of this section is that $STTT^{\mathbb{J}}_{\equiv}$ and $STTT^{\mathbb{J}}_{\dashv\models}$ are shown to be non-trivial as desired.

Note that our strict-tolerant theory $STTT^{\mathbb{J}}_{\dashv\models}$ (or $STTT^{\mathbb{J}}_{\equiv}$) undermines the motivation for the Embracing Revenge view recently developed independently by Cook (2008) and Schlenke (2010). According to the Embracing Revenge view, because of revenge paradoxes, we should accept that the class of nonclassical semantic values is indefinitely extensible. Cook (2009) describes the revenge problem as follows:

> Put simply, the problem is this: Given any semantics that purports to deal adequately with various semantic paradoxes such as the Liar and its strengthened variants, if we extend our language to include the resources for discussing the truth values assigned to statements in that semantics, then we will be able to construct a statement using these novel resources which cannot have exactly one of those truth values as its semantic value. In particular, if T, F, V_1, V_2, ... V_n are the (exclusive and exhaustive) truth values admitted by the semantics, and we extend our language by adding predicates $T(x)$, $F(x)$, $V_1(x)$, $V_2(x)$, ... $V_n(x)$ which hold a statement (of its Gödel code, etc) if and only if the statement receives the correspoding truth value (and no other value), then the corresponding Super-Liar sentence:
>
> $$SupL : F(SupL) \vee V_1(SupL) \vee V_2(SupL) \vee ...V_n(SupL)$$
>
> cannot receive any one of T, F, V_1, V_2, ... V_n as its truth value. (ibid p. 192)

Then, it seems that we are left with two options. The first one is 'to stop the regress by denying, at some point, that the semantic concepts used in the metalanguage can be legitimately added to the object language in question' (ibid). Cook deems this option undesirable:

> Such restrictions violate strong intuitions concerning the functioning of language and our apparent ability to straightforwardly express such concepts in natural language, however. In particular, it seems that we can, in fact, meaningfully (even if sometimes mistakenly) say things such as "All statements in our language are false." or "Some statements in our language are pathological." Such restrictions on what we can say — in other words, claims that we cannot express what we seem to be able to express quite easily — seem to me to

involve biting a somewhat too large and unpalatable bullet. (ibid, p. 192 - 193)

The second option is that we should 'embrace revenge as an inherent feature of semantic theorizing itself, and of the language(s) within which we carry out such endeavors' (Tourville & Cook, 2016, p. 328). That is, we should accept the following thesis:

- Given any definite collection of exclusive semantic values $\mathcal{V}$, there is a sentence that cannot receive any of the values in $\mathcal{V}$. (Tourville & Cook, 2016, p. 327 - 328)

However, as we saw, if we make use of the biconditional or the semantic equivalence to represent self-reference and the consequence relation is a non-transitive one, then we need not introduce infinitely many truth values to deal with the Liar paradox and its strengthened versions. The non-triviality proof for $STTT^{\mathbb{J}}_{\Rightarrow\models}$ (or $STTT^{\mathbb{J}}_{\equiv}$) shows that every sentence in the langauge $\mathcal{L}^{+}$ can receive one and only one value in $\mathcal{V} = \{1, \frac{1}{2}, 0\}$.

Still, one may wonder how $STTT^{\mathbb{J}}_{\Rightarrow\models}$ sheds light on the status of the Strengthened Liar sentence. The ST counterpart of the Strengthened Liar sentence (18) is this:

(19) This sentence is strictly false.

This sentence is formally represented as $\xi \Rightarrow\models^{STTT^{\mathbb{J}}_{\Rightarrow\models}} \mathbb{J}\neg T\langle\xi\rangle$ (or $\xi \equiv \mathbb{J}\neg T\langle\xi\rangle$). Since ξ always takes the value $\frac{1}{2}$, we can say that (19) is tolerantly true (and tolerantly false). Indeed, by the identity of truth, $T\langle\xi\rangle$ always takes the value $\frac{1}{2}$ as well, which means $\models^{STTT^{\mathbb{J}}_{\Rightarrow\models}} T\langle\xi\rangle$. On the other hand, since $\mathbb{J}\neg T\langle\xi\rangle$ always takes the value 0, we can also say that (19) is strictly false. However, since (19) is tolerantly true, what this means is that (19) just *tolerantly says* of itself that it is strictly false – what (19) says of itself is merely tolerantly true, not strictly true.

Our discussion seems to suggest that Priest is right in saying that the Strengthened Liar sentence is both true and just false (Alternatively, the Strengthened Liar sentence is both tolerantly true and strictly false.) Yet, this does not mean that we do not have a notion of just false (just true) which is different from the notion of falsity (truth). We have. And we can have such a notion without collapsing into triviality.

5.5 Conclusion

Typically, the self-referential behavior of of paradoxical sentences is represented by using the biconditional or the semantic equivalence. We showed that $LPTT$ extended with a Just True operator cannot deal with the revenge via the semantic equivalence, although it can deal with the revenge via the biconditional. On the other hand, we show that $STTT$ extended with a Just True operator can deal with the revenge via the biconditional and the revenge via the semantic equivalence.

Chapter 6

The Exclusion Problem and Pragmatic Implicatures

It has been argued that dialetheists have trouble in expressing disagreement to their opponents. This is known as the *exclusion problem*. According to Priest (2006), dialetheists can make use of pragmatic implicatures to communicate what they disagree about to their opponents. However, Shapiro (2004) casts doubt upon this suggesetion. He suspects that implicatures do not act upon embedded sentences.

In this chapter, we present some linguistic evidence that implicatures can arise at a sub-sentential level. The main task of this chapter is to present a pragmatic interpretation rule which is based on an exact truthmaker semantics. Such a pragmatic interpretation rule has several theoretical benefits. Firstly, it accounts for how dialetheists communicate disagreement to their opponents through implicatures. Secondly, it also accounts for how implicatures act upon embedded sentences. Thirdly, it accounts for many cases of exhaustive interpretation as well.

6.1 The Exclusion Problem and Some Attempted Solutions

6.1.1 The Exclusion Problem

In this chapter, we discuss an objection to dialetheism which can be called the *exclusion problem*: several authors (e.g., Parsons 1990, Littman & Simmons 2004, Shapiro 2004) argue that dialetheists have no way of expressing their disagreement to non-dialetheists. Parsons (1990) poses the objection as follows:

> Suppose that you say [A], and Priest replies [‘$\neg A$’]. Under ordinary circumstances you would think that he had disagreed with you. But then you remember that Priest is a dialetheist, and it occurs you that he might very well agree with you after all – since he might think that [A] and [$\neg A$] are *both* true. How can he indicate that he genuinely disagrees with you? The natural choice is for him to say ‘[A] is not

> true'. However, the truth of this assertion is also consistent with [A]'s being true – for a dialetheist, anyway. (ibid, p.345)

As Shapiro (2004) points out, *consistent* in the last sentence of this passage is a misnomer, because, even for dialetheists, 'A is not true' is inconsistent with A. Nevertheless, Parsons' problem can be reformulated. It seems what Parsons has in mind is that, for dialetheists, $\neg T\langle A\rangle$ (or $\neg A$) is logically compatible with A, that is, $\neg T\langle A\rangle$ (or $\neg A$) and A can obtain together. Because of this, asserting $\neg T\langle A\rangle$ (or $\neg A$) does not rule out A. Parsons thereby concludes that dialetheists '[have] difficulty asserting disagreement with other's views' (ibid).

One apparent way for dialetheists to indicate disagreement with A is to assert 'A is just false'. However, paraconsistent dialetheists have difficulty in interpreting what 'A is just false' means. For one thing, if 'A is just false' is formalized as $T\langle A\rangle \wedge \neg T\langle \neg A\rangle$, this can still be a contradiction in LP theories. Thus, $T\langle A\rangle \wedge \neg T\langle \neg A\rangle$ cannot get the desired interpretation. For another thing, if a LP theory of truth is extended with a Just False operator $\mathbb{F}$ such that $\mathbb{F}A$ takes the value 1, only if A takes the value 0 (otherwise, $\mathbb{F}A$ takes the value 0.), then we can construct a sentence ξ which is intersubstitutable for $\mathbb{F}T\langle \xi\rangle$. Then, the theory in question will be trivial. Thus, it appears that paraconsistent dialetheists have trouble in spelling out their disagreement with non-dialetheists.

On the other hand, strict-tolerant dialetheists are in a better position. Because $STTT$ augmented with the just true operator $\mathbb{J}$ does not suffer from revenge paradoxes: $STTT^{\mathbb{J}}$ can resist the revenge via the biconditional, as well as the revenge via the semantic equivalence. Moreover, in $STTT^{\mathbb{J}}$, a Just False operator $\mathbb{F}$ can be defined in terms of the Just True operator $\mathbb{J}$: $\mathbb{F}A =_{df} \mathbb{J}\neg A$.

In any case, dialetheists need not assert 'A is just false' to express their disagreement with A. Priest (2006) suggests that dialetheists can communicate what they disagree about to their opponents through pragmatic implicatures. The rough idea of this suggestion is that although $\neg A$ (or $\neg T\langle A\rangle$) is compatible with A in dialetheic theories, the assertion of $\neg A$ will implicate the fact that the speaker does not accept A. It is because if the speaker believed that $A \wedge \neg A$, he would have said so. The main task of this chapter is to make this idea precise. Before doing so, we first survey some other proposals to deal with the exclusion problem.

6.1.2 Arrow Falsum

Priest (2006) proposes that if dialetheists want to communicate the fact that they do not accept A, they can assert:

$$A \to \bot$$

where $\to$ is a detachable conditional, and $\bot$ is usually taken as $\forall x T(x)$. This proposal is often called the *arrow-falsum strategy*.

As Field (2008) notes, the arrow-falsum strategy suffers from Curry's paradox. Curry's paradox involves a sentence which says itself that if it is true, then anything is true. We can represent the Curry sentence as:

$$\kappa =\!|\!\models T\langle \kappa\rangle \to \bot$$

Firstly, notice that, on the dialetheic approaches, κ cannot be true. Suppose for reductio that we have $T\langle\kappa\rangle$. Since the conditional $\rightarrow$ is detachable, it follows that $\bot$. Thus, dialetheists have to rule out κ.

Then, according to the arrow-falsum strategy, to rule out κ is to assert $\kappa \rightarrow \bot$. Since κ and $T\langle\kappa\rangle$ are intersubstitutable, it follows that $T\langle\kappa\rangle \rightarrow \bot$. However, by the definition of κ, $T\langle\kappa\rangle \rightarrow \bot$ gives us κ, which amounts to $T\langle\kappa\rangle$. Since the conditional $\rightarrow$ is detachable, we have $\bot$. Thus, rejecting κ commits dialetheists to triviality, which is undesirable.

Field suggests that dialetheists might extend the original strategy to cope with κ. The new proposal is as follows. To rule out A (including κ), we can assert that:

$$(A \rightarrow \bot) \vee (A \rightarrow (A \rightarrow \bot))$$

Unfortunately, the new schema is subject to the Curry's revenge. To see this, we construct a new Curry sentence:

$$\kappa_2 \eqslantless\models (T\langle\kappa_2\rangle \rightarrow \bot) \vee (T\langle\kappa_2\rangle \rightarrow (T\langle\kappa_2\rangle \rightarrow \bot))$$

According to the new proposal, to rule out κ_2, we have to assert:

$$(\kappa_2 \rightarrow \bot) \vee (\kappa_2 \rightarrow (\kappa_2 \rightarrow \bot))$$

Since κ_2 is intersubstitutable for $T\langle\kappa\rangle$, we have:

$$(T\langle\kappa_2\rangle \rightarrow \bot) \vee (T\langle\kappa_2\rangle \rightarrow (T\langle\kappa_2\rangle \rightarrow \bot))$$

By the definition of κ_2, this amounts to κ_2. Thus, we have $T\langle\kappa_2\rangle$. Suppose that $T\langle\kappa_2\rangle \rightarrow \bot$. By modus ponens, we have $\bot$. Suppose that $T\langle\kappa_2\rangle \rightarrow (T\langle\kappa_2\rangle \rightarrow \bot)$. By modus ponens, we have $(T\langle\kappa_2\rangle \rightarrow \bot)$. Applying modus ponens again, we have $\bot$. So either way, we have $\bot$.

Accordingly, to rule κ_2 out, we need a scheme which is even more weaker:

$$(A \rightarrow \bot) \vee (A \rightarrow (A \rightarrow \bot)) \vee (A \rightarrow (A \rightarrow (A \rightarrow \bot)))$$

Yet, as expected, we can construct a new Curry sentence:

$$\kappa_3 \eqslantless\models (T\langle\kappa_3\rangle \rightarrow \bot) \vee (T\langle\kappa_3\rangle \rightarrow (T\langle\kappa_3\rangle \rightarrow \bot)) \vee (T\langle\kappa_3\rangle \rightarrow (T\langle\kappa_3\rangle \rightarrow (T\langle\kappa_3\rangle \rightarrow \bot)))$$

By familiar reasoning, we have $\bot$ again. Field suggests that dialetheists might construct a transfinite sequence of arrow-falsum schema, which mirrors his hierarchy of determinacy operators. Yet, Priest (2010a) claims that he does not want to appeal to such a hierarchy.

The second problem of Priest's arrow falsum strategy is that, as Shapiro (2004) points out, we often communicate our disagreement with A not by asserting that if A is true, then everything is true:

> In Priest's framework, the closest anyone can come to asserting something incompatible with $[A]$ is to say that if $[A]$, then $\forall x Tx$. Even this is logically compatible with $[A]$, as above, but surely $\forall x Tx$ is absurd, if anything is. There is no room (yet) for a more mild form of disagreement. Intuitively, saying 'I think, or suspect, that you may be wrong' is not the same as claiming that if you are right, then everything is true. (ibid, p. 339)

To illustrate the problem, suppose that a dialetheist may want to disagree that:

(20) There will be a sea battle tomorrow.

However, it is too strong for him to say that triviality ensues from (20). After all, even if (20) holds, the dialetheist may still believe that not everything is true.

6.1.3 Shriek Rules

Beall (2013) proposes another strategy which is similar to Priest's arrow falsum strategy. He suggests that when a dialetheist takes A to be consistent and wants to rule out A, he should add a shriek rule (i.e., a non-logical, theory-specific rule) to the theory in question. The shriek rule for A is $A \wedge \neg A \models \bot$. Once the shriek rule for A has been added to the theory, asserting $\neg A$ will rule out A out in the sense that if A holds, triviality would follow.

Parenthetical Remark. Beall uses $\vdash$, rather than $\models$. But since we focus on model theories, we use the semantic entailment $\models$ to represent the inferential relation in the shriek rules.

However, like the arrow falsum strategy, Beall's strategy is too strong (Young, 2015b). Consider:

(21) All of Nixon's utterances about Watergate are true.

Recall that the status of (21) depends on contingent factors: in some unusual circumstances, (21) is paradoxical. So, for dialetheists, there are some non-trivial circumstances in which (21) and its negation hold. Still, a dialetheist may take (21) to be consistent and want to rule it out. However, according to Beall's strategy, to rule out (21), a dialetheist should add a shriek rule such that (21) and its negation *entails* triviality. In other words, the shriek rule for (21) does not allow for the non-trivial circumstances where (21) and its negation hold. This is undesirable.

6.1.4 Primitive Exclusion and Absolute Contradiction

Berto (2014) suggests that a primitive notion of exclusion can be used to formulate a notion of absolute contradiction. He hopes that these notions can provide a clear basis for discussions between dialetheists and their opponents.

Berto's proposal begins by arguing that exclusion cannot be understood in terms of other notions:

> Exclusion should be taken as a primitive concept with a general metaphysical import. There are reasons for so taking it. First, that there must be primitive notions is uncontroversial: were all notions definable in terms of others, we would face either a bad infinite regress, or a (large) circulus in definiendo... Definitions have to come to an end...
>
> [Exclusion] is so basic to our experience of the world... It is likely to show up in the most rudimentary thing new-borns learn to do: distinguishing objects, recognizing a border between something and something else, or acknowledging that this thing's being here rules out its simultaneously being there. We know that if an ordinary material object is uniformly green, it cannot simultaneously be uniformly red; that if it's shorter than one inch then it cannot be longer than a mile. The notion is shared by the dialetheist, of course – we had examples of exclusion from his mouth ... for instance, x's catching the bus and x's *simul*, *sub eodem* missing the bus. (ibid, p.199)

Formally, we can represent the primitive exclusion relation by an operator $\star$:

- We have a primitive exclusion operator $\star$ such that $P \star Q$ means that the property P is incompatible with the property Q.

The primitive exclusion operator $\star$ allows us to talk of an incompatibility set for the property P:

- an incompatibility set for the property P: $\mathrm{I}_P = \{Q \mid Q \star P\}$

According to Berto, we can define the minimal incompatible property $\underline{P}$ of the property P in terms of I_P. The idea of the underline operator _ is that the operator _ takes a given property P as input, and outputs the minimal incompatible property $\underline{P}$ of the property P. Berto defines $\underline{P}$ as follows:

- If I_P is finite, then $\underline{P} = \bigvee\{Q \mid Q \star P\} = Q_1 \vee ... \vee Q_n$, where $Q_1, ..., Q_n \in \mathrm{I}_P$.
- if I_P is infinite, then $\underline{P}x =_{df} \exists Q(Qx \wedge P \star Q)$.

However, as Arenhart (2019) points out, it seems unclear what would be a *minimal* incompatible property of a given property P; as Berto does not provide any hint how to define an incompatible ordering between properties. It is also not clear that there would be a unique minimal property. (See also Arenhart (2019) for the critique of the definition of $\underline{P}$.)

For the moment, we set aside the details of $\underline{P}$, and see what Berto does with the underline operator _. According to Berto, we should have:

- $Px \vee \underline{P}x$ fails.
- It is never the case that $Px \wedge \underline{P}x$ for any property P and object x.

Firstly, Berto claims that $Px \vee \underline{P}x$ fails, because a given object may fail to have the property P, and does not have any property ruling out P. Secondly, according to Berto, we can define the notion of absolute contradiction as $Px \wedge \underline{P}x$; since nothing can have a property and anything ruling out that property.

Berto suggests that the notion of absolute contradiction can serve as a common ground in discussions between dialetheists and their opponents:

> Contradictions in the old negation-involving sense can be true for the dialetheist, a relevant case being provided by the various Liars (and their negations); but no absolute contradiction can. We have, in this sense, some unquestionable ground in the debate on dialetheism: a notion of contradiction... unacceptable by any involved party for any x and P. (Berto, 2014, p. 202)

Berto also claims that the underline operator _ and the notion of exclusion allow dialetheists and their opponents to communicate disagreement to each other:

> [B]y means of "_" we can express in a nonquestion-begging fashion exactly what the divergence between dialetheists and their rivals on the concept of truth consists in – thus making implausible the view that foes and friends of consistency are normally talking past each other, or that either party is just victim of a conceptual confusion, on this issue. For in general the disagreement between dialetheists and supporters of consistency has to do with the extension of a notion (whose intension) they both grasp and share: the notion of exclusion. (ibid)

Now we turn to the evaluation of Berto's proposal. The key question is: once the underline operator _ is added to a dialetheic theory of truth, does the theory suffer from any revenge paradox? The relevant revenge concerns the sentence which says of itself having some feature incompatible with truth. Formally, this sentence can be represented by:

$$\xi \dashv\models \underline{T}\langle\xi\rangle$$

Berto thinks that ξ does not pose any problem to his proposal:

> [T]he dialetheist can take [ξ] as simply false: [$T\langle\neg\xi\rangle$]; from which follows, because ["T"] is transparent, that is, via [Capture] and [Release], that it should also be taken as not true, [$\neg T\langle\xi\rangle$]. The dialetheist does not have to take [ξ] as a dialetheia or, in general, as having a designated value, just as he does not have to (and had better not) take the Curry sentence as a dialetheia or, in general, as having a designated value... [ξ] just falsely claims to have a truth-excluding feature, and its plain falsity does not entail its having a truth-excluding feature. As in general [$T\langle\neg A\rangle \nvDash \underline{T}\langle A\rangle$] and [$\neg T\langle A\rangle \nvDash \underline{T}\langle A\rangle$], the plain falsity or untruth of [ξ] need not entail an absolute contradiction. (Berto, 2014, p.203)

Berto is right in saying $T\langle\neg A\rangle \nvDash \underline{T}\langle A\rangle$ in general. A case in point is the Liar sentence. Recall that the Liar sentence is represented by stipulating that $\lambda \dashv\models \neg T\langle\lambda\rangle$. So given that λ, we have both $T\langle\lambda\rangle$ (by Capture) and $\neg T\langle\lambda\rangle$ (by the definition of λ). That is, the Liar sentence λ is both true and untrue (or false).

Yet, it is another thing to say that a sentence's plain falsity, construed as having

the value 0, does not entail its having a truth-excluding feature. Suppose that a sentence A is just false. Then, by the identity of truth, $T\langle A\rangle$ is just false as well. On the standard dialetheic approaches to truth, 'just false' is the only candidate to be the property incompatible with the property T. that is, $\underline{T}$. Arenhart (2019) notes:

> [$T\langle A\rangle$] or [$\underline{T}\langle A\rangle$] would fail only if both [$T\langle A\rangle$] and [$\underline{T}\langle A\rangle$] could be just false. But then, A could neither have the property of being true and nor be a glut (so that [$T\langle A\rangle$] fails). In this scenario, the possibility for [$\underline{T}\langle A\rangle$] to fail (be just false) would require that A has some property incompatible with truth that is not 'just false', but rather something like 'gap' or some further truth-value. (ibid, p.12)

Hence, given that a sentence A is just false, $\underline{T}\langle A\rangle$ is true (i.e., having the value 1 or $\frac{1}{2}$).

In general, although $Px \vee \underline{P}x$ fails, we still have :

$$T\langle A\rangle \vee \underline{T}\langle A\rangle$$

To see this, consider:

» Suppose that A takes the value 1 or the value $\frac{1}{2}$. By the identity of truth, $T\langle A\rangle$ takes the value 1 or the value $\frac{1}{2}$. By the usual valuation scheme, $T\langle A\rangle \vee \underline{T}\langle A\rangle$.

» Suppose that A takes the value 0. By the identity of truth, $T\langle A\rangle$ takes the value 0 as well. Hence, since A is 'just false', $\underline{T}\langle A\rangle$ takes the value 1 or the value $\frac{1}{2}$. By the usual valuation scheme, $T\langle A\rangle \vee \underline{T}\langle A\rangle$.

Accordingly, by familar reasoning, we can show that absolute contradiction follows:

1	$T\langle\xi\rangle \vee \underline{T}\langle\xi\rangle$	An instance of $T\langle A\rangle \vee \underline{T}\langle A\rangle$
2	$T\langle\xi\rangle$	Hypothesis
3	ξ	2: Release
4	$\underline{T}\langle\xi\rangle$	3: The definition of ξ
5	$T\langle\xi\rangle \wedge \underline{T}\langle\xi\rangle$	2, 4: $\wedge$-Intro
6	$\underline{T}\langle\xi\rangle$	Hypothesis
7	ξ	6: The definition of ξ
8	$T\langle\xi\rangle$	7: Capture
9	$T\langle\xi\rangle \wedge \underline{T}\langle\xi\rangle$	6, 8: $\wedge$-Intro
10	$T\langle\xi\rangle \wedge \underline{T}\langle\xi\rangle$	1, 5, 9: Reasoning by Cases

Thus, it seems that Berto's proposal is not genuinely revenge-immune.

6.1.5 Pragmatic Solutions: Denials and Implicatures

Apart from the arrow-falsum strategy, Priest (2006) also proposes two different pragmatic solutions to the exclusion problem. The first proposal is that dialetheists can make use of a speech act of denial to communicate what they disagree

about to their opponents. According to Smiley's (1996) bilateralism, denial is a speech act that cannot be understood as asserting the negation of a sentence. Rather, denial is a speech act such that it expresses *dissent* towards sentences. Suppose that a non-dialetheist asserts A and a dialetheist disagrees with that. The dialetheist can communicate the fact that he disagrees with A simply by denying A.

Another proposal is that a dialetheist can communicate what he disagrees about through pragmatic implicatures. For instance, when a dialetheist asserts $\neg A$, the assertion will *implicate* the fact that he does not accept A. Priest says:

> Suppose you say to me 'How many siblings do you have and I reply 'I have two brothers'. This may be true, but the answer is definitely misleading if the whole truth is that I have two brothers and one sister. In virtue of my answer, you may reasonably infer that I have no sisters. In the same way, suppose you ask me whether [A] and I [assert] that [$\neg A$]. If I believed [$A \wedge \neg A$], the answer would be decidedly incomplete. You may reasonably infer, there, that I do not accept [A], though what I say does not entail this. (Priest, 2006, p. 291)

However, Shapiro (2004) casts doubt upon these pragmatic solutions. He contends that denials and implicatures cannot act on embedded sentences:

> The dialetheist (or anyone else) either relies on implicature to get the point across or directly expresses disagreement with denial. But how would a dialetheist formulate a *hypothesis* that someone is mistaken? Suppose that Karl says ['A'] and his dialetheist friend Seymore does not want to disagree (yet), but he wonders if Karl is mistaken. Seymore might want to assert a conditional: 'if Karl is mistaken then [B]'. How can Seymore express this? [...] What are the conversational rules for formulating hypotheses, or for the antecedents of conditionals? Even if there are coherent and useful implicatures concerning hypotheses, they cannot be used to determine the consequences of these hypotheses. So far, we just does not have a statement equivalent to 'Karl is mistaken in asserting [A].' (ibid, p. 339 - 340)

It should be noticed that denials can act on embedded sentences. Incurvati & Schlöder (2017) offer examples of inferences with denied sentences as premisses and conclusions. Consider a context in which a and b are the only socialist candidates in the election. Then, the following inference is acceptable:

(22) a. If the election will not be won by a or b, then we will not have a socialist president.

b. Is it the case that a or b will win the election? No, a or b or c will win.

---

c. Is it the case that we will have a socialist president? No.

Notice that the answer in (22b) cannot be reduced to a negated sentence. Suppose otherwise that the 'no' in (22b) is interpreted as meaning that it is not the case that a or b will win the election. Then the whole answer would be interpreted as saying that c will win. But this is clearly too strong. Hence, (22) cannot be

analyzed as an instance of modus ponens (i.e., $\neg A \supset \neg B, \neg A \models \neg B$). Rather, a better analysis would be:

(23) a. Assert: If not A, then not B
b. Deny: A
---
c. Deny: B

The second example is as follows. Consider a context in which Franz is chairing the seminar. The following inference is plausible:

(24) a. If there is a seminar today, Franz is here.
b. Is Franz here? No, not as far as I know.
---
c. Will there be a seminar talk? No.

Similarly, the answer in (24b) cannot be interpreted as a negated sentence. The speaker of the answer expresses his dissent toward the claim that Franz is here, because he does not know that Franz is here. But he would be uncomfortable with asserting that Franz is not here. Accordingly, (24) cannot be analysed as an instance of modus tollens (i.e., $A \supset B$, $\neg B \models \neg A$). Instead, it should be construed as an instance of the following rule:

(25) a. Assert: if A, then B
b. Deny: B
---
c. Deny: A

As for implicatures, Shapiro emphasizes that we have no theory of how implicature acts on embedded sentences. However, it is important to note that (seeming) implicatures can arise at the level of sub-locutionary constituents. Consider the following examples (Recanati, 2003).

(26) Bill and Jane have three or four children.

(27) a. Bill and Jane got married and had many children.
b. Every father feels happy if his daughter gets married and gives birth to a child; much less if she gives birth to a child and gets married.

(26) implicates that Bill and Jane have exactly three or exactly four children; for if the speaker knows that they had more than exactly three or exactly four children, the speaker would have said so. (27a) implicates that Bill and Jane got married before having children. In (27b), such temporal suggestion also occurs in the antecedent of the conditionals. To account for these cases, some theorists generalize the notion of conversational implicatures. Some theorists classify such cases as pseudo-implicatures. We call cases like (26) and (27) *embedded implicatures*, no matter how theorists account for them.

In any case, Shapiro's worry should at best be conceived as calling for an account of embedded speech acts and an account of embedded implicatures. Incurvati

& Schlöder (2017) offer an analysis of embedded denials. In what follows, we would like to explore the second route: pragmatic implicatures.

Parenthetical Remark. For further discussion of the denials proposal, see Ripley (2015). Ripley argues that the bilateral notion of denials give us the resources to define a new operator $-$ on content such that an assertion of $-A$ is equivalent to a denial of A. Then we can form a new paradox with the operator $-$.

6.2 Pragmatic Implicatures: The Basic Picture

To get a sense of the pragmatic account we will present, it might help to consider a problem of Neo-Gricean pragmatics. On any Neo-Gricean approach to implicatures, we begin with an account of semantic meaning and then formulate a pragmatic mechanism based on the semantics. The pragmatic mechanism takes the semantic meaning of a sentence as input and determines the implicature the sentence gives rise.

The major problem of this approach is that classical logical equivalences cannot give rise to different implicatures. The problem can be posed by the following argument.

(P_1) The pragmatic mechanism takes only semantic meaning as input.

(P_2) If the pragmatic mechanism takes only semantic meaning as input, then sentences with the same semantic content cannot give rise to different implicatures.

Thus, we have:

(C_1) Sentences with the same semantic content cannot give rise to different implicatures.

(P_3) Logical equivalents are sentences with the same semantic content.

Thus, we have:

(C_2) Logical equivalents cannot give rise to different implicatures.

To illustrate the problem, let's consider the following examples.

(28) a. Alice or Bob was there.
b. Alice or Bob or both were there.

It is standardly supposed that (28a) and (28b) are logically equivalent and have the same semantic meaning as each other. But only the former gives rise to the scalar implicature that it is not the case that both Alice and Bob was there. Schulz & van Rooij (2006) call this the *functionality problem*.

Facing the functionality problem, we can reject either one of the premises. Some theorists (e.g., Chierchia et al. 2012) have thereby concluded that (scalar) implicatures should be calculated compositionally. That is, the calculation of (scalar) implicatures should be based upon the syntactic structures of sentences so that

sentences with the same semantic content are allowed to give rise to different implicatures. This view can be called the *local analysis of implicatures*. This strategy amounts to rejecting (P_2). In contrast, globalists insist that (scalar) implicatures are calculated independently of syntax. They might reject (P_1) and allow the pragmatic mechanism to take other factors into account. Alternatively, they might reject (P_3) and provide a semantics that could assign classical logical equivalents with different semantic content.

In what follows, we will primarily follow the approach developed by Cobreros et al. (2015a, 2017a, 2017b) and van Rooij (2017). Their proposal starts from a more fine-grained semantics than standard semantics. The fine-grained semantics makes use of the notion of (exact) truthmakers in the works of van Fraassen (1969) and Fine (2014). In this semantics, classical logical equivalents are allowed to have different semantic contents. Then, we discuss how to define the notion of pragmatic meaning based on the semantics. In our account, the pragmatic meaning of a sentence is calculated locally. In other words, our pragmatic account rejects both (P_3) and (P_2).

6.3 Exact Truthmaker

6.3.1 The Propositional Case

Van Fraassen (1969) proposes a theory of fact to provide a semantic analysis of tautological entailments. In what follows, we will make use of the notion of facts to provide a finer-grained semantics than standard semantics. The basic idea is that sentences are true (or false) in virtue of some facts in the world. For instance, the sentence 'Aristotle is a logician' is true, because the fact that Aristotle is a logician holds. The sentence 'Aristotle is a poet' is false, because the fact that Aristotle is not a poet holds.

We begin with a propositional language. We first define the set of all states of affairs as follows.

Definition 35 (States of Affairs)**.** Let $\mathcal{S}$ be a set of all states of affairs. For each atomic sentence $p \in \mathcal{L}$, there is exactly one state of affairs $\mathbf{p} \in \mathcal{S}$ and exactly one corresponding complement $\overline{\mathbf{p}} \in \mathcal{S}$ for which it holds that $\overline{\overline{\mathbf{p}}} = \mathbf{p}$.

We are primarily interested in the non-empty set of state of affairs. We call them a *fact*. We define the set of all facts as follows.

Definition 36 (Facts)**.** The set of all facts $\mathcal{F} = \wp(\mathcal{S}) - \emptyset$.

There are atomic facts. Suppose that $\mathbf{p}$ and $\mathbf{q}$ are in $\mathcal{S}$. Then $\{\mathbf{p}\}$ and$\{\mathbf{q}\}$ are atomic facts. There are conjunctive facts as well, such as $\{\mathbf{p}, \mathbf{q}\}$.

Sentences can be made true by facts. For instance, the atomic sentence p can be made true by the fact $\{\mathbf{p}\}$ and the fact $\{\mathbf{p}, \mathbf{q}\}$. But the fact $\{\mathbf{p}\}$ is a more minimal truthmaker than the fact $\{\mathbf{p}, \mathbf{q}\}$.

In what follows, we are interested in truthmakers that, to use a phrase of Fine (2014), *exactly* verify a sentence. Clearly, the exact truthmaker for the atomic sen-

tence p is the atomic fact $\{\mathbf{p}\}$. As for conjunctions, an exact truthmaker for $A \wedge B$ is the *fusion* of an exact truthmaker for A and an exact truthmaker for B. For instance, the exact truthmaker for $p \wedge q$ is the conjunctive fact $\{\mathbf{p}, \mathbf{q}\}$, where $\{\mathbf{p}, \mathbf{q}\}$ is the fusion of $\{\mathbf{p}\}$ and $\{\mathbf{q}\}$. We write $s \otimes t$ for the fusion of the fact s and the fact t, where $\otimes$ is an operation on sets such that $X \otimes Y = \{s \cup t \mid s \in X \text{ and } t \in Y\}$. Then, we have:

- s exactly makes $A \wedge B$ true, iff there are some t and u such that t exactly makes A true, u exactly makes B true and $s = t \otimes u$.

A disjunction $A \vee B$ is exactly made true by the facts that exactly make one of the disjuncts true. For instance, the exact truthmakers for $p \vee q$ are the fact $\{\mathbf{p}\}$ and the fact $\{\mathbf{q}\}$. We want to exclude the conjunctive fact $\{\mathbf{p}, \mathbf{q}\}$, because $\{\mathbf{p}, \mathbf{q}\}$ seems too strong to be qualified as an exact verifier for $p \vee q$. Thus, we have:

- s exactly makes $A \vee B$ true, iff either s exactly makes A true, or s exactly makes B true.

The negation case can be analyzed by the notion of exact falsemaker. We say:

- s exactly makes $\neg A$ true, iff s exactly makes A false.

Analogously, the notion of exact falsemaker can be easily analyzed as follows. Just as the exact truthmaker for the atomic sentence p is the fact $\{\mathbf{p}\}$, the exact falsemaker for p is the fact $\{\overline{\mathbf{p}}\}$. As for the exact falsemaker for conjunctions, disjunctions and negations, we have:

- s exactly makes false $A \wedge B$, iff either s exactly makes A false, or s exactly makes B false.
- s exactly makes false $A \vee B$, iff there are some t and u such that t exactly makes A false, u exactly makes B false, and $s = t \otimes u$.
- s exactly makes $\neg A$ false, iff s exactly makes A true.

To summarize, we have:

Definition 37 (Exact Truthmakers and Exact Falsemakers)**.** Let $\mathcal{L}$ be a propositional language. For any atomic sentence p and any A, $B \in \mathcal{L}$, the set of exact truthmakers $|A|^+$ and the set of exact false-makers $|A|^-$ are inductively defined as follows.

$$
\begin{array}{ll}
|p|^+ = \{\{\mathbf{p}\}\} & |p|^- = \{\{\overline{\mathbf{p}}\}\} \\
|\neg A|^+ = |A|^- & |\neg A|^- = |A|^+ \\
|A \wedge B|^+ = |A|^+ \otimes |B|^+ & |A \wedge B|^- = |A|^- \cup |A|^- \\
|A \vee B|^+ = |A|^+ \cup |B|^+ & |A \vee B|^- = |A|^- \otimes |B|^-
\end{array}
$$

where $\otimes$ is an operation on sets such that $X \otimes Y = \{s \cup t \mid s \in X \text{ and } t \in Y\}$.

As for conditionals, we analyze them as the material conditional $A \supset B$. That is, $|A \supset B|^+ = |A|^- \cup |B|^+$ and $|A \supset B|^- = |A|^+ \otimes |B|^-$. Before moving to the language of first-order predicate logic, let's look at some examples:

- $|p|^+ = \{\{\mathbf{p}\}\}$

- $|\neg p|^+ = \{\{\overline{\mathbf{p}}\}\}$
- $|p \vee q|^+ = \{\{\mathbf{p}\}, \{\mathbf{q}\}\}$
- $|p \wedge q|^+ = \{\{\mathbf{p}, \mathbf{q}\}\}$
- $|p \vee (q \vee r)|^+ = |((p \vee q) \vee r)|^+ = \{\{\mathbf{p}\}, \{\mathbf{q}\}, \{\mathbf{r}\}\}$
- $|((p \vee q) \wedge (r \vee s))|^+ = \{\{\mathbf{p}, \mathbf{r}\}, \{\mathbf{p}, \mathbf{s}\}, \{\mathbf{q}, \mathbf{r}\}, \{\mathbf{q}, \mathbf{s}\}\}$
- $|(p \vee (p \wedge q))|^+ = \{\{\mathbf{p}\}, \{\mathbf{p}, \mathbf{q}\}\}$
- $|(p \supset q)|^+ = \{\{\overline{\mathbf{p}}\}, \{\mathbf{q}\}\}$
- $|(\neg(p \supset q))|^+ = \{\{\mathbf{p}, \overline{\mathbf{q}}\}\}$

6.3.2 Predicates and Quantifiers

The treatment for predicates and quantifiers is straightforward. The cases for connectives are exactly the same. All we need is to replace the atomic case and define the cases of quantifiers in the obvious way.

Definition 38 (Qualified Exact Truthmakers and Exact Falsemakers)**.** Let $\mathcal{L}$ be a language of first-order predicate logic. For any atomic sentence $Pa_0, ..., a_n$ and any $A, B \in \mathcal{L}$, the set of exact truthmakers $|A|^+$ and the set of exact false-makers $|A|^-$ are inductively defined as follows.

$\|Pa_0, ..., a_n\|^+ = \{\{\mathbf{Pa_0}, ..., \mathbf{a_n}\}\}$	$\|Pa_0, ..., a_n\|^- = \{\{\overline{\mathbf{Pa_0}, ..., \mathbf{a_n}}\}\}$
$\|\neg A\|^+ = \|A\|^-$	$\|\neg A\|^- = \|A\|^+$
$\|A \wedge B\|^+ = \|A\|^+ \otimes \|B\|^+$	$\|A \wedge B\|^- = \|A\|^- \cup \|A\|^-$
$\|A \vee B\|^+ = \|A\|^+ \cup \|B\|^+$	$\|A \vee B\|^- = \|A\|^- \otimes \|B\|^-$
$\|\forall x Ax\|^+ = \bigotimes_{a \in \mathcal{D}} \|(Ax[a/x])\|^+$	$\|\forall x Ax\|^- = \bigcup_{a \in \mathcal{D}} \|(Ax[a/x])\|^-$
$\|\exists x Ax\|^+ = \bigcup_{a \in \mathcal{D}} \|(Ax[a/x])\|^+$	$\|\exists x Ax\|^- = \bigotimes_{a \in \mathcal{D}} \|(Ax[a/x])\|^-$

Again, let us look at some examples. Suppose that our domain $\mathcal{D} = \{a, b\}$. Then we have:

- $|\forall x Px|^+ = |Pa|^+ \otimes |Pb|^+ = \{\{\mathbf{Pa}, \mathbf{Pb}\}\}$
- $|\exists x Px|^+ = |Pa|^+ \cup |Pb|^+ = \{\{\mathbf{Pa}\}, \{\mathbf{Pb}\}\}$
- $|\forall x(Px \vee Qx)|^+ = |(Pa \vee Qa)|^+ \otimes |Pb \vee Qb)|^+$
  $= \{\{\mathbf{Pa}, \mathbf{Pb}\}, \{\mathbf{Pa}, \mathbf{Qb}\}, \{\mathbf{Qa}, \mathbf{Pb}\}, \{\mathbf{Qa}, \mathbf{Qb}\}\}$
- $|\exists x(Px \wedge \neg Px)|^+ = \{\{\mathbf{Pa}, \overline{\mathbf{Pa}}\}, \{\mathbf{Pb}, \overline{\mathbf{Pb}}\}\}$

6.3.3 Truth Conditional Meaning

Possible Worlds. We can use the set of exact truthmakers $|A|^+$ to define the standard truth-conditional meaning for A, which we write as $[\![A]\!]$. We can think of $[\![A]\!]$ as the set of possible worlds which has an exact truthmaker for A. Formally, we have:

Definition 39 (Classical Semantics)**.**

$$[\![A]\!] =_{df} \{w \in W \mid \exists f \in |A|^{+} : f \subseteq w\}$$

But we need to be clear about what possible worlds are. On standard approaches, possible worlds are taken to be *maximal* and *consistent* facts.

Definition 40 (Possible Worlds)**.** A possible world is a set $w \in W$ of state of affairs such that for each atomic sentence p in the language, either $\mathbf{p} \in w$ or the corresponding complement $\overline{\mathbf{p}} \in w$, but not both.

Then, standard logical consequence can be defined in terms of set-theoretic notions. We say that a set of conclusion Δ is a logical consequence of a set of premises Γ, iff the set of possible worlds that makes every $B \in \Gamma$ true is also the set of possible worlds that makes some $A \in \Delta$ true. Thus, we have:

Definition 41 (Classical Consequence)**.**

$$\Gamma \models \Delta \quad \text{iff} \quad \bigcap_{B \in \Gamma} [\![B]\!] \subseteq \bigcup_{A \in \Delta} [\![A]\!]$$

.

Impossible Worlds. We can also use the set of exact truthmakers $|A|^{+}$ to define LP semantics and ST semantics. But to define these semantics, we need to make some adjustments: we have to allow worlds to be inconsistent. We can also define the the semantics of K_3. To do so, we have to allow worlds to be incomplete. Yet, we follow Cobreros et al. (2015a, 2017a, 2017b) and do not allow incomplete worlds in what follows.

Accordingly, in our framework, there are two kinds of worlds: worlds that are maximally consistent (i.e., possible worlds) and worlds that aren't (i.e., impossible worlds). We say that W is a set of worlds, $P \subseteq W$ is the set of possible worlds and $I = W - P$ is the set of impossible worlds. Hence, we can define impossible worlds I as follows:

Definition 42 (Impossible Worlds)**.** An impossible world is a set $w \in I$ of state of affairs such that for some atomic sentence p in the language, it is the case that both $\mathbf{p} \in w$ and $\overline{\mathbf{p}} \in w$.

Then, we can connect the exact truthmaker semantics to LP (ST) models. For each atomic sentence $p \in \mathcal{L}$, any world $w \in W$, we define:

- $v_w(p) = 1 \quad \text{iff} \quad \mathbf{p} \in w \text{ and } \overline{\mathbf{p}} \notin w \quad \text{iff} \quad \exists f \in |p|^{+} : f \subseteq w \wedge \forall \mathbf{a} \in f : \overline{\mathbf{a}} \notin w.$
- $v_w(p) = \frac{1}{2} \quad \text{iff} \quad \mathbf{p} \in w \text{ and } \overline{\mathbf{p}} \in w \quad \text{iff} \quad \exists f \in |p|^{+} : \exists g \in |p|^{-} : f \cup g \subseteq w.$
- $v_w(p) = 0 \quad \text{iff} \quad \mathbf{p} \notin w \text{ and } \overline{\mathbf{p}} \in w \quad \text{iff} \quad \exists f \in |p|^{-} : f \subseteq w \wedge \forall \mathbf{a} \in f : \overline{\mathbf{a}} \notin w.$

Thus, when we interpret the connectives and quantifiers in the usual way, we have the following relations between exact truthmakers and LP (ST) models. For any sentence $A \in \mathcal{L}$ and any world $w \in W$:

- $v_w(A) = 1 \quad \text{iff} \quad \exists f \in |A|^{+} : f \subseteq w \wedge \forall \mathbf{a} \in f : \overline{\mathbf{a}} \notin w.$

- $v_w(A) = \frac{1}{2}$ iff $\exists f \in |A|^+ : \exists g \in |A|^- : f \cup g \subseteq w$.
- $v_w(A) = 0$ iff $\exists f \in |A|^- : f \subseteq w \wedge \forall \mathbf{a} \in f : \overline{\mathbf{a}} \notin w$.

Alternatively, we can define the notion of tolerant truth and strict truth in terms of facts and worlds. The tolerant truth (meaning) of A is the set of maximal worlds that contains A's exact truthmakers; whereas the strict truth (meaning) of A is the set of maximally consistent worlds that contains A's exact truthmakers. Hence, we have:

Definition 43 (Tolerant Truth)**.**

$$[\![A]\!]^t =_{df} \{w \in W \mid \exists f \in |A|^+ : f \subseteq w\}$$

Definition 44 (Strict Truth)**.**

$$[\![A]\!]^s =_{df} \{w \in P \mid \exists f \in |A|^+ : f \subseteq w\}$$

Then, we can define the notion of consequence for LP theories and ST theories. LP consequence can be construed as the preservation of tolerant truth; whereas ST consequence goes from a strictly true set of premises to a tolerantly true conclusion. Accordingly, we have:

Definition 45 (LP Consequence)**.**

$$\Gamma \models^{LPTT} \Delta \text{ iff } \bigcap_{B \in \Gamma} [\![B]\!]^t \subseteq \bigcup_{A \in \Delta} [\![A]\!]^t$$

Definition 46 (ST Consequence)**.**

$$\Gamma \models^{STTT} \Delta \text{ iff } \bigcap_{B \in \Gamma} [\![B]\!]^s \subseteq \bigcup_{A \in \Delta} [\![A]\!]^t$$

6.4 Pragmatic Meaning

6.4.1 Strongest Meaning Hypothesis

As previously discussed, the purpose of giving a finer-grained semantics than standard semantics is to account for pragmatic implicatures. Specifically, we will define a notion of pragmatic meaning based on the fact-based semantics.

To begin with, we consider a pragmatic principle called the Strongest Meaning Hypothesis (SMH) used by Cobreros et al. (2012). According to the SMH, speakers interpret a sentence in the semantically strongest possible way. In our setting, the SMH comes down to this principle: the pragmatic meaning of A amounts to the set of minimally inconsistent worlds – the set of worlds that contain no more inconsistencies than required to exactly make A true. We say that a world v is less inconsistent than a world w (i.e., $v \ll w$), if the inconsistent states of affairs (i.e., states of affairs that are inconsistent to each other) that v contains is less than that of w contains. Formally, $v \ll w$ iff $_{df} \{x \in SOA : \{x, \overline{x}\} \subseteq v\} \subset \{x \in SOA : \{x, \overline{x}\} \subseteq w\}$. Then, our pragmatic principle can be formulated as follows.

- $Prag_1(A) =_{df} \{w \in [\![A]\!]^t \mid \neg\exists v \in [\![A]\!]^t : v \ll w\}$

Notice that $Prag_1$ amounts to the claim that the pragmatic meaning of A is its strict meaning, if it is possible to interpret strictly; otherwise, A's pragmatic meaning should be interpreted tolerantly. That is, $Prag_1$ can also be formulated as this:

- $Prag_1(A) = \begin{cases} [\![A]\!]^s, \text{ if } [\![A]\!]^s \neq \emptyset. \\ [\![A]\!]^t, \text{ otherwise.} \end{cases}$

$Prag_1$ successfully accounts for some recent experimental data. Some recent studies (e.g., Ripley 2011; Alxatib & Pelletier 2011) show that naive speakers find some logical contradictions such as 'John is smart and not smart' acceptable. $Prag_1$ can account for the acceptability of *John is smart and not smart*: the contradiction $Sa \wedge \neg Sa$ cannot be interpreted as true under the strict notion of truth. Nor can it be interpreted as true in a classical model. Nevertheless, it can be interpreted as true tolerantly.

It is important to notice that the SMH operates at the sentence level. According to the SHM, the entire sentence should be interpreted strictly, if it is possible to do so; otherwise, it should be interpreted tolerantly. However, as Alxatib et al. (2013) note, there are examples where the sentence should in part be evaluated strictly, and in part be evaluated tolerantly. Here's an example of this kind:
(We say that a is the borderline case of S, if Sa is neither strictly true, nor $\neg Sa$ is strictly true. Moreover, we write $A \leftrightsquigarrow B$, if A and B are pragmatically equivalent, that is, have the same pragmatic meaning.)

(29) John is smart and not smart, or Mary is rich.
$\leftrightsquigarrow$ John is borderline smart, or Mary is strictly rich.

To see (29), consider the case in which we know that Mary is not rich at all. Given that John is smart and not smart, or Mary is rich holds, we can conclude that John is borderline smart.

However, the SMH would predict that *John is smart and not smart, or Mary is rich* is equivalent to *Mary is strictly rich*. Specifically, we would have:

- $Prag_1((Sa \wedge \neg Sa) \vee Rb) = Prag_1(Rb)$
$= \{w \in W : v_w(Rb) = 1 \text{ and } v_w(Sa) \neq \frac{1}{2}\}$

(For simplicity, we ignore atomic sentences other than Sa and Rb.)

Let us see how $Prag_1$ interprets John is smart and not smart, or Mary is rich by considering the following case. Suppose that $W = \{w_1, w_2, w_3\}$, where $w_1 = \{\mathbf{Sa}, \overline{\mathbf{Sa}}, \mathbf{Rb}\}$, $w_2 = \{\mathbf{Sa}, \mathbf{Rb}\}$ and $w_3 = \{\mathbf{Sa}, \overline{\mathbf{Sa}}, \overline{\mathbf{Rb}}\}$. Firstly, notice that $(Sa \wedge \neg Sa) \vee Rb$ is at least tolerantly true in all of these worlds. But w_2 is the least inconsistent world; since both w_1 and w_3 contain an inconsistent state of affairs. Thus, in this case, $Prag_1((Sa \wedge \neg Sa) \vee Rb) = \{w_2\}$. However, we have $Prag_1(Rb) = \{w_2\}$ as well. Thus, $Prag_1((Sa \wedge \neg Sa) \vee Rb) = Prag_1(Rb) = \{w \in W : v_w(Rb) = 1 \text{ and } v_w(Sa) \neq \frac{1}{2}\}$. Thus, $Prag_1$ would wrongly interpret *John is smart and not smart, or Mary is rich* as meaning that *Mary is strictly rich*.

According to Alxatib et al.'s (2013) diagnosis, the essential problem of the SHM lies in the fact the SMH strengthens the semantic meaning at the sentence level:

> The SMH is a principle of linguistic pragmatics and therefore is applied at the sentence level. Though one of the disjuncts in ['John is smart and not smart, or Mary is rich'] is a classical contradiction, the full [sentence is] not. The SMH predicts, then, that only the strict interpretation can apply [...] It therefore seems necessary to evaluate sentences in part relative to strict evaluation and in part relative to the tolerant notion of truth. (ibid, 2013, p. 624)

Although the above comment that the principle of linguistic pragmatic must be applied at the sentence level is questionable, it seems that the suggestion that the meaning of sentences is interpreted locally should be followed.

Parenthetical Remark. One important feature of disjunctions is ignored by Alxatib et al (2013): $A \vee B$ gives rise to scalar implicature. If a speaker say $A \vee B$, we can pragmatically infer that not both A and B. It is because the speaker would have said so otherwise. We will come back to scalar implicatures after considering $Prag_2$.

6.4.2 Meaning Strengthening and Exact Truthmaker

Accordingly, Cobreros et al. (2017) propose another pragmatic principle according to which meaning is strengthened locally:

- $Prag_2(A) =_{df} \{w \in W \mid \exists f \in |A|^+ : f \subseteq w \text{ and } \neg\exists v \supseteq f : v <_f w\}$

where $v <_f w$ iff$_{df}$ $\{x \in SOA : x \in f \text{ and } \overline{x} \in v\} \subset \{x \in SOA : x \in f \text{ and } \overline{x} \in w\}$. Notice that $<_f$ is relativised to a specific exact truthmaker f. Intuitively, the notion of $<_f$ means that a world v is less inconsistent than a world w with respects to a specific fact f, iff the f-inconsistent states of affairs (i.e., the states of affairs that are inconsistent to the states of affairs in f) that v contains is less than that of w contains. Thus, according to $Prag_2$, the pragmatic meaning of A is the set of minimally inconsistent worlds with respects to each of A's exact truthmakers.

Let's see how $Prag_2$ accounts for (29).

- $Prag_2((Sa \wedge \neg Sa) \vee Rb) = \{w \in W : v_w(Sa) = \frac{1}{2} \text{ or } v_w(Rb) = 1\}$

(Again, we ignore atomic sentences other than Sa and Rb.)

Consider the following model: $W = \{w_1, w_2, w_3, w_4\}$, where $w_1 = \{\mathbf{Sa}, \overline{\mathbf{Sa}}, \mathbf{Rb}\}$, $w_2 = \{\mathbf{Sa}, \mathbf{Rb}\}$, $w_3 = \{\mathbf{Sa}, \overline{\mathbf{Sa}}, \overline{\mathbf{Rb}}\}$ and $w_4 = \{\overline{\mathbf{Sa}}, \mathbf{Rb}\}$. Firstly, note that $(Sa \wedge \neg Sa) \vee Rb$ has two exact truthmakers: $\{\mathbf{Sa}, \overline{\mathbf{Sa}}\}$ and $\{\mathbf{Rb}\}$. Clearly, for the exact truthmaker $\{\mathbf{Rb}\}$, w_1, w_2 and w_4 are a minimally inconsistent world. (Note that although w_1 contains both **Sa** and $\overline{\mathbf{Sa}}$, it still is a minimally inconsistent world for $\{\mathbf{Rb}\}$; since it does not contain any state of affair that is inconsistent to **Rb**.) For the exact truthmaker $\{\mathbf{Sa}, \overline{\mathbf{Sa}}\}$, w_1 and w_3 are equally inconsistent. Thus, in this case, $Prag_2((Sa \wedge \neg Sa) \vee Rb) = \{w_1, w_2, w_3, w_4\} = \{w \in W : v_w(Sa) = \frac{1}{2} \text{ or } v_w(Rb) = 1\}$. That is, *John is smart and not smart, or Mary is rich* is equivalent to *John is borderline smart, or Mary is strictly rich*.

$Prag_2$ accounts for how dialetheists communicate disagreement by implicatures. For instance, *John is smart* and *John is smart and is not smart* give rise to different pragmatic interpretations. The former means that John is strictly smart; whereas the latter is interpreted as saying that John is borderline smart. Thus, if a dialetheist asserts that John is smart, his opponent can reasonably infer that the dialetheist does not accept that John is not smart; unless the dialetheist asserts that John is not smart explicitly.

Similarly, the pragmatic meaning of *John is not smart* is different from *John is smart and is not smart*: *John is not smart* is interpreted as saying that John is not even tolerantly smart. Unless the speaker explicitly states that John is smart, his assertion that John is not smart should be interpreted as meaning that he does not accept that John is smart. To summarize, we have these cases:

(30) a. John is smart.
$\leftrightsquigarrow$ John is strictly smart.

b. John is not smart.
$\leftrightsquigarrow$ John is not even tolerantly smart

c. John is smart and is not smart.
$\leftrightsquigarrow$ John is borderline smart.

$Prag_2$ successfully predicts these cases. It is easy to check that we have:

- $w \in Prag_2(Sa)$ iff $v_w(Sa) = 1$
- $w \in Prag_2(\neg Sa)$ iff $v_w(Sa) = 0$
- $w \in Prag_2(Sa \wedge \neg Sa)$ iff $v_w(Sa) = \frac{1}{2}$

We also have some cases which are more complicated:

(31) a. John is smart and not smart, and Mary is rich.
$\leftrightsquigarrow$ John is borderline smart, and Mary is strictly rich.

b. John is smart and not smart, or Mary is rich.
$\leftrightsquigarrow$ John is borderline smart, or Mary is strictly rich.

In these cases, the latter conjunct in (31a) and the latter disjunct in (31b) are interpreted strictly. These cases suggest that meaning strengthening qua inconsistency minimization can occur at a sub-sentential level.

$Prag_2$ accounts for the pragmatic meaning of (31a) and (31b). Specifically, we have:

- $w \in Prag_2((Sa \wedge \neg Sa) \wedge Rb)$ iff $v_w(Sa) = \frac{1}{2}$ and $v_w(Rb) = 1$
- $w \in Prag_2((Sa \wedge \neg Sa) \vee Rb)$ iff $v_w(Sa) = \frac{1}{2}$ or $v_w(Rb) = 1$

The general recipe of $Prag_2$ is that it looks for each of the exact truthmakers f of a sentence A, and tries to minimize inconsistencies with respect to f. It is this feature of $Prag_2$ that helps account for how meaning strengthening qua inconsistency minimization occurs at a sub-sentential level.

6.4.3 Meaning Strengthening and Exhaustive Interpretation

Exhaustive Interpretation. So far so good. But cases like (30a), (30b) (31a), (31b) are not the only way how we strengthen the semantic meaning of utterances; we expect that the notion of pragmatic meaning characterizes how we enrich the semantic meaning in general. To get sense of the general phenomenon of meaning strengthening, let us start with some examples. Recall that Priest (2006) motivates the use of pragmatic implicatures to express disagreement by the following example.

(32) How many siblings do you have?
   a. Two brothers.
      $\rightsquigarrow$ only two brothers
   b. Two brothers.
      $\not\rightsquigarrow$ Two brothers and one sister.

In many context, the answer in (32) is interpreted as exhausting the predicate in question. The answer in (32) is interpreted saying that I have two brothers only; one cannot take the answer as saying that I have two brothers and one sister.

Similarly, in the following dialogue, the answer is read as exhausting the predicate *came to the party* (Schulz & van Rooij 2006).

(33) Who came to the party?
   a. John and Mary.
      $\rightsquigarrow$ only John and Mary.
   b. John, or Mary.
      $\rightsquigarrow$ only John, or only Mary.

The answer (33a) is interpreted as saying that John and Mary are the only people that came to the party. The answer (33b) is interpreted as saying that either John or Mary is the only people that came to the party. We call the linguistic phenomenon like (32) and in (33) the *exhaustive interpretation of answers*.

Note that exhaustive interpretation occurs in the case of multiple disjunctions (van Rooij & Schulz, 2004).

(34) Who knows the answer?
   a. Peter, (or) Mary, or Sue.
      $\rightsquigarrow$ only Peter, Mary, or Sue knows the answer.
   b. John or Peter, and Mary, or Sue.
      $\rightsquigarrow$ only John and Mary, or only John and Sue,
      or only Peter and Mary, or only Peter and Sue.

To summarize, we have the following data:

(35) a. $p \vee q$ $\rightsquigarrow$ Not both p and q
   b. $p \wedge q$ $\rightsquigarrow$ Only p and q
   c. Two students passed $\rightsquigarrow$ Exactly two students passed

d. $p \vee q \vee r$ $\rightsquigarrow$ Only one of p, q, and r

e. $(p \vee q) \wedge (r \vee s)$ $\rightsquigarrow$ Only one of $(p \wedge r)$, $(p \wedge s)$, $(q \wedge r)$, or $(q \wedge s)$

Classical Semantics Is Too Coarse-Grained. Before accounting for these data, we have to note that classical semantics is too coarse-grained to account for (some cases of) exhaustive interpretation.

(36) Who came to the party?

a. John $\rightsquigarrow$ not Mary.

b. John, or John and Mary. $\not\rightsquigarrow$ not Mary.

In classical semantics, John came to the party (i.e., p) and John, or John and Mary came to the party (i.e., $p \vee (p \wedge q)$) are equivalent. Hence, if our pragmatic mechanism takes only semantic meaning as input, it cannot account for why the answer (36a) and the answer (36b) should be interpreted differently. The answer (36a) should be interpreted as saying that only John came to the party; whereas the answer (36b) should be interpreted as saying that only John came to the party, or both John and Mary came to the party. One cannot interpret the answer (36b) as saying that only John came to the party. This is precisely the functionality problem we saw in §6.2 of this chapter.

The following examples also highlight the same problem:

(37) How many people passed the examination?

a. Two students passed.
(i.e., $\exists x \exists y(Px \wedge Py \wedge x \neq y)$)
$\rightsquigarrow$ exactly two students passed.

b. Two or three students passed.
(i.e., $\exists x \exists y(Px \wedge Py \wedge x \neq y) \vee \exists x \exists y \exists z(Px \wedge Py \wedge Pz \wedge x \neq y \wedge x \neq z \wedge z \neq z)$)
$\rightsquigarrow$ Exactly two or exactly three students passed.

In classical semantics, the answer (37a) and the answer (37b) are equivalent. However, the former implicates that exactly two students passed; no more, no less. The latter implicates that exactly two or exactly three students passed.

It should be noted that exhaustive interpretation also interacts with determiners (Schulz & van Rooij 2006).

(38) How many people passed the examination?

a. Two students passed .
$\rightsquigarrow$ exactly two students passed.

b. At least two students passed.
$\not\rightsquigarrow$ Exactly two students passed.

In this example, we see that exhaustive interpretation is sensitive to *at least*. Without this determiner, one can infer from *two students passed* that exactly two students passed. (But notice that *at least* does not cancel all sorts of exhaustive interpretation: one can still infer from *at least two students passed* that nobody besides

students passed.) Nevertheless, in classical semantics, the answer (38a) has the same meaning as (38b).

To sum up, the following data cannot be accounted by any explanation based on classical semantics, whether the explanation is semantic or pragmatic in nature:

(39) a. $p \vee (p \wedge q)$ $\rightsquigarrow$ Only p, or (only) $p \wedge q$

b. 2 or 3 students passed passed $\rightsquigarrow$ Exactly 2 or exactly 3 students

c. At least 2 students passed $\not\rightsquigarrow$ Exactly 2 students passed

A Pragmatic Account of Exhaustive Interpretation. We begin by discussing the general skeleton of our pragmatic account. Firstly, the pragmatic account must account for meaning strengthening qua inconsistency minimization. In particular, it must preserve the explanatory power of $Prag_2$ and accounts for (30) and (31). Secondly, it must account for exhaustive interpretation. That is, it must accounts for (35) and (39). Van Rooij (2017) offers a global analysis for exhaustive intepretation based on the exact truthmakers semantics we previously introduecd. Our pragmatic account will based on $Prag_2$ and van Rooij's work.

According to $Prag_2$, when we interpret a sentence, we first find out the exact truthmakers f of it. Then, we look for the minimally inconsistent worlds with respects to each f. There is nothing wrong with the direction of $Prag_2$. To see this, consider $p \vee (p \wedge q)$. Suppose that $W = \{w_1, w_2, w_3, w_4, w_5\}$, where $w_1 = \{\mathbf{p}, \mathbf{q}\}$, $w_2 = \{\mathbf{p}, \overline{\mathbf{q}}\}$, $w_3 = \{\overline{\mathbf{p}}, \mathbf{q}\}, w_4 = \{\mathbf{p}, \overline{\mathbf{p}}, \mathbf{q}\}, w_5 = \{\mathbf{p}, \mathbf{q}, \overline{\mathbf{q}}\}$. Again, we ignore atomic sentences other than p and q. First, notice that $|p \vee (p \wedge q)|^+ = \{\{\mathbf{p}\}, \{\mathbf{p}, \mathbf{q}\}\}$. Then, we look for the minimally inconsistent worlds for the exact truthmakers $\{\mathbf{p}\}$ and $\{\mathbf{p}, \mathbf{q}\}$. For $\{\mathbf{p}\}$, w_1, w_2, and w_5 are minimally inconsistent worlds. For $\{\mathbf{p}, \mathbf{q}\}$, w_1 is the minimally inconsistent world. There is nothing wrong to look for the minimally inconsistent worlds for each exact truthmaker.

However, as far as the exact truthmaker $\{\mathbf{p}\}$ is concerned, what we are after is the worlds where only p is (strictly) true. In this sense, w_2 is more minimal than w_1 and w_5 for $\{\mathbf{p}\}$: whereas both w_1 and w_5 contain more positive state of affair than necessary to make p true, w_2 contains just enough to make p true. This means that some of the minimally inconsistent worlds are not sufficiently minimal. Accordingly, among these minimally inconsistent worlds, we should continue to look for the minimal worlds such that only $(|f|)$ can be interpreted as true as possible, where $(|f|)$ is the proposition that corresponds to f.

Now let's formalize the idea. We say that a world v is more $(|f|)$-minimal than a world w, iff the positive state of affairs which are not in f that v contains is less than that of w contains . Formally, $v \prec_{(|f|)} w$ iff $_{df} \{x \in SOA : x \notin f \text{ and } x \in v\} \subset \{x \in SOA : x \notin f \text{ and } x \in w\}$. Then, we define the pragmatic meaning of A to be the set of worlds w such that w is a minimally inconsistent world with respects to each of A's exact truthmakers f and w is the $(|f|)$-minimal world among the minimally inconsistent worlds.

Definition 47 (Pragmatic Meaning)**.**

$$\begin{aligned} Prag(A) =_{df} \{w \in W \mid & \exists f \in |A|^{+} : f \subseteq w \text{ and} \\ & \neg\exists v \supseteq f : v \in W \wedge v <_{f} w \text{ and} \\ & \neg\exists u \supseteq f : (\neg\exists v \subseteq f : v <_{f} u) : u \in W \wedge u \prec_{(|f|)} w\} \end{aligned}$$

Let's see how our new pragmatic interpretation rule accounts for $p \vee (p \wedge q)$. To continue our previous example, notice that, for $\{\mathbf{p}\}$, w_2 is the minimally inconsistent world such that it does not contain more positive state of affairs than necessary to make p true. Moreover, notice that, for $\{\mathbf{p}, \mathbf{q}\}$, w_1 is the minimal world we are after. Thus, $Prag(p \vee (p \wedge q)) = \{w_1, w_2\}$.

It is helpful to take a look at other examples. Since literals and conjunctive sentences have only one exact truthmaker f, our pragmatic interpretation rule looks for the worlds that make f as true as possible and contains no more than necessary positive state of affairs to make $(|f|)$ true.

- $Prag(p) = \{w \in W : v_w(p) = 1\}$
- $Prag(\neg p) = \{w \in W : v_w(p) = 0\}$
- $Prag(p \wedge \neg p) = \{w \in W : v_w(p) = \frac{1}{2}\}$
- $Prag((p \wedge \neg p) \wedge q) = \{w \in W : v_w(p) = \frac{1}{2} \text{ and } v_w(q) = 1\}$

As for disjunctive sentences, they have more than one exact truthmaker. Our pragmatic interpretation rule singles out the minimally inconsistent worlds for each exact truthmaker f that only makes $(|f|)$ true.

- $Prag(p \vee \neg p) = \{w \in W : v_w(p) = 1 \text{ or } v_w(p) = 0\}$
- $Prag(p \vee q) = \{w \in W : (v_w(p) = 1 \text{ and } v_w(q) = 0) \text{ or } (v_w(p) = 0 \text{ and } v_w(q) = 1)\}$
- $Prag((p \wedge \neg p) \vee q) = \{w \in W : (v_w(p) = \frac{1}{2} \text{ and } v_w(q) = 0) \text{ or } (v_w(p) = 0 \text{ and } v_w(q) = 1)\}$

Let us illustrate our pragmatic interpretation rule $Prag$ with $((p \wedge \neg p) \vee q)$. To see the pragmatic meaning of $((p \wedge \neg p) \vee q)$, first notice that $((p \wedge \neg p) \vee q)$ has two exact truthmakers: $\{\mathbf{p}, \overline{\mathbf{p}}\}$ and $\{\mathbf{q}\}$. According to our pragmatic principle $Prag$, the second step is to look for the minimally inconsistent worlds with respects to these exact truthmakers. For the exact truthmaker $\{\mathbf{p}, \overline{\mathbf{p}}\}$, the minimally inconsistent worlds are the worlds where p takes the value $\frac{1}{2}$. For the exact truthmaker $\{\mathbf{q}\}$, the minimally inconsistent worlds are the worlds where q takes the value 1.

The third step is to look the minimal worlds for $p \wedge \neg p$ and the minimal worlds for q. Since the minimal worlds for $p \wedge \neg p$ do not contain any positive state of affairs than necessary to make $p \wedge \neg p$ true, such worlds do not contain any positive state of affair to make q true. Thus, the minimal worlds for $p \wedge \neg p$ are the worlds where p takes the value $\frac{1}{2}$ and q takes the value 0. Similarly, the minimal worlds for q are the worlds where q takes the value 1 and p takes the value 0; because the minimal worlds for q do not contain any positive state of affair to make p true. Then, we finish calculating the pragmatic meaning of $((p \wedge \neg p) \vee q)$: the pragmatic meaning

of $((p \wedge \neg p) \vee q)$ amounts to the minimal worlds for $p \wedge \neg p$ and the minimal worlds for q.

In summary, our pragmatic principle can be construed as an algorithmic procedure. The algorithm is as follows. To calculate the pragmatic meaning of A, we have to:

i. look for each of the exact truthmakers f of A.

ii. look for the minimally inconsistent worlds for each truthmaker f.

iii. look for the $(|f|)$-minimal worlds among the minimally inconsistent worlds for each f. The $(|f|)$-minimal worlds are the worlds where only the proposition $|f|$ which corresponds to f is interpreted as strictly true, if it is possible to so interpret. If it is not possible to interpret $|f|$ as strictly true, then $|f|$ is interpreted as tolerantly true.

6.5 Conclusion

To conclude the chapter, we point out some theoretical benefits of our pragmatic interpretation rule $Prag$. Firstly, the pragmatic interpretation rule accounts for how a dialetheist can communicate his disagreement to his opponent. For instance, when a dialetheist asserts $\neg p$, what he is trying to say is that p is strictly false. The interlocutor of the dialetheist can reasonably assume or interpret that the dialetheist does not think that $p \wedge \neg p$ is the case.

Secondly, the pragmatic interpretation rule $Prag$ accounts for how implicatures arise at a sub-sentential level. As we saw, $Prag$ accounts for embedded implicatures of negations, disjunctions and conjunctions, which means that it accounts for embedded implicatures of the whole language.

The pragmatic interpretation rule $Prag$ can answer Shapiro's (2004) skepticism about the use of implicatures. Recall that Shapiro suspects that implicatures do not act on the antecedents of conditionals. Shapiro (2004, p. 340) also claims that 'even if there are coherent and useful implicatures concerning hypotheses, they cannot be used to determine the consequences of these hypotheses'. To answer Shapiro's skepticism, consider the following conversation:

(40) a. Liverpool is going to win the Premier League 2019.
b. Liverpool will achieve a record points total if they are going to win the Premier League 2019.

Suppose that someone asserts (40a). The dialetheist interlocutor has his doubt about (40a). So the dialetheist asserts (40b). This seems to be an ordinary football conversation. Let (40a) be p and (40b) be $p \supset q$. Now let us calcaute the pragmatic meaning of the dialetheist's assertion (40b). According to our pragmatic intepretation rule, we first notice that $|p \supset q|^+ = \{\{\overline{\mathbf{p}}\}, \{\mathbf{q}\}\}$. Then, our pragmatic interpretation rule minimizes inconsistencies for $\{\overline{\mathbf{p}}\}$ and $\{\mathbf{q}\}$. Finally, our pragmatic interpretation rule looks for the worlds where $\neg p$ is strictly true and q is strictly false, as well as the worlds where only q is strictly true. Then, we have:

- $Prag(p \supset q) = \{w \in W : (v_w(p) = 0 \text{ and } v_w(q) = 0) \text{ or } (v_w(p) = 0 \text{ and } v_w(q) = 1)\}$

This seems to capture the intended meaning of (40b). For one thing, $Prag(p \supset q)$ does not contain any inconsistent world. For another thing, $Prag(p \supset q)$ does not contain any world where p is strictly true. This precisely captures the dialetheist's doubt about (40a).

Thirdly, our pragmatic interpretation rule $Prag$ accounts for many cases of exhaustive interpretation. Recall that Priest (2006) motivates the use of pragmatic implicatures by drawing an analogy between exhaustive interpretation of answers and meaning strengthening qua inconsistency minimization. Our pragmatic interpretation accounts for both phenomena: the first and the second step of our pragmatic interpretation algorithm deals with meaning strengthening qua inconsistency minimization; while the third step deals with exhaustive interpretation of answers.

Chapter 7

Conclusion

It's time to take stock. In the thesis, we provide two dialetheic solutions for the issue concerning the Liar paradox, expressive limitations and revenge paradoxes. One is the strict-tolerant solution. Another is the pragmatic solution. We first discuss the former.

7.1 The Strict-Tolerant Solution

In chapter 1, we distinguished two different projects, corresponding to the Liar paradox in natural languages and its formal counterpart. One is the semantic characterization project. The Liar paradox makes us wonder how we can exhaustively and exclusively characterize all meaningful and declarative sentences in natural languages. So we are interested in explaining how we can so characterize sentences. Or at least, if sentences cannot be exhaustively and exclusively characterized, we want an explanation why sentences cannot be so characterized.

Another project is the non-triviality project. If a logic obeys some familiar classical laws and respects the naive principles of truth, it can be shown by liar reasoning that the logic is trivial. Accordingly, formal theorists of truth are interested in showing how a formal language which allows for self-reference is expressive enough to express truth (as well as some important semantical notions) without being trivialized by the Liar paradox (and it's revenge).

In what follows, we evaluate the strict-tolerant solution with respect to both projects. We begin with the non-trivality project.

7.1.1 The Non-Triviality Project

It is often suggested that revenge paradoxes show that any theory of truth must face a dilemma. The first horn of the dilemma is to admit that the theory in question is expressively incomplete. The second horn of the dilemma is that if the theory is augmented with extra connectives so that the theory can express some important semantical notions, the augmented theory must be trivialized by revenge paradoxes. As far as the non-triviality project is concerned, we argued for two conclusions, which can be formulated as:

i. Both the paracomplete gap approaches and the paraconsistent dialetheic approaches must face such a dilemma. That is, both the paracomplete gap approaches and the paraconsistent dialetheic approaches are either expressively incomplete, or trivialized by revenge paradoxes.

ii. Among the theories we considered, the strict-tolerant dialetheic approaches are the only approaches that can escape the dilemma: $STTT$ augmented with a Just True operator $\mathbb{J}$ can resist the revenge via the material biconditional, and the revenge via the semantic equivalence.

In chapter 3, we saw that K_3TT cannot report the status of the Liar sentence. The claim that the Liar sentence is neither true nor false cannot be true (i.e., cannot have the value 1) in K_3TT, if the claim is formalized as $\neg(T\langle\lambda\rangle \vee T\langle\neg\lambda\rangle)$. We saw that if K_3TT is augmented with a (bivalent) determinacy operator $\mathbb{D}$ which is defined as:

- $v_{\mathcal{M}^{+\mathbb{D}}}(\mathbb{D}A) = \begin{cases} 1, \text{ if } v_{\mathcal{M}^{+\mathbb{D}}}(A) = 1. \\ 0, \text{ if } v_{\mathcal{M}^{+\mathbb{D}}}(A) \neq 1. \end{cases}$

then, a Strengthened Liar sentence can be formed by the semantic equivalence relation in the augmented theory $K_3TT^{\mathbb{D}} : \xi \dashv\models^{K_3TT^{\mathbb{D}}} \neg\mathbb{D}\xi$. Then, we can show that $K_3TT^{\mathbb{D}}_{\dashv\models}$ is trivial by some familiar paradoxical reasoning. (Notice that K_3-based theories cannot use the material biconditional $\equiv$ to represent self-reference. For instance, the Liar sentence cannot be represented by stipulating that $\lambda \equiv \neg T\langle\lambda\rangle$, since $\lambda \equiv \neg T\langle\lambda\rangle$ cannot be true.)

In chapter 4, we saw that Field's theory tries to improve K_3TT. His non-idempotent and non-bivalent determinacy operator D is primarily motivated by the intuition that paradoxical sentences are in some sense defective. For instance, the Liar sentence λ_0 can be characterized as not determinately true and not determinately false. The Strengethened Liar sentence λ_1 can be characterized as not determinately determinately true and not determinately determinately false.

However, while Field's theory is able to characterize paradoxical sentences, it seems that the theory should still be counted as expressively incomplete. The theory does not allow for certain semantical notions such as bivalent determinateness, a general notion of truth-value gaps and exclusion negation. Moreover, adding extra connectives/predicates to model such notions will give rise to revenge paradoxes, trivializing the augmented theory. Nevertheless, Field raises a number of philosophically complicated arguments to justify for choosing the expressive incompleteness horn. We argued against those arguments in chapter 3 and 4. In chapter 3, we saw that Field argues that notions like bivalent determinateness are unmotivated. According to Field, we cannot motivate bivalent determinateness by the model-theoretic notion of designatedness, because the extension of the value 1 and the extension of model-independent truth diverge. However, semantical notions like bivalent determinateness, truth-value gaps and exclusion negation need not to be motivated by the characteristics of the model-theoretic semantics. They can be motivated by philosophical theories such as Strawson's analysis of presuppositions. They can also be motivated by empirical evidence. There is evidence for the fact that exclusion negation is the way we use

'not' in usual situations, as we saw in chapter 4.

Field insists semantical notions like bivalent determinateness and exclusion negation are not coherent. It is because we do not need such notions to characterize sentences: we can use the non-idempotent and non-bivalent determinacy operator D to characterize sentences. Moreover, such notions lead to paradoxes, if Field's theory is augmented with extra connectives/predicates to model such notions. At this point, it seems what Field tries to suggest is that if a notion breeds paradoxes for a theory, then it is not coherent (or not understandable for anyone advocating the theory). As Rossi (2018) points out, such a suggestion is 'deeply problematic':

> [The suggestion] turns any attempt to compare theories into a futile exercise: the advocates of [a theory] would declare every notion incompatible with [the theory] to be simply nonsense. Several fundamental debates in theories of truth – including the debates on which is the 'right' non-classical logic of naive truth, and the debates on whether truth is naive – would also have to be considered completely pointless. (ibid, p. 12)

> [The suggestion] applies across the board, and not just to revenge-breeding semantic notions. Its application to naive truth shows that this notion was not sufficiently understandable before the development of suitable non-classical theories in relatively recent times. At the very least, [the suggestion] entails that naive truth was not sufficiently understandable when it was first formulated as a limitative result (see e.g. Tarski 1936) in that it yields triviality once added to the accepted formal systems of first-order arithmetic (or some other sufficiently expressive theory) formulated in classical logic. (ibid, p. 13)

In any case, Field offers no independent evidence against the coherence of semantical notions like bivalent determinateness and exclusion negation. Neither does he offer any independent evidence for attributing some notions that are defined by the iteration of his determinacy operator to ordinary speakers. On the other hand, there is plenty of evidence for the fact we use exclusion negation, a general notion of truth-value gaps. If a theory disallows such intuitively appealing semantical notions, and we do often use such notions in natural languages, it seems that the theory is far too remote from natural languages.

As for the dialetheic approaches, in chapter 5, we saw that $LPTT$ and $STTT$ cannot properly interpret the claim that A is just true: if the claim that A is just true is formalized as $T\langle A\rangle \wedge \neg F\langle A\rangle$, then it cannot get the desired interpretation; since $T\langle A\rangle \wedge \neg F\langle A\rangle$ can still be a contradiction (i.e., $T\langle A\rangle \wedge \neg F\langle A\rangle$ can still be assigned the value $\frac{1}{2}$).

However, if we extend both $LPTT$ and $STTT$ with a Just True operator $\mathbb{J}$ defined as:

- $v_{\mathcal{M}^{+\mathbb{J}}}(\mathbb{J}A) = \begin{cases} 1, \text{ if } v_{\mathcal{M}^{+\mathbb{J}}}(A) = 1. \\ 0, \text{ if } v_{\mathcal{M}^{+\mathbb{J}}}(A) \neq 1. \end{cases}$

the augmented theories might not suffer from revenge paradoxes. In chapter 5, we carefully distinguished various ways to represent self-reference. The strong self-referential procedures make use of the identity $=$, or a denotation function to represent self-reference. The weak self-referential procedures require a self-referential sentence to be equivalent to a sentence that talks about the first one. The weak self-referential procedures can make use of the semantic equivalence in the theory; they can also make use of biconditionals. Accordingly, revenge paradoxes can be formulated by making use of different self-referential procedures.

While the augmented theories $LPTT^{\mathbb{J}}$ and $STTT^{\mathbb{J}}$ cannot resist the revenge via strong procedures, both theories do better in resisting the revenge via weak procedures. In particular, $LPTT^{\mathbb{J}}$ can resist the revenge via the material biconditional. Unfortunately, it cannot resist the revenge via the semantic equivalence. On the other hand, $STTT^{\mathbb{J}}$ can resist the revenge via the material biconditional and the revenge via the semantic equivalence.

Since revenge paradoxes are often formed by biconditionals and semantic equivalences, it seems that the strict-tolerant dialetheic approaches are doing quite well in dealing with revenge paradoxes. In any case, the strict-tolerant dialetheic approaches do a better job than the paracomplete gap approaches and the paraconsistent dialetheic approaches: among the three different approaches we considered, the strict-tolerant dialetheic approaches are the only approaches that survive from revenge paradoxes.

Some theorists such as Field confess that if a theory has the resources to express some revenge-breeding notions without being trivialized, such a theory has an advantage over those theories that opt for the expressive incompleteness horn:

> Note however that this argument cannot very well be advocated by the classical theorist, since the classical theorist has no such unified notion either. Nor can it very well be advocated by the proponent of any other solution to the paradoxes in which such a notion is unavailable. Indeed, I'm not sure that there are any demonstratively consistent theories (or even non-trivial dialetheic ones) that have such a notion available and hence are in a position to advocate this argument. I'm willing to concede (for the moment anyway) that it would be a point in favor of a solution to the paradoxes that it had a unified notion of defectiveness. (Field, 2007, p. 144)

Thus, it seems that the non-triviality results for $STTT^{\mathbb{J}}$ give a significant advantage for the strict-tolerant dialetheic approaches.

7.1.2 The Semantic Characterization Project

The Liar paradox shows that some sentences cannot be exhaustively and exclusively characterized as true or false. The standard strategy to deal with the Liar paradox is to posit a category other than truth and falsity. Kripke's theory K_3TT posits truth-value gaps; whereas dialetheic theories posit truth-value gluts. (At least, these categories are what such theories intend to posit, despite their expressive limitations.)

However, it turns out that introducing a new category does not help. For K_3-based theories, the sentence 'this sentence is gappy or false' cannot be exhaustively and exclusively characterized as either true, false or gappy. As a result, Field rejects exhaustive characterization. He suggests that there can be no semantical notions for which excluded middle can be assumed. For dialetheic theories, the sentence 'this sentence is just false' cannot be exhaustively and exclusively characterized as just true, just false, or glutty. It is glutty and just false. Priest suggests that we should reject exclusive characterization: if there is an argument that shows a sentence has more than one semantic status, we should simply accept the argument. This strategy is compatible with the spirit of dialetheism. Our strict-tolerant solution follows Priest's suggestion. But is rejecting the possibility of exhaustively and exclusively characterizing sentences an undesirable feature of a theory? Not necessarily, because our strict-tolerant solution can explain why sentences cannot be exclusively characterized.

Recall that the Strengethened Liar sentence 'this sentence is strictly false' can be represented as $\xi \dashv\models^{STTT^{\mathbb{J}}}_{\dashv\vDash} \mathbb{J}\neg T\langle\xi\rangle$ (or $\xi \equiv \mathbb{J}\neg T\langle\xi\rangle$). The Strengthened Liar sentence 'this sentence is strictly false' can be said to be both tolerantly true, and strictly false, because ξ always takes the value $\frac{1}{2}$ and $\mathbb{J}\neg T\langle\xi\rangle$ always takes the value 0. But since the Strengthened Liar sentence is tolerantly true, rather than strictly true, it is merely *tolerantly says* of itself that it is strictly false. No sentence can strictly say of itself that it is strictly false.

This sheds some light on why we cannot exclusively characterize sentences. The Liar and its revenge do not show that we do not have a coherent notion of just true (just false). We do have such notions, but we are still unable to exclusively (and exhaustively) characterize sentences. According to our strict-tolerant solution, it is because the self-referential relation is non-transitive: some sentences can tolerantly talk about themselves and tolerantly say of themselves that they are just false.

7.2 The Pragmatic Solution

The issue of expressive limitations is sometimes posed as a problem of communicating disagreement: it is argued that dialetheists cannot express their disagreement to their opponents. The argument is that if a dialetheist disagrees with A, he cannot express his disagreement by asserting $\neg A$ (or $\neg T\langle A\rangle$); because $\neg A$ (or $\neg T\langle A\rangle$) is compatible with A. The problem is known as the exclusion problem. Some might even think that dialetheists have no way to solve the exclusion problem. It is commonly thought that dialetheists cannot have an exclusion-expressing device, because they cannot model the notion of just true and the notion of just false.

Priest (2006) suggests that dialetheists can express their disagreement to their opponents through pragmatic implicatures. According to Priest, despite the fact that $\neg A$ (or $\neg T\langle A\rangle$) is compatible with A, asserting $\neg A$ (or $\neg T\langle A\rangle$) will implicate the fact that the speaker does not accept A. It is commonly accepted (e.g., Shapiro 2004, Priest 2006, Berto 2014) that this proposal does not work, because implicatures do not act upon embedded sentences (in particular, conditionals).

Making use of an exact truthmaker semantics, we defined a pragmatic interpretation rule that accounts for how a dialetheist can express his disagreement to his opponent through implicatures, and accounts for embedded implicatures for the whole language. In particular, the pragmatic interpretation rule accounts for meaning strengthening qua inconsistency minimization. It also accounts for meaning strengthening qua exhaustive interpretation.

As far as the exclusion problem is concerned, the pragmatic solution is a safe bet for dialetheists. Paraconsistent dialetheists cannot properly model the notion of just true and the notion of just false, because of the expressive limitations of their theories and revenge paradoxes. However, they can still communicate what they disagree about to non-dialetheists through implicatures.

The pragmatic solution is also compatible with the strict-tolerant solution. While strict-tolerant dialetheists can model the notion of just true and the notion of just false, they need not make use of them to communicate disagreement to their opponents. (Indeed, it is rare that one says A *is just true/just false* to indicate that he does not believe $\neg A$.) They can simply assert A is not the case (or A is not true/ A is false), even if what they have in mind is that A is just false. Their assertion implicates the fact that they do not accept A.

Bibliography

Alxatib, Sam, Peter Pagin, and Uli Sauerland (2013). "Acceptable Contradictions: Pragmatics or Semantics? A Reply to Cobreros, et al". In: *Journal of Philosophical Logic* 42.4, pp. 619–634.

Alxatib, Sam and Francis Jeffry Pelletier (2011). "The Psychology of Vagueness: Borderline Cases and Contradictions". In: *Mind and Language* 26.3, pp. 287–326.

Arenhart, Jonas Rafael Becker (2019). "On Material Exclusion and Absolute Contradiction". In: *Axiomathes*, pp. 1–13.

Atlas, Jay David (1989). *Philosophy Without Ambiguity: A Logico-Linguistic Essay*. Oxford University Press.

Bacon, Andrew (2013). "Non-classical Metatheory for Non-Alassical Logics". In: *Journal of Philosophical Logic* 42.2, pp. 335–355.

— (2015). "Can the Classical Logician Avoid the Revenge Paradoxes?" In: *Philosophical Review* 124.3, pp. 299–352. DOI: 10.1215/00318108-2895327.

Barrio, Eduardo, Federico Pailos, and Damian Szmuc (2018). "A Recovery Operator for Nontransitive Approaches". In: *Review of Symbolic Logic*, pp. 1–25.

Barrio, Eduardo, Lucas Rosenblatt, and Diego Tajer (2015). "The Logics of Strict-Tolerant Logic". In: *Journal of Philosophical Logic* 44.5, pp. 551–571.

Barwise, Jon and John Etchemendy (1987). *The Liar: An Essay on Truth and Circularity*. Oxford University Press USA.

Beall, Jc (2009). *Spandrels of Truth*. Oxford University Press.

— (2011). "Multiple-Conclusion LP and Default Classicality". In: *Review of Symbolic Logic* 4.2, pp. 326–336.

— (2013). "Shrieking Against Gluts: The Solution to the 'Just True' Problem". In: *Analysis* 73.3, pp. 438–445.

— (2015). "Free of Detachment: Logic, Rationality, and Gluts". In: *Noûs* 49.2, pp. 410–423.

Beall, Jc and Michael Glanzberg (2008). "Where the Paths Meet: Remarks on Truth and Paradox &Ast". In: *Midwest Studies in Philosophy* 32.1, pp. 169–198. DOI: 10.1111/j.1475-4975.2008.00171.x.

Beall, Jc, Michael Glanzberg, and David Ripley (2017). "Liar Paradox". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward Zalta. Fall 2017. Metaphysics Research Lab, Stanford University. URL = < https://plato.stanford.edu/archives/fall2017/entries/liar-paradox/>.

— (2018). *Formal Theories of Truth*. Oxford University Press.

Belnap, Nuel (1962). "Tonk, Plonk and Plink". In: *Analysis* 22.6, pp. 130–134.

Berto, Francesco (2014). "Absolute Contradiction, Dialetheism, and Revenge". In: *Review of Symbolic Logic* 7.2, pp. 193–207.

Brady, Ross (1989). "The Non-Triviality of Dialectical Set Theory". In: *Paraconsistent Logic: Essays on the Inconsistent*. Ed. by Graham Priest, Richard Routley, and Jean Norman. Philosophia Verlag, pp. 437–470.

Cattermole, Thomas (2016). "Paraconsistent Logics and Identity – a Pragmatic Approach". Master's Thesis. University of Amsterdam.

Chartier, Cian (2011). "Tarski's Threat to the T-Schema". Master's Thesis. University of Amsterdam.

Chierchia, Gennaro, Danny Fox, and Benjamin Spector (2012). "The grammatical view of scalar implicatures and the relationship between syntax and semantics". In: *Semantics: An International Handbook of Natural Language Meaning* 1.

Cobreros, Pablo, Paul Égré, David Ripley, and Robert van Rooij (2012). "Tolerant, Classical, Strict". In: *Journal of Philosophical Logic* 41.2, pp. 347–385.

— (2013). "Reaching Transparent Truth". In: *Mind* 122.488, pp. 841–866.

— (2014). "Priest's Motorbike and Tolerant Identity". In: *Recent Trends in Philosophical Logic*. Springer, pp. 75–83.

— (2015a). "Pragmatic Interpretations of Vague Expressions: Strongest Meaning and Nonmonotonic Consequence". In: *Journal of Philosophical Logic* 44.4, pp. 375–393.

— (2015b). "Vagueness, Truth and Permissive Consequence". In: *Unifying the Philosophy of Truth*. Ed. by Kentaro Fujimoto, José Martínez Fernández, Henri Galinon, and Theodora Achourioti. Springer Verlag, pp. 409–430.

— (2017a). "A Pragmatic Analysis of What Is Communicated (Though Not Said)". manuscript.

— (2017b). "Tolerant Reasoning: Nontransitive or Nonmonotonic?" In: *Synthese*, pp. 1–25.

Cook, Roy (2009). "What is a Truth Value And How Many Are There?" In: *Studia Logica* 92.2, pp. 183–201. DOI: 10.1007/s11225-009-9194-1.

Égré, Paul, David Ripley, and Steven Verheyen (2015). "The Sorites Paradox in Psychology". In: *The sorites paradox*. Ed. by Elia Zardini and Sergi Oms. Springer Verlag, pp. 1–28.

Estrada-González, Luis (2012). "Models of Possibilism and Trivialism". In: *Logic and Logical Philosophy* 21.2, pp. 175–205. DOI: 10.12775/LLP.2012.010.

Field, Hartry (2003). "A Revenge-Immune Solution to the Semantic Paradoxes". In: *Journal of Philosophical Logic* 32.2, pp. 139–177.

— (2007). "Solving the Paradoxes, Escaping Revenge". In: *Revenge of the Liar: New Essays on the Paradox*. Ed. by Jc Beall. Oxford University Press.

— (2008). *Saving Truth From Paradox*. Oxford University Press.

Fine, Kit (2014). "Truth-Maker Semantics for Intuitionistic Logic". In: *Journal of Philosophical Logic* 43.2-3, pp. 549–577.

Fine, Kit and Mark Jago (2018). "Logic for Exact Entailment". In: *Review of Symbolic Logic*.

French, Rohan (2016). "Structural Reflexivity and the Paradoxes of Self-Reference". In: *Ergo: An Open Access Journal of Philosophy* 3. DOI: 10.3998/ergo.12405314.0003.005.

Gamut, L. T. F. (1990). *Logic, Language, and Meaning, Volume 1: Introduction to Logic*. University of Chicago Press.

Geis, Michael and Arnold Zwicky (1971). "On Invited Inferences". In: *Linguistic Inquiry* 2.4, pp. 561–566.

Glanzberg, Michael (2004). "A Contextual-Hierarchical Approach to Truth and the Liar Paradox". In: *Journal of Philosophical Logic* 33.1, pp. 27–88. DOI: 10.1023/B:LOGI.0000019227.09236.f5.
— (2015). "Complexity and Hierarchy in Truth Predicates". In: *Unifying the Philosophy of Truth*. Ed. by Kentaro Fujimoto, José Martínez Fernández, Henri Galinon, and Theodora Achourioti. Springer Verlag.
Gödel, Kurt (1992). *On formally undecidable propositions of Principia Mathematica and related systems*. Courier Corporation.
Goodship, Laura (1996). "On Dialethism". In: *Australasian Journal of Philosophy* 74.1, pp. 153–161.
Grice, Paul (1975). "Logic and Conversation". In: *The Semantics-Pragmatics Boundary in Philosophy*. Ed. by Maite Ezcurdia and Robert Stainton. Broadview Press, p. 47.
Groenendijk, Jeroen and Martin Stokhof (1984). "Studies on the Semantics of Questions and the Pragmatics of answers." PhD thesis. University of Amsterdam.
Gupta, Anil and Nuel Belnap (1993). *The Revision Theory of Truth*. MIT Press.
Halbach, Volker (1997). "Tarskian and Kripkean Truth". In: *Journal of Philosophical Logic* 26.1, pp. 69–80.
— (2011). *Axiomatic Theories of Truth*. Cambridge University Press.
Heck, Richard (2005). "Truth and Disquotation". In: *Synthese* 142.3, pp. 317–352. DOI: 10.1007/s11229-005-3719-6.
— (2007). "Self-Reference and the Languages of Arithmetic". In: *Philosophia Mathematica* 15.1, pp. 1–29.
— (2012a). "A Liar Paradox". In: *Thought: A Journal of Philosophy* 1.1, pp. 36–40.
— (2012b). "More on 'A Liar Paradox'". In: *Thought: A Journal of Philosophy* 1.4, pp. 270–280.
Horn, Laurence (1989). *A Natural History of Negation*. University of Chicago Press.
— (2000). "From if to iff: Conditional Perfection as Pragmatic Strengthening". In: *Journal of Pragmatics* 32.3, pp. 289–326.
Incurvati, Luca and Julian Schlöder (2017). "Weak Rejection". In: *Australasian Journal of Philosophy* 95.4, pp. 741–760.
Incurvati, Luca and Peter Smith (2010). "Rejection and Valuations". In: *Analysis* 70.1, pp. 3–10.
Jago, Mark (2018). *What Truth Is*. Oxford: Oxford University Press.
Jenny, Matthias (2016). "Classicality Lost: K3 and LP After the Fall". In: *Thought: A Journal of Philosophy* 5.4.
Kabay, Paul (2008). "A Defense of Trivialism". PhD thesis. The University of Melbourne.
Ketland, Jeffrey (2003). "Can a Many-Valued Language Functionally Represent its Own Semantics?" In: *Analysis* 63.4, pp. 292–297. DOI: 10.1093/analys/63.4.292.
Kirkham, Richard (1992). *Theories of Truth: A Critical Introduction*. Cambridge, MA: MIT Press.
Kremer, Michael (1988). "Kripke and the Logic of Truth". In: *Journal of Philosophical Logic* 17.3, pp. 225–278.
Kripke, Saul (1975). "Outline of a Theory of Truth". In: *Journal of Philosophy* 72.19, pp. 690–716.

Kripke, Saul A. (2019). "Ungroundedness in Tarskian Languages". In: *Journal of Philosophical Logic* 48.3, pp. 603–609. DOI: 10.1007/s10992-018-9486-x.
Leitgeb, Hannes (2007). "On the Metatheory of Field's 'Solving the Paradoxes, Escaping Revenge'". In: *Revenge of the Liar: New Essays on the Paradox*. Ed. by Jc Beall. Oxford University Press.
Littman, Greg and Keith Simmons (2004). "A Critique of Dialetheism". In: *The Law of Non-Contradiction*. Ed. by Graham Priest, Jc Beall, and Bradley Armour-Garb. Oxford University Press, pp. 1–226.
Martin, Robert and Peter Woodruff (1975). "On Representing 'True-in-L' in L". In: *Philosophia* 5.3, pp. 213–217.
McGee, Vann (1991). *Truth, Vagueness and Paradox*. Hackett.
Mortensen, Chris (2005). "It Isn't So, But Could It Be?" In: *Logique Et Analyse* 48.
Parsons, Terence (1984). "Assertion, Denial, and the Liar Paradox". In: *Journal of Philosophical Logic* 13.2, pp. 137–152.
— (1990). "True Contradictions". In: *Canadian Journal of Philosophy* 20.3, pp. 335–353.
Plebani, Matteo (2015). "Could Everything Be True? Probably Not". In: *Philosophia* 43.2, pp. 499–504. DOI: 10.1007/s11406-015-9584-8.
Priest, Graham (1979). "The Logic of Paradox". In: *Journal of Philosophical Logic* 8.1, pp. 219–241.
— (1984a). "Hyper-Contradictions". In: *Logique Et Analyse* 27.7, p. 237.
— (1984b). "Logic of Paradox Revisited". In: *Journal of Philosophical Logic* 13.2, pp. 153–179.
— (1984c). "Semantic Closure". In: *Studia Logica* 43.1-2, pp. 117–129. DOI: 10.1007/BF00935745.
— (1990). "Boolean Negation and All That". In: *Journal of Philosophical Logic* 19.2, pp. 201–215. DOI: 10.1007/BF00263541.
— (1991a). " Minimally Inconsistent LP". In: *Studia Logica* 50.2, pp. 321–331.
— (1991b). "Intensional Paradoxes". In: *Notre Dame Journal of Formal Logic* 32.2, pp. 193–211. DOI: 10.1305/ndjfl/1093635745.
— (2000). "Could Everything Be True?" In: *Australasian Journal of Philosophy* 78.2, pp. 189–195. DOI: 10.1080/00048400012349471.
— (2005). *Doubt Truth to be a Liar*. Oxford University Press.
— (2006). *In Contradiction: A Study of the Transconsistent*. Oxford University Press.
— (2007). "Revenge, Field, and ZF". In: *Revenge of the Liar: New Essays on the Paradox*. Ed. by Jc Beall. Oxford University Press, p. 225.
— (2008a). *An Introduction to Non-Classical Logic: From If to Is*. Cambridge University Press.
— (2008b). "Spiking the Field Artillery". In: *Deflationism and Paradox*. Ed. by Jc Beall and Bradley Armour-Garb. Oxford University Press.
— (2010a). "Hopes Fade for Saving Truth". In: *Philosophy* 85.1, pp. 109–140. DOI: 10.1017/S0031819109990489.
— (2010b). "Non-Transitive Identity". In: *Cuts and Clouds: Vagueness, its Nature, and its Logic*. Ed. by Richard Dietz and Sebastiano Moruzzi. Oxford University Press, pp. 406–416.
— (2014). *One: Being an Investigation Into the Unity of Reality and of its Parts, including the Singular Object which is Nothingness*. Oxford University Press.

Priest, Graham, Francesco Berto, and Zach Weber (2018). "Dialetheism". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward Zalta. Fall 2018. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2018/entries/dialetheism/>.

Priest, Graham and Richard Sylvan (1992). "Simplified Semantics for Basic Relevant Logics". In: *Journal of Philosophical Logic* 21.2, pp. 217–232.

Raatikainen, Panu (2018). "Gödel's Incompleteness Theorems". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward Zalta. Fall 2018. Metaphysics Research Lab, Stanford University. URL = < https://plato.stanford.edu/archives/fall2018/entries/goedel-incompleteness/>.

Rayo, Agustín and Philip Welch (2007). "Field on Revenge". In: *Revenge of the Liar: New Essays on the Paradox*. Ed. by Jc Beall. Oxford University Press.

Recanati, François (2003). "Embedded Implicatures". In: *Philosophical Perspectives* 17.1, pp. 299–332.

Ripley, David (2011). "Contradictions at the Borders". In: *Vagueness in Communication*. Ed. by Rick Nouwen, Robert van Rooij, Uli Sauerland, and Hans-Christian Schmitz. Springer, pp. 169–188.

— (2012). "Conservatively Extending Classical Logic with Transparent Truth". In: *Review of Symbolic Logic* 5.2, pp. 354–378.

— (2015). "Embedding Denial". In: *Foundations of Logical Consequence*. Ed. by Colin Caret and Ole Hjortland. Oxford University Press, pp. 289–309.

Rossberg, Marcus (2013). "Too Good to Be 'Just True'". In: *Thought: A Journal of Philosophy* 2.1, pp. 1–8. DOI: 10.1002/tht3.50.

Rossi, Lorenzo (2018). "Model-Theoretic Semantics and Revenge Paradoxes". In: *Philosophical Studies*, pp. 1–20.

Scambler, Chris (2018). "Ineffability and Revenge". In: *Review of Symbolic Logic*, pp. 1–14. DOI: 10.1017/S1755020318000473.

Scharp, Kevin (2013). *Replacing Truth*. Oxford University Press UK.

Schlenker, Philippe (2010). "Super Liars". In: *Review of Symbolic Logic* 3.3, pp. 374–414. DOI: 10.1017/s1755020310000067.

Schulz, Katrin and Robert van Rooij (2006). "Pragmatic Meaning and Non-Monotonic Reasoning: The Case of Exhaustive Interpretation". In: *Linguistics and Philosophy* 29.2, pp. 205–250.

Shapiro, Lionel (2011). "Expressibility and the Liar's Revenge". In: *Australasian Journal of Philosophy* 89.2, pp. 297–314. DOI: 10.1080/00048401003695156.

Shapiro, Stewart (2004). "Simple Truth, Contradiction, and Consistency". In: *The Law of Non-Contradiction*. Ed. by Graham Priest, Jc Beall, and Bradley Armour-Garb. Oxford University Press.

Shaw, James (2013). "Truth, Paradox, and Ineffable Propositions". In: *Philosophy and Phenomenological Research* 86.1, pp. 64–104. DOI: 10.1111/j.1933-1592.2011.00530.x.

Shaw, James R. (2014). "What is a Truth-Value Gap?" In: *Linguistics and Philosophy* 37.6, pp. 503–534. DOI: 10.1007/s10988-014-9160-x.

Simmons, Keith (1993). *Universality and the Liar: An Essay on Truth and the Diagonal Argument*. Cambridge University Press.

— (2018). *Semantic Singularities: Paradoxes of Reference, Predication, and Truth*. Oxford University Press.

Smiley, Timothy (1996). "Rejection". In: *Analysis* 56.1, pp. 1–9.

Smith, Peter (2012). *An Introduction to Gödel's Theorems*. Cambridge University Press.
Steinberger, Florian (2019). "Three Ways in Which Logic Might Be Normative". In: *Journal of Philosophy* 116.1, pp. 5–31.
Strawson, Peter (1950). "On Referring". In: *Mind* 59.235, pp. 320–344.
Tarski, Alfred (1944). "The Semantic Conception of Truth and the Foundations of Semantics". In: *Philosophy and Phenomenological Research* 4.3, pp. 341–376.
— (1956). "The Concept of Truth in Formalized Languages". In: *Logic, Semantics, Metamathematics*. Ed. by Alfred Tarski. Oxford University Press, pp. 152–278.
Teijeiro, Paula (2012). "Circularity is Still Scary". In: *Análisis Filosófico* 32.1, pp. 31–35.
Tourville, Nicholas and Roy Cook (2016). "Embracing the Technicalities: Expressive Completeness and Revenge". In: *Review of Symbolic Logic* 9.2, pp. 325–358. DOI: 10.1017/s175502031600006x.
van Fraassen, Bas (1966). "Singular Terms, Truth-Value Gaps, and Free Logic". In: *Journal of Philosophy* 63.17, pp. 481–495.
— (1968). "Presupposition, Implication, and Self-Reference". In: *Journal of Philosophy* 65.5, pp. 136–152.
— (1969). "Facts and Tautological Entailments". In: *Journal of Philosophy* 66.15, pp. 477–487.
van Rooij, Robert (2017). "A fine-grained global analysis of implicatures". In: *Linguistic and Psycholinguistic Approaches on Implicatures and Presuppositions*. Springer, pp. 73–110.
van Rooij, Robert and Katrin Schulz (2004). "Exhaustive Interpretation of Complex Sentences". In: *Journal of Logic, Language and Information* 13.4, pp. 491–519.
Wang, Wenfang (2013). "Filtering Theories of Truth: Compositionality as a Criterion". In: *Frontiers of Philosophy in China*, pp. 156–170.
Welch, Philip (2008). "Ultimate Truth Vis-À-Vis Stable Truth". In: *Review of Symbolic Logic* 1.1, pp. 126–142.
— (2011). "Truth, Logical Validity and Determinateness: A Commentary on Field's Saving Truth From Paradox". In: *Review of Symbolic Logic* 4.3, pp. 348–359.
— (2014). "Some Observations on Truth Hierarchies". In: *Review of Symbolic Logic* 7.1, pp. 1–30.
Whittle, Bruno (2017). "Truth, Hierarchy and Incoherence". In: *Reflections on the Liar*. Ed. by Bradley Armour-Garb. Oxford University Press.
Yablo, Stephen (2004). "New Grounds for Naive Truth Theory". In: *Liars and Heaps: New Essays on Paradox*. Ed. by Jc Beall. Clarendon Press, pp. 312–330.
Young, Gareth (2015a). "Revenge: Dialetheism and its Expressive Limitations". PhD thesis. University of Glasgow.
— (2015b). "Shrieking, Just False and Exclusion". In: *Thought: A Journal of Philosophy* 4.4, pp. 269–276.