MASTER'S THESIS

Can silence be a proper response to the liar paradox?

Li, Dilin

Date of Award:
2020

HONG KONG BAPTIST UNIVERSITY

Master of Philosophy

THESIS ACCEPTANCE

DATE: November 18, 2020

STUDENT'S NAME: LI Dilin

THESIS TITLE: Can Silence be a Proper Response to the Liar Paradox?

This is to certify that the above student's thesis has been examined by the following panel members and has received full approval for acceptance in partial fulfilment of the requirements for the degree of Master of Philosophy.

Chairman:	Prof Hjort Mette Dean, Faculty of Arts, HKBU
Internal Members:	Dr Lee Siu Fan Associate Professor, Department of Religion and Philosophy, HKBU (Designated by the Head of Department of Religion and Philosophy)
	Prof Palmquist Stephen R Professor, Department of Religion and Philosophy, HKBU
External Examiner:	Prof Priest Graham Distinguished Professor of Philosophy The Graduate Center The City University of New York

Issued by Graduate School, HKBU

Can Silence be a Proper Response to the Liar Paradox?

LI Dilin

A thesis submitted in partial fulfilment of the requirements
for the degree of
Master of Philosophy

Principal Supervisor:
Prof. PALMQUIST Stephen R (Hong Kong Baptist University)

11.2020

DECLARATION

I hereby declare that this thesis represents my own work which has been done after registration for the degree of MPhil at Hong Kong Baptist University, and has not been previously included in a thesis or dissertation submitted to this or any other institution for a degree, diploma or other qualifications.

I have read the University's current research ethics guidelines, and accept responsibility for the conduct of the procedures in accordance with the University's Research Ethics Committee (REC). I have attempted to identify all the risks related to this research that may arise in conducting this research, obtained the relevant ethical and/or safety approval (where applicable), and acknowledged my obligations and the rights of the participants.

Signature: 李幸霖

Date: 2020.11.18

Abstract

Many attempts at solving the liar paradox involve either rejecting some principles in classical logic so as to block the argument that leads to the contradiction or modifying the notion of truth so that the liar sentence can be classified as true in one aspect while false in another. However, the prominent approaches based the above strategies may suffer from the revenge problem. That is, while they solve the pristine liar paradox, the introduction of the solution triggers another one with the same structure. In this dissertation, three prominent approaches to the liar paradox are first introduced and examined. In particular, they are, first, the Tarskian hierarchical approach, whose main idea can be roughly characterized as that a natural language is a hierarchy of a series of languages and the liar sentence is true at one level of the hierarchy and is false at another; second, Saul Kripke's paracomplete approach, whose main idea can be roughly characterized as that the liar sentence is ungrounded and has no classical truth value at all; finally, Gupta and Belnep's revision theory of truth, the main idea of which is that truth is a circular concept and that the truth predicate is circularly defined. With a new semantics and logic for circular concept and definition, one can classify the liar sentence as not categorical. Based on two general patterns that give rise to the revenge paradox by Graham Priest, it is shown that none of the above approaches can escape the revenge paradox, at least, not satisfactorily. After the examination of three prominent approaches, I provide an initial characterization of a kind of approach which I call the silence approach. The main idea of the silence approach is that, perhaps what the liar paradox teaches us is that the semantic status of the liar sentence is eventually not classifiable, in the sense that the accepted or correct semantic theory for natural language simply does not apply to the liar sentence. There are two theoretical possibilities that can evoke the failure of classification. Either there is just no semantic category that fits the liar sentence or the necessary principles for the classification do not apply to the sentence. In either case, the silence approach suggests that although the liar sentence could have a semantic status according to the accepted or correct semantic theory, but given that we cannot classify it, we cannot know it. In this dissertation, I do

not provide a detailed and well-developed theory of the silence approach. Instead, after the initial characterization of this approach, I go on to introduce and examine two current theories on the liar paradox which I think satisfy at least part of my characterization of the silence approach. The first theory is the semantic epistemicism by Paul Horwich. The second one is what I call exceptional theory, which is given by Thomas Hofweber. The result of the examination is that, both theories can indeed be interpreted as a silence approach. However, although they can block both the pristine liar paradox and the revenge paradox, they suffer severely from the problem of being ad hoc. The current conclusion of this dissertation about the silence approach thus is that, it is possible to construct a silence approach which can block the pristine liar paradox and the revenge, but it is hard to find a rationale for the solution. That is, it is hard to answer the question as to why the liar sentence is not classifiable. Finally, as an overlook to the future development of the silence approach, I suggest that even if we can solve the problem of ad hocness, there remains a question as to whether the incompleteness of classification is a symptom revealing that the accepted semantic theory is defective, or it is a symptom showing that there is just no possible semantic theory that can eventually do the job. Without answering this question, the silence approach still lacks a plausible theoretical ground.

Acknowledgements

The main idea of this dissertation (especially chapter 4) emerged in the first semester of my study in HKBU. And after that I have been discussing the idea of the silence approach with my principal supervisor, Stephen R. Palmquist. I want to thank professor Palmquist for his valuable comments and encouragement, without which I may have withdrawn from the university two years ago. I also want to thanks Ling Chow in the Graduate School and Ling Yam in the Religion and Philosophy department, for their constant assistances. Finally, I thank Qiang Siwei for her helping me buy the two-ring folders as well as print and submit my dissertation.

The initial idea of this dissertation was first presented as a paper in the Graduate Seminar in the Religion and Philosophy department in 2017. An extension of that paper was presented in the conference *International Postgraduate Roundtable and Research Forum cum Summer School 2017*, held by The Education University of Hong Kong, and won the prize "Outstanding Paper Award" in the conference. A shorter and revised version was presented in the conference *Chinese National Conference on Modern Logic 2017*, held by the Modern Logic chapter of The Chinese Association of Logic and Zhe Jiang University in Zhejiang, China.

Table of Contents

Chapter 1 Introduction 1

Section 1 Liar paradox and revenge 1

Section 2 Literature review 6

Section 3 Silence approach? 17

Section 4 Aim and structure 18

Chapter 2 Methodology 20

Section 1 Problem identification 20

Section 2 Solution 24

Section 3 Language model 28

Section 4 The scope 35

Section 5 Summary 38

Chapter 3 Some Prominent Approaches 40

Section 1 Tarskian hierarchical approach 40

Section 2 Kripke's paracomplete approach 59

Section 3 Gupta and Belnap's revision theory 74

Section 4 Summary 105

Chapter 4 Silence Approach 107

Section 1 Silence position and its three main worries 107

Section 2 What is silence? 110

Section 3 Horwich: Semantic Epistemicism 115

Section 4 Hofweber: Exception theory 139

Section 5 Summary and future development 148

References 152

CURRICULUM VITAE 158

List of Figures

Figure3-1 89

Figure 3-2 89

Figure 3-3 90

Figure 3-4 90

List of Symbols

The following list includes signs which have fixed usage throughout this dissertation:

$\vdash$: derivation

$\wedge$: conjunection

$\vee$: disjunction

$\sim$: negation

$\rightarrow$: material conditional

iff, $\leftrightarrow$: material biconditional

$\forall$: universal quantifier

$\exists$: existential quatifier

Σ: a set of sentences

$\in$: membership

$\subseteq$: subset

$\leq$: smaller or equal to

$\geq$: greater or equal to

$=$: identity

$\varnothing$: empty set

$<L, M, \sigma>$: L, syntax information;

M, model

σ , valuation schema

$<D, I>$: D, universe of domain

I, interpretation function

D^n: the set of n-tuples in the domain D

$D^n \rightarrow D$: a function from D^n to D

$D^n \rightarrow \{1,0\}$ a function from D^n to $\{1,0\}$

V_M: valuation function, relative to a model M.

κ: Strong Kleene semantic schema

θ, τ, ν: variable representing formula

Can Silence be a proper response to the liar paradox?

Chapter 1: Introduction

1. Liar paradox and revenge

The origin of the liar paradox perhaps can date back to ancient Greece, when Eubildes asked whether a man is lying when the man himself says that he is lying.[1] Suppose that the man is telling the truth, then what he says must be the case, so he is lying and so he is not telling the truth; suppose on the other hand that he is lying, then since this is exactly what he is saying, he is telling the truth. In its modern form, and also a much simpler form, the liar paradox is often formulated with the concept of truth and a name of a self-referential sentence:

S: *S* is false

Suppose that S is true, we derive that S is false. Conversely, if we suppose that S is false, then we can derive that it is true. The sentence thus forces us to conclude that it is both true and false, which is a contradiction. It is a paradox because we can derive a seemingly unacceptable conclusion from seemingly plausible acceptable premises via seemingly valid rules of inference. A solution to the paradox would therefore need to point out exactly which premise is not true or which step of inference is invalid if it does not accept the consequent contradiction. If on the other hand, the solution accepts the contradictory consequence, then it must tell us how to live with it.

With a careful observation we can pick up several prominent assumptions that lead us to this paradoxical consequence:

(1) Classical logic holds.

Alternatively, classical semantic principles hold.[2]

---

[1] This story about the origin of the liar paradox is widely known. I saw this one in Gupta and Belnap (1993), p.6.

[2] By "classical semantic principles", I mean the semantic principles for classical logic. For example, the principle of bivalence: every sentence is either true or not true. There can be two versions of the liar paradox. First, it is proof-theoretical (using classical rules of inference and the T-schema in its rule form) and the other is semantic (using classical semantic principles and

(2) Two truth-related rules of inference:[3]

T-intro: A ⊢ "A" is true

T-elim: "A" is true ⊢A

Correspondingly, we have a proof-theoretical form of the T-schema:[4]

"A" is true ⇔A

or, we may use the semantic principles (if the semantics for the language is classical two-valued) rather than the above rules:

T-out: "A" is true→A

T-in: A→ "A" is true

Correspondingly, we have a conditional form of the T-schema:

"A" is true ↔ A

(3) Some sort of self-referential apparatus is allowed.[5]

Premise (1) is about logic, premise (2) is about the concept of truth and premise (3) is about the syntactic resources of the language, which allows us to construct the liar sentence. If all these three are based on our normal conceptions of logic, truth and language, then the liar paradox shows that there may be something wrong with our ordinary conceptions. So the liar paradox may not be merely a linguistic riddle, but perhaps is some important symptom revealing some inadequacy of any one of these three aspects or our conceptions of them.

---

the T-schema in its *conditional* form).

[3] A represents a sentence and "A" is its quotation mark name. "⊢" is the sign for "derivability", A ⊢ B means that B can be derived from A. "⇔" represents "provable equivalence". "A⇔B" means that A and B are provably equivalent. Here we omit the reference to some particular deduction system. "→" represents conditional. "A" is true→ A means that if "A" is true, then A; A→ "A" is true means that if A then "A" is true. "↔" represents material equivalence. "A↔B" means A and B are materially equivalent. These two groups of truth-related principles (except the two forms of T-schema) can be found in Scharp (2013), p.23 and Beall (2007), p.1. Following Beall (2007, 1-2), I will sometimes refer to T-in as Capture (in its conditional form) and to T-out as Release (in its conditional form) and to T-intro as Capture (in its rule form) and Release (in its rule form).

[4] It seems that calling the following displayed schematic sentence, T-schema, is a tradition since Tarski (1933). See Beall, J.C. Glanzberg, M. and Ripley, D. (2018), p.17, for introduction.

[5] Yablo's paradox (Yablo, 1993) may show that self-reference is not a necessary condition for liar-like paradox. But I do not want to expand the scope of this dissertation. Here I only consider one simple kind of liar paradox, which takes the form like that of sentence S. For more on this see chapter 2.

A quick response is to reject premise (1). Perhaps the correct logic and semantics is not what the classical theory suggests. For example, the liar sentence *S* is neither true nor false.

But this alone does not help us a lot. For it is quite easy to reconstruct a stronger version of the liar sentence, using the semantic concept which we use to classify the pristine liar sentence. Consider the following one:

R: *R* is *either neither true nor false or false.*[6]

Suppose that it is neither true nor false or that it is false, then in either case, by premise (2), it is true. And since being neither true nor false or being false implies being *not* true, we have a contradiction. Suppose that it is true, then by premise (2), it is *not* true. Either way, we have a contradiction. This is the inconsistency problem. The problem can also be viewed from another way. Suppose that we characterize *R* as being neither true nor false, then it seems that we should endorse the claim that *R* is either neither true nor false or false, which is exactly *R* itself. So we are endorsing a sentence that according to the theory, is neither true nor false. This is the problem of self-refutation. The former problem is about inconsistency of truth value, the latter is about how to express the theory itself.[7]

This phenomenon is not uncommon. Quite often that when a theory tries to solve the pristine liar paradox, a new semantic concept (perhaps, a sequence of semantic concepts) will be introduced into the language. And with the help of these newly introduced semantic concepts, one can construct another sentence which is *structurally similar* to the pristine liar sentence, but which cannot be handled by the theory itself. This phenomenon is generally called *revenge paradox*, and the sentence reconstructed via the newly introduced semantic concepts (either use the

---

[6] Strengthened liar sentence of this kind is frequently discussed among the literatures to the liar paradox, as an initial introduction to the issue. For example, one can find a relevant introduction in Beall (2007), p.3.

[7] These two kinds of problems were discovered quite a long time ago. A clear distinction of these two problems can be found in Scharp (2013), p.100. For relevant discussion over whether Gap theorist can express their own theory, see Parson (1990) and Priest (1995).

new concept directly, or some other concept that is defined in terms of the newly introduced concept) is called *revenge sentence*.

Many modern theories on the liar paradox make use of some highly sophisticated method.[8] Normally, they will construct an object-language (via some set-theoretical notions) which contains no semantic concepts and then examine whether or not, under their theories on either truth, semantics or logic, an additional one-place predicate can be interpreted as the truth predicate in such a language without triggering inconsistency (We will see more details on this method in the methodology chapter). There is one important criterion in examining the adequacy of these language models. That is, the object-language so constructed must be *similar enough* to natural languages at least in some relevant aspects in order to show that how natural languages, while *having such and so feature*, can still contain a truth predicate with desired features. The revenge argument although admits that the object-language can contain its own truth predicate, it shows that it cannot contain at least some notions (on pain of inconsistency or some other paradoxicality) which are seemingly expressible in natural languages. In the above example, the revenge paradox caused by the sentence *R* can be used to show that the semantic concept "*neither true nor false*" is not expressible in the object-language, for otherwise, we have a contradiction. Thus one form of the revenge argument is that, the object-language fails to be similar enough to natural languages for it is *essentially expressively weaker* than natural languages. The upshot of a revenge argument against a theory is thus to establish certain expressibility gap between the language model and natural languages. Note that, given that the revenge sentence is structurally extremely similar to the pristine liar sentence, it is normally not considered as a different paradox to the liar paradox. Rather, if a solution to the liar paradox cannot avoid the revenge paradox, it is generally considered as unsatisfactory—it does not really solve the liar paradox but just transforms it into a

[8] The following general remarks on the sophisticated method employed in modern theories on the liar paradox is mainly from Beall (2007), appendix to chapter 1.

different form.

However, it is not the case that as long as we can prove that some semantic concept, either it is used by the theory or not, is not expressible in the object-language, the theory will necessarily be damaged. Beall (2007, 11) argues (and I agree with him) that while it is easy to establish the inexpressibility of certain concept in a language model, it is *too easy* (in a negative sense) to conclude that the language model is inadequate, for the relevance of the inexpressibility of the concept is sometimes not obvious. Beall's idea is, if my understanding is correct, that the relevant semantic concept is often a model-dependent[9] classical concept. Its inexpressibility in the object-language may therefore be due to the concept itself rather than the object-language. Take the above revenge sentence *R* as an example. We derive the conclusion that *R* is both true and not true *only* when we assume that the concept "neither true nor false" has a *classical semantics* and follows *classical logic*. In particular, we assume that for any sentence in the object-language, either it is *neither true nor false* or it is not *neither true nor false*. But perhaps the classical "neither true nor false" concept is itself *paracomplete*, that is, there is some sentence that is neither *neither true nor false* nor *not neither true nor false*—the classical version of this semantic concept is so defective.[10] On the other hand, there may be some non-model-dependent concepts which are not directly used by the theory but are said to be expressible in natural language. To these concepts, Beall (2007, 13) argues that one still needs to establish the intelligibility of them or at the very least, one needs to establish that we *need* to recognize those concepts. But exactly what does it mean by "intelligibility"? What is it to call a concept intelligible? In what

[9] A model-dependent concept is a concept that is defined particularly in the theory in order to describe some theoretical feature of the sentence, which cannot be understood without the theory. See Beall (2007), chapter 1, section 1.4 for relevant discussion.

[10] It seems that in Beall's idea, the inexpressibility of such a concept may be *irrelevant*, and one may simply leave it without any further explanation. See Beall (2007), p10. I cannot make sense of this position. For me, I think that as long as the concept is expressible in natural language, one should response to it, if it will eventually cause inconsistency. Even if eventually the inconsistency caused by the concept is not about truth or semantics, still, it is a problem as to how to characterize the sentence, if it cannot be characterized in the normal way, or in the way the theory at issue suggests.

situation a concept is not intelligible? Beall does not provide a definition. However, it can be observed that some solutions to the liar paradox, when they realize that there is a revenge problem for them, they end up denying the legitimacy of some concepts that form the very revenge sentence in order to avoid it. The situation here is, just like Beall (2007, 13) says, delicate. For on the one hand, objectors may argue that since those concepts are intelligible, such and so theory is wrong for it fails to accommodate those concepts; but the theorists may reply that since their theories are right, those concepts are not intelligible at all! Beall (2007, 13-14) concludes that therefore the burden of proof falls on both sides. The objector needs to provide evidence for the intelligibility of relevant concepts while the theorists need to explain why despite the apparent intelligibility, the relevant concepts are actually not intelligible. We will see some examples later.

2. Literature review

In this section we will quickly review some well-known solutions to the liar paradox, without focusing on the technical details in those theories. And we will also discuss some general patterns that give rise to the revenge paradox. More detailed examination will be given in chapter 3.

As mentioned, there are at least three kinds of assumptions in the liar paradox—assumptions about logic, truth and language. A solution may therefore try to solve the problem by denying one or more than one of them.[11] The quick solution we discuss above denies classical logic and/or the corresponding semantic principles. In particular, it denies that the law of excluded middle/principle of bivalence is valid—it is not the case that either the liar sentence is true or that it is false. One question for this kind of theory is that although we may agree that the liar sentence *can* be neither true nor false, it is not at all obvious why it *is* neither true nor false.

[11] These options are only a sketchy classification of solutions to the liar paradox and they are not mutually exclusive at all. For a detailed and sophisticated classification, see Scharp (2013), chapter 1.

Martin (1967) rejects the liar sentence as being either true or false based on his theory on semantical correctness, a kind of category mistake. Only semantically correct sentences can be either true or false and the liar is not, so it is neither true nor false.[12] Priest (2006, 14) argues that the approach is ad hoc for it rules out the liar sentence as being semantically incorrect only by applying a special clause in the decision procedure for semantical correctness. But I think the clause may not be ad hoc, especially because it seems to be a general principle that is designed to assess most self-referential sentences including both pathological sentences and those that are healthy, rather than declaring directly that this or that sentence is not semantically correct.[13] But anyway, this is not the main issue, the trouble for the theory perhaps is more on the revenge problem. Another kind of rationale is to argue that the liar sentence is neither true nor false because it does not express a proposition/statement. A defense of this theory is from Goldstein (2009), who based on Strawson's distinction between *sentence* and *the use of sentence* (Strawson, 1950), argues that sentences are neither true nor false and only the statement made by the use of a sentence can be either true or false and since the liar sentence is a sentence, it is neither true nor false. Further, the liar sentence can never be used to make a statement, so it does not say something that is either true or false. One virtue of this position is that, since the liar-like sentences, including the revenge, are taken to be stating nothing, we can simply call them neither true nor false without getting further inconsistency.[14] For the revenge argument relies on the assumption that

[12] Strictly speaking, Martin's theory does not show that the liar is neither true nor false because it does not show that the liar is not semantically correct. What the theory establishes is that the sentence cannot be shown to be semantically correct. And so we have no reason to accept that it is either true or false.

[13] Note that what Martin is doing is to use several criteria to characterize what he think is semantically incorrect, a phenomenon that is independent from the liar paradox. Whether his solution to the liar paradox is ad hoc will then depend on whether the criteria he provides are ad hoc. The ideal set of criteria is the one that would apply to all sorts of sentences in a language, but which can rule out problematic sentences and preserve the good ones. But what sort of criteria would be ad hoc? I do not have a clear answer now. But I think a kind of criteria, like "this or that sentence is problematic", which applies only to a particular sentence, would be definitely unacceptable as being ad hoc. And Martin's clause for assessing self-referential sentences is not of this kind. So I think there may still have room for debate over whether his criteria are ad hoc or not. But in any case, I do not insist on this point.

[14] But one may still have question as to whether it will evoke the problem of self-refutation

"being neither true nor false is what the revenge sentence says", but now the revenge sentence does not state anything, so no revenge follows. There have been many objections on the idea that the liar sentence does not express a proposition. Most of the objections[15] are based on the idea that the liar sentences seem to satisfy some sufficient conditions for a sentence to express a proposition/statement. For example, one can *believe* in a paradoxical sentence; one can also *modalize* a paradoxical sentence; one can *do deductive reasoning* with a paradoxical sentence. It seems that if the liar sentence is meaningful but does not state a proposition, like an imperative sentence, the above behaviors are not possible. For some philosophers, these are enough to show that the liar sentence does express a proposition. And because of this, Goldstein's theory for deciding whether a sentence succeeds in making a statement will be rejected by them as some ad hoc move.[16] Eventually which party will win the debate perhaps depends on a more general theory on truth bearers and I will not enter that topic in this dissertation any further.

Another sort of approach that denies classical logic can be found in Priest (1979),[17] which denies law of non-contradiction and argues that the liar paradox shows that there is true contradiction and the liar sentence is simply both true and false. The

---

(see chapter 1, section 1). Consider the revenge sentence:
R: R is neither true not false or false.
Now Goldstein's position is that, R does not make a statement, so R is neither true nor false. And if we can assert that R is neither true nor false, we can also assert that R is neither true nor false or false. But a puzzle occurs when we realize that this very assertion, using exactly the same words in the revenge sentence, is now true, rather than being neither true nor false.
Goldstein's reply is that, if my understanding is correct, our assertion that R is neither true nor false is a different sentence from R itself. While R is a self-referential sentence, our assertion isn't; it points to R, rather than itself. And while R fails to make a statement, our assertion about it succeeds. See Goldstein (2009, 386).

[15] See Burge (1979, 182), Gupta and Belnap (1993, 8-9), Burgess and Burgess (2011, 57) and also Beall (2001, 126).

[16] Roughly speaking, Goldstein's criterion of deciding whether a sentence can be used to make a statement is based on whether the T-biconditional for the sentence is informative. Without defining the term "informative", he argues that the T-biconditionals for both the truth-teller "This sentence is true", and the liar sentence "This sentence is not true" are both uninformative, so they cannot be used to make statement. See Goldstein (2009, 383) for relevant description and see Scharp (2013, 58-9) for some objections.

[17] See also Priest (2002), Priest (2006), and Beall (2009).

idea is now called dialetheism. Since the theory acknowledges true contradiction, it seems to be able to accommodate the revenge sentence as just another contradictory sentence.[18] Aside from the revenge problem, some point out that the theory may have trouble in delivering its own position. For the claim that *the liar sentence is both true and false* is itself paradoxical (that is, both true and false) according to its own theory (Simmons, 1993, 81). Thus in Simmons's view, the dialetheist approach also has the problem of self-refutation. But this may not be a serious problem for dialetheists, for the main point of dialetheism is not to eliminate paradoxical assertion. It may not be unacceptable for dialetheists to assert paradoxical sentences. This is different from the situation that Gap theories face.

Another sort of approach denies premise 2. That means, they reject our ordinary understanding of the concept of truth. According to this sort of solution, the unrestricted T-schema, either in their conditional form or in their rule form,[19] may only be an approximate characterization of the concept of truth, which works most of the time but may fail in some special situation. A more correct conception of truth may then suggest some more complicated and restricted version of the T-schema. An orthodox attempt of this sort in the last century comes from Alfred Tarski (1933) who makes a strict distinction between object-language (the language we speak of) and meta-language (the language we speak in), and in turn generates a hierarchy of languages. Each language at some level in the hierarchy contains a truth predicate which uniquely applies on the language at previous level and so the truth predicate in natural language is actually ambiguous—we use the same term to express different concepts of truth. The liar sentence, according to this kind of

---

[18] One can think of Priest's theory as proposing three semantic categories: true only, false only and both true and false. The revenge sentence for his theory is:
R: R is not true only.
In Saka (2007, 226-227), the negation of R is taken to be "R is true only" instead of "R is either true only or both true and false". In Saka's analysis, then, assuming that R is true only, false only, both true and false respectively, will give us contradictions. For more on this see Littmann and Simmons (2004), section 3 and 4. It seems to me that there is no essential difficulty for dialetheists to accept the accompanying contradiction as another true contradiction, but I do not intend to defend dialetheism here.

[19] See chapter 1, section 1, fn3, for these terminologies.

theory, will either be ruled out as ungrammatical or will not generate inconsistency for while it says of itself as being false at some level, it is just false (or true) at some higher level.

The idea is said to dominate the philosophy community until around 1970s, when solutions with truth value gap became popular. The most influential approach of this kind and also the most severe critique on Tarskian hierarchy comes from Kripke (1975). Kripke raises up several well-known objections to Tarskian hierarchical solution: Tarskian hierarchy is highly artificial and ad hoc; there is no reason to assume that natural languages are of a hierarchical structure; the truth predicate is univocal. On the other hand, there are reasons to believe that natural languages are not in a hierarchical structure. Kripke shows that sometimes it is difficult or even impossible to assign a correct level to the truth predicate we use and sometimes, and there are sentences whose paradoxicality is completely an empirical matter, which cannot be decided a priori by focusing on its syntax. For these reasons Kripke rejects Tarskian hierarchy and turns to his paracomplete approach. Kripke's approach has two merits. First, from technical side, he shows us that under three-valued semantics, certain formalized language can indeed contain its own truth predicate.[20] Second, from philosophical side, he defines a semantic phenomenon—ungroundedness, and the liar sentence is then understood as just one of the many ungrounded sentences. But as mentioned, the revenge paradox disrupts this kind of approach. Some concepts (e.g. groundedness) that are employed to spell out the theory cannot be expressed in the object-language that Kripke constructs on pain of being inconsistent. Thus Kripke admits that "the ghost of Tarski's hierarchy is still with us" (Kripke, 1975, 714).

Tyler Burge (1979) formulates what now is called the contextualist theory, which tries to solve the liar paradox by introducing a context parameter—the truth

[20] See Gupta and Belnap (1993), chapter 2, for discussion on this point.

predicate is context-sensitive. Burge's primary target is not the pristine liar sentence but the revenge sentence for gap theory. For him, the pristine liar sentence is just pathological and so not true. But the corresponding revenge sentence, "this sentence is not true" is not merely not true, but also true. The seemingly contradiction is resolved by parameterizing the truth predicate, so that the revenge sentence is not true in one context and true in another. One problem of Burge's approach is that, while he tells us that if the truth predicate is context-sensitive then we can resolve the liar paradox, he does not explain exactly why the truth predicate is context-sensitive, what constant nature of truth that makes its extension change from context to context (Saka, 2007, 233). Aside from the mysterious context-sensitivity, it seems that the solution cannot escape the revenge as well. One can formulate the revenge sentence "This sentence is not true in any context" and derive a contradiction after all. Burge's reply is to deny the legitimacy of the concept *being true in some context* (Burge, 1979, 192).[21] It seems to me that this is not very plausible for it seems to be a plain fact that one can easily understand his theory and his theory is that every sentence, if is true, then is *true in some context*. Rejecting the legitimacy of this concept seems to limit our ability to describe the theory.

Anil Gupta (1982) offers a revision theory, which tries to solve the liar paradox by classifying the liar sentence as unstable based on a revision interpretation of the truth predicate. Gupta and Belnep (1993) further develop the revision theory. They find a philosophical explanation for the unstability of the truth value of the liar sentence. The paradoxical behavior of the sentence is explained as just one special case of a more general phenomenon—circular definition. Truth is a circular concept, and the truth predicate is a circularly defined predicate. They argue that if we accept circular definition, then based on their semantic theory and a modified logic we can accommodate the truth predicate in a classical and even semantically closed language. The theory has a virtue in that it is less ad hoc. For it is preliminarily a

[21] Burge's original word is "true at a level" (Burge, 1979, 192), where "level" here represents "context" in his theory.

theory on definition and then is a theory on the concept of truth. Even if its solution to the liar paradox is eventually unsuccessful, this alone is not enough to undermine their general theory on definition. As to the revenge problem, the situation is a little complicated here. On the face of it, the theory introduces a new semantic concept "categorical" (or "stable", in the early version of the theory) and so a revenge sentence can be formed as "this sentence is not categorical (not stable) or not true". Some[22] has used this sentence to attack the revision theory. But this is a *too-easy* revenge for it assumes that the categoricalness concept has a classical semantics, which is denied by Gupta and Belnap (Gupta and Belnap, 1993, 255). In fact, the concept of categoricalness in their theory is also circular and so the revenge sentence surely has a similar behavior as the pristine liar sentence. To describe the semantic behavior of the revenge sentence, one can use a higher-order categoricalness concept by applying the theory on the pristine categoricalness concept. And of course, each time we form a higher-order categoricalness concept, a new revenge sentence can be constructed and the process can keep going on. Does this eventually avoid the revenge problem without triggering inconsistency or any other pathologicality? My current idea is that it may not, when the above process is carried into infinite stages. We will discuss more about this in chapter 3.

All these theories give us new insight as to what truth is, and how our language looks like if it contains its own truth predicate. However, the revenge phenomenon is very persistent in those theories and the responses to the revenge problem are often considered as being not very satisfactory. In fact, the phenomenon is so persistent that some (like Priest) have conjectured that no consistent approach to the liar paradox can be satisfactory in the sense that it will either fail to solve the revenge, or it will be ad hoc, or it solves the problem only by limiting the expressibility power of the language.[23] The conjecture is first dubbed as

[22] See Bacon (2015), p.305; Scharp (2013), p.86.

[23] Actually the conjecture has one more clause, that is, a consistent solution may fail to be satisfactory because it cannot provide a uniform solution to other self-referential paradoxes. Priest (2002) defenses what he calls *the principle of Uniform Solution* which requires that a

Dialetheist's Conjecture by Armour-Garb (2005, 94). And we will keep reviewing this conjecture later when we are examining some theories to the liar paradox.

Graham Priest (and some others) has been analyzing the structural reason as to why the revenge phenomenon seems to be inevitable. Here is one quotation from Priest (2006):

> Nor is this phenomenon [revenge] purely accidental; this situation is ultimately inevitable with any purported solution. To see this, let us start by considering the significance of extended paradoxes [revenge paradox]. The paradox phenomenon starts with a set of *bona fide*, truths, which are assertible. (Normally these will be just the plain truth, but each solution may allow us to describe them in slightly different terms—true at their rank, stably true, etc.) Those that are left over we will call "the Rest". The essence of the liar paradox is a particular twisted construction which forces a sentence, if it is in the *bona fide* truths, to be in the Rest (too); conversely, if it is in the Rest, it is in the *bona fide* truths. The pristine liar 'This sentence is false' is only a manifestation of this problem arrived at by taking the Rest to be the false. In this case, we can get out of the problem by insisting that the false is only a proper part of the Rest. This creates a gap in which the liar can conveniently lie. But this solves the problem only at the cost of showing that it was inadequately posed; for, if the false is only a proper part of the Rest, then the pristine liar is not the correct formulation of the problem. What paradoxes, such as 'This sentence is false or neither true nor false', do is remind us of this fact…In virtue of this, the only move which will produce consistency is that which bans the expressibility of certain key concepts (truth, Value gaps, stable truth etc.) from the language. (Priest, 2006, 24)

In short, the pattern Priest provides can be represented as the interaction between two sets, the Truths and the Rest. Each time when we introduce new semantic concepts, which are normally not in the set of truths, we expand the set of the Rest. The revenge sentence can be constructed by forming a sentence which says of itself as being in the Rest. Then by the T-schema, if the sentence is in the Rest, it will be

---

satisfactory solution to the liar paradox should provide a uniform solution to all self-referential paradoxes including both set-theoretical paradox like Russell's paradox and semantic paradoxes like Berry's paradox, etc. Priest argues that these paradoxes, unlike the traditional division, are essentially of one kind and so if a solution to any one of these paradoxes fails to provide a way out of the others, then the solution is not satisfactory. For a relatively short introduction, see Priest (1994). In this dissertation, my primary concern is the liar paradox and issues regarding the Principle of Uniform Solution as well as the general structure about all self-referential paradoxes are beyond the scope of this paper and so I will leave the topic here. See chapter 2, for the scope of this dissertation.

carried into the Truths and vice versa and so create inconsistency or self-refutation unless the solution bans the expressibility of relevant concepts in the object-language.[24]

Priest (2002) provides another pattern that applies to a particular kind of solutions which we may call *Parameterism*. A solution of this kind solves the problem by introducing a parameter to the truth predicate and so classifies the liar sentence as being true relative to one parameter and as being false relative to another. In Priest's idea, Tarskian hierarchy solution and some contextualism (like Burge's) all belong to this kind. The general way to form a revenge sentence for this sort is to define a concept which is the logical sum of all parameterized semantic concepts and then form a sentence which says of itself that it does not satisfy that concept. To present it formally, let c be the variable representing the certain sort of parameter (it can be, for example, "level" in Tarskian hierarchy, "context" in contexutalism), let i be the variable quantifying over the instances of that parameter and let T be the truth predicate. So for a sentence to be T_{ci} it means that it is true relative to parameter c_i. The logical sum of all parameterized concepts is:

$T_s=T_{c1}\vee T_{c2}\vee T_{c3}\ldots$

The revenge sentence is:

R_{Ts}: R_{Ts} is not T_s.

Normally, in a theory of parameterism, if a sentence is true, then it must be true relative to some parameter (like being true at level i, being true in context i, etc.). But if it is true relative to some parameter, then it is T_s. Thus assuming that R_{Ts} is true relative to some parameter will lead us to the conclusion that it is T_s, and so it contradicts with what it says. On the other hand, if we assume that it is not T_s, then it will be true again.

[24] Bacon (2015) offers something similar but is much more formal and only applies to classical theories. To put it informally and more concisely, his idea is that, each classical theory will diagnose the set of paradoxical sentences as having some features. He calls them The Unhealthys, which I think corresponds to the set of The Rests in Priest's analysis. The revenge sentence there is constructed by giving a sentence saying of itself as being unhealthy or not true.

For our purpose, it would be useful if we formalize the two patterns a little bit. Suppose that a theory introduces[25] a sequence of semantic concepts (that are in the Truths): $T_0, T_1, T_2, T_3,\ldots$then in the first pattern, the troublesome concept "being in The Rest" is constructed by forming a logical sum of the negation of each concept in the sequence:[26]

Being in the Rest= $\sim T_0 \vee \sim T_1 \vee \sim T_2 \vee \sim T_3 \ldots$[27]

The revenge sentence is formed by giving a sentence saying of itself as being in the Rest:

R_{rest}: R_{rest} is $(\sim T_0 \vee \sim T_1 \vee \sim T_2 \vee \sim T_3 \ldots)$

In the second pattern, the troublesome concept T_s is constructed by forming a logical sum of each introduced concept:

$T_s = T_0 \vee T_1 \vee T_2 \vee T_3 \ldots$

And the revenge sentence is formed by giving a sentence saying that it does not satisfy T_s:

R_{Ts}: R_{Ts} is *not* $(T_0 \vee T_1 \vee T_2 \vee T_3 \ldots)$

I will call the first type of revenge sentence, *type S* and the *second type, type P*. It is not clear at this moment whether there is some intrinsic connection between these two types of revenge, but at least they are not logically equivalent.

Priest's analysis works in many cases, but there may be exceptions. Recall Beall's warning about too-easy revenge. Simply proving that certain relevant concepts are not expressible on pain of causing inconsistency in the object-language by itself

---

[25] Some of the concepts in the sequence need not to be newly introduced by the theory. For example, one of the concepts here can be the concept of truth, or the concept of "either true or false". The reason I speak of them uniformly as some concepts being introduced is just for simplicity.

[26] Let A be a concept. By the negation of A, I mean not-A, represented by "~A". For example, if A is the concept of truth, then the negation of this concept is not-true. In the case of truth value gap, I take the semantic concept neither true nor false to be the negation of the concept "either true or false".

[27] Note that while the set The Rest is described as a conjunction: ~T1 and ~T2 and ~T3…, the concept *Being in The Rest* is defined as a disjunction of being ~T1, being ~T2, being ~T3…

does not necessarily damage the theory. The defender may argue that the relevant semantic concepts do not have classical semantics or that the relevant concepts (like R_{rest} and R_{Ts}) are not intelligible at all.[28] In the former case, even though the revenge sentence can be constructed, it may not cause inconsistency, at least, not semantical inconsistency. In the latter case, the theory can blame the problematic concept for causing inconsistency. We will have some detailed examination of some theories in chapter 3.

There is one more possibility to avoid Priest's two patterns. If a theory, while allowing the construction of the relevant revenge sentence, denies the applicability of relevant instance of the T-schema, then one can successfully classify the revenge sentence as being in the Rests without having any pressure to further classify it into the Truths. I believe that Horwich's *Semantic Epistemicism*, which will be discussed in chapter 4, belongs to this kind. One consequence of this strategy is that we will have some sentences that are assertable, but we cannot call them truths. Another problem is that since the revenge sentence is classified as being pathological in some sense, the theory suggests that we can assert some pathological sentence (this is the refutation problem). Whether these two consequences can be explained or explained away depends on the details of concrete theories and I will not go on to discuss it here.

---

[28] I should mention that there is one stronger objection to the absolute concepts like R_{rest} and R_{Ts}. That is, they are not expressible at all, even in natural language. The idea is that the set of possible truth value is, according to their term, *indefinitely extensible*, and so there can be no way to express what we intend to express by using phrases like "all truth values". See Cook (2009) for such a position. Note that, the position is not that these notions are not expressible for otherwise we will have a contradiction in certain artificial language, but that they are not expressible at all even in natural language. For otherwise, the above argument for the absolute revenge still holds, and one can only resolve the revenge by limiting the expressibility of natural language, which is not against the *Dialetheist's Conjecture*. The position indeed deserves further examination, which may take us into much more complicated subject involving the concept of infinity. And so I do not plan to provide a full and final response to this position in this paper. For now, my quick response is that, there seems to be little hope in maintaining the inexpressibility of any such absolute concept. Even Cook himself understands that, if we take the phrase like "all truth values" to mean what it intuitively means, an absolute revenge is inevitable. So I tend to believe that, the position at best, leads us to a normative project, rather than a descriptive one (for distinction of these two projects, see chapter 2). And it is sufficient, for my current purpose, to accept that these terms are expressible in natural language.

Though it remains a question that whether those theories will be severely damaged by the revenge problem, many philosophers have been trying to think of another way which can be revenge free. What is certain now is that being revenge free is indeed one widely accepted criterion of being an adequate solution to the liar paradox.

3. Silence approach?

Let us summarize what we have seen so far. The liar sentence causes inconsistency in our ordinary understanding of the notion of truth and logic and also language. To avoid inconsistency, philosophers have been trying to figure out exactly what is wrong in our ordinary conceptions. Solutions normally will define a different semantic feature in contrast to the classical semantic feature (e.g. any sentence is either true or false). Some propose that perhaps, unlike what we think, the liar sentence is actually neither true nor false. Some propose that perhaps the truth predicate is ambiguous (e.g. it is context-sensitive) or that the truth predicate is circular. These theories result in different semantic characterizations of the liar sentence. The trouble is that, each time we try to solve the problem by a different characterization of the liar sentence, we find that a revenge sentence can be constructed using the newly introduced semantic concept, and which will probably cause inconsistency or some other problem. Although the relevance of the revenge sentence cannot be established a priori, one can almost be certain that such a sentence can be constructed using the two patterns provided above.

Now the question is, can there be a solution to the liar paradox that can be known a priori to be revenge free in the sense that at least no revenge sentence can be constructed? This seems to be possible. If the revenge is caused by the introduction of some new semantic concepts, then one way to avoid it is perhaps is to avoid introducing those concepts. If the liar sentence will be a problem when we try to describe its semantic status, then perhaps giving up such an attempt is a way to

avoid it. In other words, perhaps, what the liar sentence teaches us is that there is some limits to our thought that cannot be described or thought without falling into inconsistency. And the solution to it is just to give up the attempt and acknowledge that the truth status of the liar sentence is something that cannot be identified. And I would like to call this approach as *silence approach*. A similar idea has been mentioned in some literature. Beall (2007, 4) mentions a kind of approach which he calls quietist approach whose main idea is that given most of the attempts to solve the liar paradox fail, it seems we should give up the attempt to classify the truth status (or truth category, as Beall calls) of liar-like sentences, and "whereof one cannot truly classify, thereof one must…be silent" (Beall, 2007, 4). Beall thinks that since this kind of approach does not engage in the classification, that is, it remains quiet as to what is the truth value/truth status of the liar-like sentences, it does not have a chance to evoke the revenge problem, but "it offers no clear account of truth or the paradoxes at all", and thus little can be said about it.

The silence approach is the topic of chapter 4. My study on the silence approach is in an initial stage. So instead of providing a well-developed theory of the silence approach, I will simply sketch the idea in some general way, and then examine some current theories that seem to satisfy at least part of my characterization of silence approach and see whether there is any hope to develop it any further in the future.

4. **Aim and structure**

The purpose of this dissertation consists in two tasks. First, I want to discuss some prominent approaches to the liar paradox. The main aim of this task is to show that the revenge problem is a general challenge without trying to settle the concrete debate of whether they can eventually solve or avoid the problem. Second, I want to characterize what I have called silence approach which I think may escape the revenge problem and then examine some of the candidate theories that I think fit at least part of my characterization. The aim of this task is to see whether there is any theoretical advantage of this sort of theories against the revenge problem that make

them worth future development.

The structure of this dissertation is as follows:

In chapter 2, I will discuss some methodology. I will define precisely what liar paradox is and the problem it poses. I will also sketch the formal tool that is employed by many theories to resolve the liar paradox. And finally, I will provide some criteria for assessing a solution. The scope of this paper will also be delineated

In chapter 3, we will see some concrete solutions to the liar paradox, some of which have been sketched in the literature review. In this time, I will provide a more detailed discussion on the theories, especially the revenge problem they face.

In chapter 4, the basic idea of silence approach will be characterized and some candidate theories that seem to adopt a silence position toward the liar paradox will be examined. And a conclusion will be given.

Chapter 2 Methodology

This chapter focuses on the methodology. We are looking for a solution to the liar paradox. But before we look for a solution to a problem, we should first identify the problem itself. Precisely what is the liar paradox, and what are the main problems the liar paradox poses? And if there is a solution, in what sense is it a solution? In other words, what is necessary for a solution to solve the liar paradox? And what counts as a good solution to the paradox? In addition to these questions, I will also introduce (though not in any detailed way) the technical tool, language model, that many scholars have employed in giving their own solutions. I should mention that the use of the technical tool is not necessary. At least, the tool does not appear in *every* literature aiming at solving the liar paradox, but it is widely employed, especially for those who require precision. Finally, I will point out the scope of this dissertation.

1. Problem identification

1.1 The paradox

By *paradox*, I mean an argument or reasoning, which begins with some seemingly acceptable premises, following with some seemingly acceptable rules of inference, but leads us to a seemingly unacceptable conclusion. By *liar paradox*, I mean the paradox generated by a kind of sentences, which I will refer to as liar-like sentences. The distinctive feature of this kind of sentences is that, they all (at least, most of them) are equivalent to some sentences which say of them (the liar-like sentences) as being false, or not true, or something alike.

Consider the following sentence:

The first displayed sentence of this chapter is false.

Now let us make the following reasonable assumptions:

First, reasoning by cases is valid;[29]

[29] Reasoning by cases is a kind of reasoning with the following forms:

Second, law of excluded middle is valid;

Third, law of non-contradiction is valid;

Fourth, the truth predicate obeys *Capture* and *Release* in conditional form, where Capture and Release are the following conditionals:

Capture: A→ T"A"

Release: T"A"→ A

where T is the truth predicate, "→" is a conditional (material conditional, for example). A is any sentence with "A" its quotation name.[30]

Assumptions 1 to 3 constitute one understanding of our logic (or the notion of validity—what is valid and what isn't); assumption 4 constitutes one understanding of the notion of truth and the predicate denoting it. If one indeed understands our logic and the notion of truth in the above way, then, if one also accepts that the first displayed sentence is one of the legitimate (and meaningful) sentences in natural language, then a contradiction will arise, causing an inconsistency in the above seemingly reasonable four assumptions.

By 1, we can reason by cases, taking into consideration both the case where the first displayed sentence is true and where it is false; by 2 the sentence is either true or false; by 1 and 4, one can derive that the sentence is true iff it is false, and so by 2 it is both; by 3, it can't be both.

So from the seemingly acceptable premises, reasoning via seemingly acceptable rules of inference, we are led to a contradiction, that the first displayed sentence is both true and false (which is incompatible with assumption 3). But things get even worse if one also accepts totally a classical framework of logic. In particular, if one accepts that the inference from a contradiction to everything as a valid inference,

---

Suppose we assume A, and from A we derive C; now suppose we assume B, and from B we derive C, so we conclude that either A or B, we have C.

[30] See chapter 1, section 1, for these conditions.

then from the above contradiction (that the liar sentence is both true and false), one can infer everything, which means that everything is true according to this understanding. Let us call the inference from a contradiction to everything, *explosion*.[31] And logic that validates explosion will be said to be *explosive*. And let us call the position that everything is true, *trivialism*.[32] Classical logic is explosive, and so if the above argument is accepted, then trivialism follows. This is the liar paradox.

1.2 The problems

Below I briefly sketch three problems posed by the liar paradox, which I take to be the fundamental issues the liar paradox generates.

The first problem is the problem of triviality of our language. This dissertation assumes, without further justification, that trivialism is irrational and so should be rejected. Given that trivialism is bad, any theory that leads to it is problematic. But if we accept that our language (natural language) obeys classical logic, and it contains a truth predicate which obeys Capture and Release in their conditional form, then it seems that the language is trivial.[33] So the first problem posed by the

---

[31] The term "explosion" is from Priest, Tanaka, and Zach (2018). The inference from a contradiction to everything is called *ex contradictione quodlibet* and can be found in Priest, Tanaka, and Zach (2018). Here is my summary of how it goes:

1. $A \wedge \sim A$ (premise)
2. A (from 1 and conjunction elimination)
3. $A \vee B$ (from 2 and disjunction introduction)
4. $\sim A$ (from 1 and conjunction elimination)
5. B (from 3 and 4 and disjunctive syllogism)

where B is an arbitrary statement. The reasoning shows that from a contradiction one can infer everything.

[32] The term "trivalism" is from Priest (2000).

[33] One may find it odd to say of a language as being trivial. Beall (2006, 190, fn5) explains that a language or theory is trivial if according to it everything is true. Some, like Burge (1979) and Scharp (2013) may have question about whether it really makes sense to say of a language as being trivial or as being inconsistent, since language itself does not postulate anything—a language is not a theory. However, if we consider a language as a combination of syntax, semantics and logic, then it does make sense to treat a language as a theory, especially when we take into consideration the fact that there are many sentences in the language that are regarded as being true or false simply because what they mean or because of the ground logic. A similar idea is held by Ray (2006, 168), where Ray suggests that the logic we specify for a language can be regarded as identifying the meaning of relevant logical constants (e.g., logical

liar paradox is how to avoid the triviality of our language, given that it enjoys the above features (see last subsection). More *generally* put, the hard challenge the liar paradox gives us, is to show that despite the fact that our language enjoys some *desired* syntactic and semantic features,[34] and that some *desired* logic holds for the language, the language is not trivial.[35]

The second problem posed by the liar paradox arises from an intuitive idea, that is, it seems quite reasonable to say that we can truly and exhaustively classify every sentence in our language into the following three significant semantic categories: true, false, and *other*, where "*other*" here stands in for some semantic categories (Beall, 2006, 191). The classical picture has it that every sentence (those declarative sentences) in our language is either true or false, in which case, the above category "other", is in effect, an empty set.[36] But the liar paradox makes this position very doubtful. For if every sentence is indeed either true or false, then with the above understanding of our language and that of the notion of validity and truth, we are forced to accept that the liar sentence is also both true and false, which is incompatible with the classical picture. But if the semantic feature of the liar sentence is not both truth and falsity, then what is it? Some more categories may need to be supplied.[37]

---

connectives), and that the logical relations between sentences are obtained in virtue of what those sentences mean. In this way, it makes sense to say that the ground logic constitutes part of the language at issue.

[34] That the language enjoys a truth predicate which obeys Capture and Release in conditional form is just one (perhaps an intuitive one) conception of our language. Actually, in solving the liar paradox, those who blame the naïve theory of truth for causing the liar paradox have tried to solve the problem by giving new conception of truth. For more on this point, see next section.

[35] See Beall (2006), section 3, and Beall (2007), pp.6-7, for Beall's description on the first and second problems. My presentation here is a summary on both references listed here.

[36] That is, there is no semantic category other than "true" and "false".

[37] Beall (2006, 191) refers to categories like "being true", "being false" as significantly semantic categories. Beall admits that the notion is imprecise. Actually, even the term "semantics" is imprecise. Tarski takes semantics to be a discipline dealing with "*certain relation between expressions of a language and the objects (or "states of affaris") 'referred to' by those expressions.*" (Tarski, 1944, 345). Truth is defined in terms of semantic concepts (concepts that express the relation between syntax and world), so truth is a semantic concept. I have no idea whether what Beall has in mind is anything similar to this definition, but it seems to be nothing wrong to treat "semantic categories" as categories that are related to the categories of being true, in one way or the other.

The third problem stems from the second one. That is, suppose that we eventually come up with some semantic theory of our natural language, which provides some new semantic category, say, A, for truly classifying those liar-like sentences. It seems that introducing A into the language at issue generates a new kind of liar sentences which cannot be truly classified into category A. For consider the following sentence:

The second displayed sentence in this chapter is not true or A.

Following exactly the *same* sort of reasoning generating the liar paradox, we can easily obtain that the second displayed sentence in this chapter is both true and not true (if being A implies being not true). This phenomenon is the revenge problem that we have seen in chapter 1. At least on the face of it, the occurrence of the revenge phenomenon shows that the language remains inconsistent or trivial (if the logic for the language is explosive). Thus the third problem takes us back to the first.[38]

In this thesis, I will refer to the first problem as the triviality problem, the second, the classification problem and the third, the revenge problem. I take these three problems to be fundamental, in the sense that any solution to the liar paradox will not count as adequate if it fails to solve any one of them. For without solving the first our concern about the triviality of our language remains there, without solving the second, our puzzle caused by the liar remains there, and failing to solve the third reveals that the solution fails to capture the essence of the liar paradox—it fails to prevent the reoccurrence of contradiction generated by liar-like sentences.

2. Solution

Before we seek for a solution, we should make it clear what counts as a solution (or a good solution) to the liar paradox. In particular, we should answer, in what

[38] The seriousness of the revenge problem is widely accepted. For example, Burge (1979, 173) takes the revenge problem very seriously. In his view, the recurrence of the same kind of paradox is not a drawback to the solution, but a sign showing that the solution fails to account for the basic phenomenon.

situation, the liar paradox can be regarded as solved.

Recall that a paradox is an argument or reasoning that begins with some seemingly acceptable premises, but leads us to some seemingly unacceptable conclusion via some seemingly valid rules of inference. In the case of the liar paradox, it is our conception of the notion of truth, logic, and language that leads us to the unacceptable conclusion—that the liar sentence is both true and false, and which further implies that the language is trivial. Now, solutions to the liar paradox can be roughly divided into two kinds (though I do not mean to be exhaustive): the first kind rejects the argument, and so rejects its conclusion (that the liar sentence is both true and false); the second accepts the argument and so accepts the conclusion. If it is the first kind, given that it rejects the conclusion, it should point out exactly which premise is unsound (say, not true), or which rule of inference is invalid, and what's more important, *why*. In short, it should tell us exactly what's *wrong* in the liar paradox that leads us to the contradiction and it should give us an independent reason as to why it is wrong (Priest, 1979, 220). In some sense, answering this "why" question is of most importance in solving *any* paradox. For first, without answering why those seemingly acceptable premises are actually unacceptable, or why those seemingly valid rules of inference are actually invalid, the solution is ad hoc. An ad hoc solution seems to be able to "solve" any problem. Second, without providing answer to this "why" question, our puzzle caused by the paradox remains there, for it forces us to give up some assertions or rules of inference that we normally take to be unproblematic. As Barwise and Etchemendy nicely put, "we see that they are false, without understanding why" (Barwise and Etchemendy,1989,7). Following Chihara (1979),[39] I will refer to this "why" problem as the *diagnostic problem*, and call any project aiming at solving the diagnostic problem as the diagnostic project.

[39] Chihara takes the diagnostic problem very serious: "…it is clear that nothing should be called a solution (or resolution) of the paradox that does not solve its diagnostic problem." Chihara (1979, 591).

On the other hand, there are solutions that accept the conclusion that the liar sentence is both true and false. Theories of this kind accept both the premises and the rules of inference employed. In other words, our conceptions[40] of truth, logic, and our language, at least, in relevant aspects, are *correct*, in the sense that what we understand about them correctly describes or reflects the very nature or features of these entities. It would be wrong, on the contrary, to reject any one of them if one aims at describing *faithfully* our natural language and relevant notions. But if trivialism is to be rejected, then this kind of approach should tell us, exactly how, despite having a true contradiction in our language, the language can avoid triviality. At the very least, it should provide insight as to how one can live with some true contradiction without causing any practical problem.[41] Alternatively, if a theory accepts, altogether, the triviality of the language, then the solution may provide new conception of relevant notions which can satisfy our practical usage.[42] In other words, this second road abandons the attempt either to show that our natural language is good (since they believe that it is not) or that we can live with contradiction, but choose to *revise* the language by replacing relevant notions with some unproblematic ones. I will from now on refer to these two kinds of approaches (approaches that accept the conclusion that the liar is both true and false) together as the *inconsistency* approach.

The distinction between the above two kinds of inconsistency approach reveals a distinction between two more general projects. Borrowing the terms used in Gupta and Belnap (1993), I will refer to these two general projects as *descriptive* project

---

[40] I use the term "conception" in a very loose way. It can be used to refer to our general understanding of a concept, the logical behavior of the predicate denoting the concept, or some facts about the concept at issue.

[41] Gupta and Belnep (1993) rejects Chihara (1979)'s inconsistency theory of truth exactly because of this reason. Gupta believes that Chihara does not provide an account as to how ordinary people could use the notion of truth without any trouble, if the notion is inconsistent. See Gupta and Belnap (1993), chapter 1, section 3.

[42] When I say, a theory accepts the triviality of the language in use, it does not mean that the theory believes that therefore our natural language is unproblematic, or that trivialism is good. What I mean here is just that, the theory, believes that the language at issue *is* indeed trivial, as a matter of fact.

and *normative* project respectively. Descriptive project aims at giving a solution for language *in actual use*. That is, it tries to faithfully tell us, exactly what is happening, when the liar paradox occurs. If it rejects the conclusion, then what it tells us will be that we *misuse* some rules of inference, or due to some reason, we *wrongly* believe in some of the premises, or that the notion of truth has a different feature from what the naïve theory of truth[43] tells us, or if the solution is an inconsistent one, it tells us that our language is, in fact, not explosive, etc. All in all, a descriptive project solves the problem not by making any *change* to our language but by *clarifying* relevant conceptions. On the other hand, a normative project may not aim at giving a solution to natural language—language in actual use. The systems constructed in this kind of project all find out one way or the other to block the occurrence of contradiction or triviality in an artificial language. But the artificial language may not perfectly represent our natural language (or a fragment of it). In most of the cases, this is so because the artificial language is expressively weaker than natural language. One interesting thing is, most of these systems constructed do not aim at a normative project at the beginning, but due to the revenge problem, they end up having trouble in justifying their *descriptive adequacy*[44]—that their theories are descriptive enough for our natural language. Some normative solution give up the attempt to solve the problem for natural language, but argue that natural language is hopelessly inconsistent (or trivial), and what the liar teaches us is that we should replace relevant notions with some new ones.[45]

In summary, if a solution rejects the conclusion of the liar paradox, then it should solve the diagnostic problem; if it accepts the conclusion, then it should find a way

[43] By "the naïve theory of truth" I mean a kind of theory that implies unrestricted usage of Capture and Release.

[44] Simmons (1993) has offered substantial arguments (based on what he calls diagonal argument) to show that many leading solutions fail to be descriptive adequate, in the sense that they fail to give a model for language at actual use, due to expressive limitation. See chapters 3 to 4.

[45] Scharp (2013) holds this view, see Scahrp (2013), pp.1-2, p.8 and p35 for relevant claims. Tarski (1933) seems to hold a similar view. He thinks that natural language is inconsistent, but he does not require for a replacement. See Tarski (1933), pp.164-165.

to prevent trivialism or at least provide insight as to how to live with contradiction; if the solution is a descriptive one, then it should justify its descriptive adequacy; if it is a normative one, then the basic task for it is to maintain that natural language is inconsistent and that the new language obtained (either by limiting relevant notions or replacing them with ones) is consistent or not trivial.

3. Language model[46]

3.1 The need for a precision tool

Now since we have an idea of what the problem is, and what kind of solution counts as a solution to it, I will go on to introduce the tool that many scholars employ when giving their own solutions to the liar.

Recall that the hard challenge (the first problem) the liar paradox gives us, is to show that despite the fact that our language enjoys certain *desired* syntactic and semantic features, and that some *desired* logic holds for the language, the language is not trivial. But what are these desired features? In section 1.1, as an example, the language at issue is said to contain a truth predicate which obeys Capture and Release in their conditional form; the liar sentence is constructed via self-reference (using definite description); the logic at play is said to be classical which is explosive and validates the law of excluded middle. But these may not be the features or the only feature that one may desire. To begin with, the truth predicate may be characterized as one that obeys Capture and Release in their *rule* form rather than conditional form, or according to one salient view,[47] the essential feature of truth is its transparency, which means A and T "A" are *intersubstitutable* in all non-opaque[48] contexts, for all sentence in the language; some reject that the notion of

[46] The main point about the formal method—language model, in this section is basically a summary (with some extension from other materials) on Beall (2006), pp.195-196 and Beall (2007), chapter 1, section 1.3.2.2.

[47] See Beall (2006), pp.187-188. But note that, whether transparency is the essential feature of truth is disputable. See Priest (2006), section 4.9.

[48] Beall uses the terms "non-opaque context" or "transparent context", in some places, e.g., Beall (2006), p.188 and p. 191, fn11, and Beall (2009), p.1. But he never provides a precise definition of it. For our current purpose, we can understand a non-opaque context as such a

truth is univocal (and so reject unrestricted intersubstitutability of truth), instead, they may insist that truth is *indexical*, say, being *context-dependent* (recall Burge's contextualism). Second, due to Gödel's work,[49] we now know that a liar-like sentence needs not to be a self-referential one, what is required is only that, there is a sentence that is *equivalent* to a sentence saying it (the liar sentence) as being not true. This means that the syntactic resources the language needs in order to construct a liar sentence is very minimal—as long as the language contains the *language of arithmetic*,[50] liar-like sentences can be constructed without using any explicitly self-referential device like naming or definite description.[51] Third, the logic holds for the language at issue needs not be the classical one. There are many alternative logical systems in the market in addition to classical logic (recall the truth value gap theory and the inconsistency theory). Putting aside the issue of descriptive adequacy, a solution to the liar paradox thus demands a proper arrangement and combination of the above features (features of the language, truth and logic, etc.) in a way that the whole language system can achieve consistency or avoid triviality.

Above I say that according to one conception of truth, the essential feature of the truth predicate is its transparency. But exactly what does this feature amount to?

---

context that the principle of composition holds, that is, the truth value of a compound sentence depends solely on the truth value of its constituents. An example for the opaque context is belief context. The truth value of the sentences "the morning star appears at dawn" and "Venus appears at dawn" are exactly the same (they are both true, at least in the actual world); but the sentences "John believes that the morning star appears at dawn" and "John believes that Venus appears at dawn" may have different truth values, since John may not know that the morning star is Venus. In the latter case, the truth value of the whole sentence does not solely depends on the truth value of its components.

[49] Here "Gödel's work" I mean Gödel's work on Gödel's theorems. Gödel(1931) invents a technique which is now normally called Gödel numbering. The technique is essentially a coding method, which enables one to construct an arithmetical sentence (sentence talking about numbers) that is equivalent to a syntactical sentence (sentence talking about the syntax of the language coded). In such a way, as long as a language contains the language of arithmetic, it can generate a sentence which, though it talks about numbers, is extensionally equivalent to a sentence talking about the language itself. In this way, the language can generate a sentence talking about its syntax without using explicit referring devices like naming. For more on Gödel numbering, see Smith (2013), chapter 19.

[50] For the syntax of language of arithmetic, see Smith (2013), chapter 5, section 5.2.

[51] For an introduction of three different reference devices, see Beall, J.C. Glanzberg, M. and Ripley, D. (2018), chapter 4, section 4.2.

Also, I say that as long as the language at issue contains the language of arithmetic, it can talk about its own syntax. But what is it for a language to "contain" another? And before that, what is a language, how to represent it? Third, I have been talking about some logic (classical logic) as holding for a certain language. In what sense do we say a logic *holds* for a certain language? What's more, I say that a solution to the liar paradox demands a proper arrangement and combination of the above features (features of the language, truth and logic, etc.), but how to represent this "arrangement"? And how can we know that the features we select for our language do not themselves conflict with each other? In other words, how to make sure that the final language we construct, which combines all the desired features of language, logic and truth, is indeed consistent?

It is not the right place to give detailed answers to the above questions, what I can do here is to introduce the tool many scholars use when they try to give a precise account to our language, logic and the notion of truth. The tool is the language model.[52]

3.2 Language model

The use of model is quite common in science. For example, in order to study the motion of a car, we need not to actually look at a real car, which has extension in three-dimensioned space. Rather, we can study the motion of a car by studying a two-dimensioned geometric model which *represents* the motion of the car at least in some relevant aspects. One can use a geometric model with a line representing the distance the car traveled and a dot representing the car itself; or more abstractly, we can simply represents the motion of a car by some mathematical formula, which can be regarded as a mathematical model for motion. These descriptions of model for motion may not be accurate or correct, but the upshot is that, instead of study the object directly, we can study some abstract model which is said to represent the

[52] Beall (2007, 7) uses "model language", but I somehow find that "language model" is more natural.

real object (instead of studying a real car directly, we study a dot that represents the car).

In the study of language, the situation is similar. The liar paradox raises the question concerning how our language (natural language) can enjoy several desired features (e.g. contain a transparent truth predicate, be syntactically rich, obey classical logic…). But natural language is a mess. We need to abstract ourselves a little bit by focusing on a formal language which is said to be similar enough to natural language at least in the aspects that we care about. Beall (2006) has a nice introduction:

> …even though our aim is 'real language', we must none the less abstract a bit from the mess. The aim of formal accounts of truth, at least those concerned with 'real truth' in natural language (or the very language 'we' speak), is not to give an account of truth, but rather truth-in-L, for some formal 'model-language' L. The relevance of such an account is that 'real truth' is supposed to be 'similar enough' to L-truth, at least in relevant respects, to gain answers to NTP and, perhaps ECP. (Beall, 2006, 196)[53]

Here NTP and ECP correspond to the first and second problem in section 1.2. The upshot is that, if we want to show that our natural language does enjoy several desired features, then we do that by showing that an artificial language L, which models our natural language, does enjoy those features. Normally, we assume that a *semantic-free* language (a language with no semantically significant predicate like truth predicate) is paradox-free[54] (that is, free from semantic paradox like the liar), and so the construction begins from a semantic-free language, a base language, say, L, and then we proceed by trying to expand the language to a richer language, say, L^+, which is exactly the same language as L expect that it contains a predicate T. Our job is to show that, despite the fact that L^+ has certain desired features, we can interpret the predicate T, as a truth predicate (with desired features, like being transparent) for L^+ itself, and do so without generating the liar paradox or

[53] See Chihara (1979), p. 616, for a similar point.
[54] This assumption is harmless, for the priory task is to study the paradox, paradox arises from semantically significant predicate.

trivializing L^+.

Now the question is, exactly how do we specify a language model?[55] Normally, one can consider an *interpreted* language as a combination of three elements: syntax, semantics, and the logic that holds for the language.[56] In practice, an interpreted *formal* language is specified by a triple <L, M, σ>,[57] where L provides the syntax information of the language, M provides an interpretation on the syntax (according to a semantic theory), and σ is valuation schema, which provides information about how to evaluate the truth value of a compound sentence from its components.

For L, normally, we use the syntax of first-order language[58] and stipulate (for our purpose), that the language does not contain any semantically significant predicate, in particular, we stipulate that it does not contain T (which is to be interpreted as the truth predicate for *L*). For M, which is called an interpretation or a model, normally, it is specified by an ordered pair <D, I>, where *D* is *domain*, the universe of discourse, which consists of a set of objects. Intuitively, a domain is the set of objects that we can refer to in the language. I, is an interpretation function which does the following assignments:

For any name in *L*, I assigns to each name an object in the domain D.

---

[55] The following specifications of a language model are quite standard, and can be found in many textbooks on logic. The specification here is from Beall (2007), p.23.

[56] The idea that in order to specify a language one needs to specify the logic it obeys, as far as I know, comes from Tarski (1944).

[57] Gupta and Belnep (1993) has a nice explanation of the significance of this kind of construction:

"…the specification of these triples does not fix the *meanings* of the nonlogical constants, only their significations [general version of "extension"]. One cannot determine the meaning of an expression from its signification, since expressions with identical signification may differ in meaning. Nor can one determine signification from meaning alone one needs to know the relevant facts also. Hence, the constructs we are calling 'interpreted languages' are not *interpreted* languages in the ordinary sense of the term. Nonetheless, these constructs are useful, since they carry the kind of information that is needed to fix the truth or the falsity of every sentence of the language." Gupta and Belnep (1993, 45).

[58] First-order language contains four logical operators and two quantifiers. It can be used to formalize most of our declarative sentences, where no modal operator, like belief operator, possibility and necessity operators, etc., is involved. And the quantifiers it contains can only be used to quantify over objects, not over properties of object. One can find a standard introduction to first-order language in any textbook for modern logic, at least, in those that cover predicate logic.

For n-ary function symbol, I assigns to each function symbol a member of $D^n \rightarrow D$, that is, a function from n-tuples of D into D.

For n-ary predicate symbol where the language adopts a two-valued semantics, I assigns to each predicate a member of $D^n \rightarrow \{1,0\}$, that is, a function from n-tuples of D into {1,0};

When extension is concerned,

I assigns to each n-ary predicate symbol F an n-tuples $<a_1, \ldots a_n>$ of D, such that $I(F)(<a_1,\ldots a_n>)=1$.

When anti-extension is concerned,

I assigns to each n-ary predicate symbol F an n-tuples $<a_1, \ldots a_n>$ of D, such that $I(F)(<a_1,\ldots a_n>)=0$.

The semantic schema is the classical one (just as an example). In particular, we have:

A negation of a sentence is true iff the sentence is not true;

A conjunction is true iff both conjuncts are true;

A disjunction is true iff one of the disjunct is true;

A universal quantification $\forall(x)F(x)$ is true if and only if for every object d in the domain D, we have F(**d**) is true, where **d** is the name for the object d (For simplicity, here we assume that each object in the domain has a name in L for it).[59]

Given any model *M*, and semantic schema, we can further define the key model-theoretic notion *truth-in-L* and the notion of validity. For example, we define a valuation function *V* such[60] that it assigns value in the set {1,2} to sentences in L (according to M and the semantic scheme above). For any sentence *p* in L, *p* is said to be true in L if $V_M(p)=1$. If for any model M, $V_M(p)=1$, then *p* is said to be a logical truth. For the notion of validity, which is often represented by the notion of *semantic consequence,* we have the definition: for any set of sentences Σ (possibly empty) and any sentence *p*, *p* is a semantic consequence of Σ if and only if for any

[59] The semantics for quantifier in first-order logic is a little complicated. For a strict version, see Boolos, Burgesss and Jeffrey (2007) or Sider (2010).

[60] A full definition of valuation functions involve recursive clauses for compound sentences, but for our current purpose, we need not go that far.

model M, if $V_M(q)=1$, for each $q \in \Sigma$, then $V_M(p)=1$.[61]

For a language specified in this way, it can be easily checked (but not here!) that some familiar rules of inference like the law of excluded middle, the law of identity are valid in the above sense. In next chapter we will see another kind of language model in which these two laws are not valid. But here let us say something about the philosophical issue of the above semantic theory.[62]

One may have question about the relation between the above semantic theory and our real language. For example, what is the relation between the set {1,0} and our real language? Why posit these two numbers? Beall (2009) holds an instrumentalist position, according to which, this kind of semantics is nothing more than a useful tool to *establish* certain facts about our language (e.g. the logic holds for it, or some behaviors of relevant notion, like truth), and we need not give too much philosophical sense on these theories (Beall, 2009, 116). But I do not think that this is the right position. In my opinion, a language model does not only *establish* certain facts about our language, it *illuminates* it. The very need of abstraction (using formal language model) is to free us from many unnecessary elements in the real objects and to help us focus on the targeted features that we want to study. Priest (2008) complains about the above kind of instrumentalist position:

> …there is something very unsatisfactory about this, as there is about all instrumentalisms. If a mathematical 'black box' gives what seem to be the right answers, one wants to know why. There must be some relationship between how it works and reality, which explains why it gets things right…The most obvious explanation in this context is that the mathematical structures that are employed in interpretations *represent* something or other which underlies the correctness of the notion of validity. (Priest, 2008, 28)

Why positing number 1 and 0 in the above two-valued semantics? The answer is

---

[61] Beall (2007) does not include a definition for semantic consequence. This definition is a modification of the one given by Sider (2010). See Sider (2010), p.122.

[62] The term "semantic theory" is indeed not very precise. I do not have a standard definition of "semantic theory", nor did I find any so far. Here I use the term "semantic theory" to refer to theories specifying the second and the third elements in the triple <L, M, σ> and some related semantic notions, like truth-in-L and validity.

that, exactly which number we use does not matter at all. All that matter is that the behaviors of number 1 and 0 in the above semantic theory can represent the behaviors of being true and being false in our real reasoning and in our real language. And it does, so it works.

4. The scope

The liar paradox is a very complicated problem and a comprehensive study of the paradox is clearly beyond the scope of one short dissertation. So I have to limit the scope of this paper and focus on a relatively small aspect. In particular, there are at least two problems that I will not discuss in this dissertation.

The first is about the general structure of liar-like paradoxes and the uniform solution to all self-referential paradoxes.

The liar paradox is not a single paradox. It is actually a family of liar-like paradoxes, each one of which is generated by a different kind of liar-like sentence. In addition to the pristine liar sentence in the introduction chapter (which is the simplest form of liar-like sentences), many variants have been found and discussed.

For example, we have some empirical liar-like sentences, whose paradoxicality depends on empirical facts:

S_1: *The first displayed sentence in this section is false.*

We also have some liar-like sentences which have more complex logical structure:

S_2: This sentence is not true and 1+1=2.

Or a combination of the above two types of liar-like sentences:

S_3: This sentence is not true and snow is white.

We also have loop-liar paradox:

S_4 (asymmetric pair):

Socrates: what Plato says is false.

Plato: what Socrates says is true.

Or its variant

S_5 (symmetric pair):

Socrates: what Plato says is false.

Plato: what Socrates says is false.

S_6 (asymmetric open pair):

Socrates: what Plato says is false.

Plato: if what Plato says is false, then what Socrates says is false.

We also have chain paradox:

S_7:

(1) the next sentences are false

(2) the next sentences are false

(3) the next sentences are false

…

These variants of the pristine liar paradox have significant influence in our understanding of liar paradox. For example, S_1 teaches us that there seems to be no way to determine whether a sentence is paradoxical by merely focusing on its syntax (Kripke, 1975). S_5 with the principle of symmetry[63] seems to suggest that each sentence in the pair should receive the same non-classical truth value, but the same treatment cannot be applied in S_6, since the two sentences in S_6 are not symmetric (Armour-Garb and Woodbridge, 2006). S_7 shows that perhaps self-reference is not a necessary condition for generating a liar-like paradox (Yablo, 1993).[64] These variants, at least, involve the concept of truth, even though not all of them are structurally similar to the pristine liar paradox. But some have argued that the liar paradox is essentially the same kind of paradox as other self-referential semantic paradoxes which do not involve the concept of truth, like Berry's paradox and even set-theoretical paradox like Russell's paradox (Priest, 2002). If so, then according to the principle of uniform solution, a theory on the pristine liar paradox should be able, at least in a general way, to provide a solution to other paradoxes of

[63] The principle of symmetry is that, if two sentences appear in some symmetry form, then if there is no independent reason, the two sentences should receive the same truth value.

[64] For an objection to taking Yablo's paradox as non-self-referential, See Priest (1997).

the same kind. To examine whether the above paradoxes are really of the same kind and whether the principle of uniform solution is a good methodological principle will require quite a lot of additional work and I don't think it is a good place to carry out this project here. In this paper, I will only focus on the pristine liar paradox and some less controversial variants of it. In particular, I will not work on Yablo's paradox and semantic paradoxes which involve no direct use of the concept of truth.

Another problem that this paper will not go into depth is the problem of the truth-bearer. Exactly which is the primary truth-bearer in virtue of which other entities can be true or false, sentence (sentence type or sentence token) or proposition (or statement, belief, etc.)? The former is a kind of linguistic entity while the latter is normally considered as what is expressed by the former. There is reason to favour taking proposition/statement/belief rather than linguistic entities like sentences to be the primary truth-bearer. For example, the very same sentence type (whether it is a sound or a writing) can have different sentence tokens and each sentence token may be used to express different propositions and thus have different truth value. The sentence type "I'm fine" may be true when uttered by Dean but may be false when uttered by somebody else. "The present king of France is bald" may be true when uttered by people who live in a period where there is a king of France and will be false or even fails to make a statement *at present*.[65] Thus in some early literatures about truth in the last century like Russell (1912), Austin (1950), and Strawson (1950) do not take sentences to be the primary truth-bearer or even do not take them as truth-bearers at all. But theories of the liar paradox, especially those who favour a formal method, would be more likely to take sentences as truth-bearer. One reason is for convenience. For the objects of formal method are always formal sentences, which can be precisely defined.[66] Another reason perhaps is that they do not think that this is a significant issue in solving the liar paradox. For the liar

[65] All these objections can be found in Strawson (1950), pp.325-326.
[66] See Kirkham (2001), chapter 2, section 2.5 for more advantages of taking sentences (tokens) to be truth bearers.

paradox can be generated whichever we choose as primary truth-bearer. For example, if truth is a property of propositions instead of sentences, then the problematic expression used to construct a liar sentence will be "express truth" instead of "is true" and their solutions on "is true" can then be replaced by a parallel study on "express truth".[67] Of course, there are some theorists who think that the key to solve the liar paradox exactly consists in choosing a correct truth-bearer and this issue will then be highly significant for them. In this dissertation I do not want to defense or reject any view on the choice of truth-bearer. Instead, given any theory of the liar I will simply take its choice for granted and assess it by the criteria that have been discussed above in this chapter.

Finally, about technical details. The study on the liar paradox has been carried out by many formal methods, and the discussion of those technical details is beyond my current capacity. So instead of digging into the technical details, in most of the cases I will simply point out the significance and provide relevant references where they are fully worked out.

5. Summary

The Liar paradox is understood in this thesis as a reasoning or argument which is generated by liar-like sentences, where liar-like sentences are sentences which are equivalent to some sentences saying of them (the liar-like sentences) as being not true. Problems posed by the liar paradox that are taken to be fundamental in this thesis are the triviality problem, characterization problem and revenge problem. Any solution to the liar paradox should solve every one of the above problems, if it is to be adequate. If a solution rejects the liar paradox (as a valid and sound argument), then it should solve the diagnostic problem, that is, it should tell us exactly which premise is untrue or which rule of inference is invalid in the argument and what's more important, it should tell us why. If a solution accepts the conclusion,

[67] This kind of idea can be found in Gupta (1982), p.4.

then it should tell us exactly how such a language can be non-trivial or at least provide insight as to how to live with inconsistency. Moreover, if a solution is a descriptive one, then it should justify its descriptive adequacy; if it is a normative one, then it should argue for the claim that natural language is inconsistent and that the new language obtained (either by limiting relevant notions or replacing them with new ones) is consistent or not trivial. All in all, the task in solving the liar paradox is to give a precise account of how our language (real language) can enjoy certain desired features (of syntax, semantics and logic) without falling into triviality or having any practical trouble. The account is often given by the technical tool, language model. Finally, this paper will mainly focus on the pristine liar sentence and the corresponding revenge to each particular theory. Yablo's paradox and other semantical paradoxes which involve no concept of truth will not be in our discussion. As to the issue of the truth-bearer, this dissertation does not take a position on it.

All these positions will be held throughout this dissertation without further justification. They may be debatable or controversial, but they serve as the basic points or big premises for this dissertation.

Chapter 3 Some Prominent approaches

1. Tarskian hierarchical approach

In Tarski (1933), Tarski develops a method for defining the notion of truth. He introduces the method by giving a concrete definition for certain formalized languages. The method can clearly be applied to many other formalized languages without significant changes. My primary interest in this dissertation is not Tarski's definition, but a resulting solution to the liar paradox, which I will refer to as *Tarskian hierarchical solution*. Nevertheless, with a proper understanding of Tarski's definition of truth, it is much easier to understand the hierarchical solution and also the limitation of it.

In this section I will first provide a simplified (and perhaps over simplified in some aspect) version of a Tarskian definition of truth. And second, I will introduce the main idea of a resulting Tarskian hierarchical solution to the liar paradox. Finally we will see some of the common objections to it.

1.1 Tarskian definition of truth

It is widely recognized that in defining the notion of truth, Tarski has multiple goals to achieve. Here I identify three of them. First of all, his project aims at the *semantic conception of truth*, which, at least, in Tarski's own idea, is captured by the classic Aristotelian conception and is a "correspondence" conception of truth (Tarski, 1944, 342).[68] The task of his definition thus is to make precise this old conception.[69]

---

[68] I cannot think of a better word for this. By "correspondence conception", I mean the conception that a correspondence theory of truth reveals. In Tarski's own word, " 'true' signifies the same as 'correspondence with reality" Tarski (1936, 401).

[69] Though Tarski thinks he is just making a conception more precise, it is quite controversial whether his theory is really a correspondence theory of truth. Haack (1976), Blackburn (2018), chapter 5, and Black Max (1948) argue that Tarski's theory is not a correspondence theory for actually it is neutral to most theories on the nature of truth in the sense that those theories can accept Tarski's convention T with indifference—they are all compatible with the set of T-sentences. What they disagree with each other is in virtue of what, a sentence is true and this is what a Tarskian definition does not tell us. Kirkham (2001) also agrees that Tarski's theory (at least the one Tarski ends up with) is not a correspondence theory of truth, but for different reason. See Kirkham (2001), chapter 5. Actually, Kirkham denies that convention T is compatible to other theories on the nature of truth. For more on this, see Kirkham (2001), p.172

Second, he wants his definition of truth to be immune from the threat of semantic paradoxes. For example, the liar paradox. However, in Tarski's own view, natural language is hopelessly threatened by the liar paradox, so the truth predicate he intends to define is not for natural languages, but for certain formalized languages, languages that are sufficient for presenting scientific theories (Tarski, 1969, 68). Third, he wants to define the truth predicate in a way that will respect the requirement of what is called physicalism.[70] Loosely speaking, physicalism is a position that no term other than those that are or can be reduced to physical, mathematical and logical terms, is legitimate.[71] A definition of truth will respect physicalism only if it makes use of terms only from physics, mathematics and logic or terms that can eventually be reduced to terms belonging to these subjects.

To accomplish the first goal, Tarski introduces *convention T*. It is a necessary condition that any definition of truth should meet in order to be adequate. To accomplish the second goal, Tarski distinguishes strictly the language about which we talk from the language in which we talk. The truth predicate for the language we are talking about, is defined only in the language we are talking in, which is *richer* than the former. To meet the physicalist requirement, Tarski removes any semantic terms from the language we are talking in, and so if the definition of truth can be given in such a language, it surely will not make use of any semantic terms that cannot be first reduced to non-semantical ones.

---

and pp.184-5.

[70] Tarski does not express explicitly that his project aims at a physicalist definition of truth. My statement on this point is based on the following reasons: First, some textual clues show that Tarski does have an intension to define the notion of truth in a way that would respect physicalism. See Tarski, (1936), p.406. Second, that Tarski has such a goal is commonly accepted, although there may not be consensus on exactly what kind of physicalism Tarski wishes to respect. See Kirkham (2001), chapter 5, Soames (1984), pp.400-401, and Lynch (2001), introduction to chapter V, for descriptions of Tasrki's general goals and his physicalist intention. Third, Tarski (1933) states, as a goal of his project, that in defining the notion of truth, he will not use any semantic term unless it can be first reduced to non-semantic terms (Tarski, 1933, 153). With all these, I think it would be reasonable to assume, in our discussion, that Tarski is indeed working on a physicalist definition of truth.

[71] See Kirkham (2001), section 5.1 for a statement of this version of physicalism. See Neurath (1931, 53-54) for an original formulation of physicalism.

1.1.1 Semantic conception of truth[72]

In ordinary discourse, the truth predicate is ambiguous. Sometimes we use it as a pragmatic term, say, to *endorse* some statement by saying "that is true". Sometimes we use it to express the notion of genuineness. For example, we say that someone is a *true* friend. The notion of truth that Tarski wishes to define has the *semantic conception* which is captured by Aristotle's famous statement: "To say of what is that it is not, or what is not that it is, is false, while to say of what is that it is, and of what is not that it is not, is true." (Aristote: $1011^{b}25$) Modern correspondence theories of truth have some definitions which aim at making this old semantic conception precise. Tarski himself provides some examples: "The truth of a sentence consists in its agreement with (or correspondence to) reality." (Tarski, 1944, 343)

Or, "(1) *a true sentence is one which says that the state of affairs is so and so, and the state of affairs indeed is so and so*." (Tarski, 1933,155)

Tarski calls the kind of definition of truth which bears the form like (1), *semantical definition*. He does not have any *essential* objection to these definitions. He thought that the meanings of these characterizations are intuitively clear. But still, they are unsatisfactory, because "From the point of view of formal correctness, clarity, and freedom from ambiguity of the expression occurring in it, the above formation obviously leaves much to be desired" (Tarski, 1933, 155).[73] The task of a semantical definition is to make the above conception (semantical conception) of

[72]In Tarski's terminology, semantics is a discipline dealing with notions about relations between linguistic entities and objects in the world (individual objects, states of affairs, properties, etc.), in contrast to pragmatics, which deals with notions about relations between linguistic entities and actions and to syntactics, which deals with notions about relations among linguistic entities. See Tarski (1944), section 5 and Tarski (1936), p.401. The conception of truth revealed in Aristotle's statement is a familiar relation between sentences and states of affairs, which is often characterized as a kind of *correspondence* relation. And that is the reason why the above conception is *semantical*.

[73] Ironically, I found that Tarski's usage of these terms "formal correctness", "clarity", and "ambiguity", to be themselves very unclear and imprecise, since he never provides some concrete definition of these terms. It is not clear exactly what kind of definition will or will not overcome the problem of being formally incorrect, unclear and ambiguous. But perhaps, we can only appreciate the impreciseness of the above definition of truth when we appreciate the preciseness of Tarski's definition.

truth more precise and accurate.[74]

1.1.2 Convention T

A starting point is to look at some particular occurrence of the truth predicate. If we ask in what situation the sentence "snow is white" is true, then a natural answer is that, it is true if and only if snow is white. In other words, we have the following equivalence:

"Snow is white" is true *if and only if* snow is white.

Equivalences of this kind are normally called T-biconditionals or T-sentences. In Tarski's view, each T-biconditional can be regarded as a partial definition of the notion of truth. The above equivalence explains in a precise way, what it means to say that the sentence "snow is white" is true. Similarly, if we replace the sentence "snow is white" with some other sentences, we will have more partial definitions of the truth predicate, each one of which determines partially the correct usage of the predicate. In general, every instance of the following schema, which is normally called *T-schema* or *schema T*, in Tarski's idea, is a partial definition of the truth predicate:

(T) *X* is true if and only if *p*,[75]

where X is the name (of some sort) of the sentence at issue, and p is the sentence (or, as we will see, it can be a sentence that has the same meaning as the one named

---

[74] Tarski states this goal explicitly in at least two places. See Tarski (1944), p.343 and Tarski (1969), p.64.

[75] There is a question regarding how to read the "if and only if" in Tarski's T-schema. This term can be used to express certain equivalence relations, but there are different types of equivalence relations and Tarski does not specify explicitly which kind of equivalence relation he refers to when using this term. Tarski (1944) tells us that his project "aims to catch hold of the actual meaning of an old notion" (Tarski, 1944, 341). See also Tasrki,(1944) p.351. But the term "meaning" is ambiguous as well. It can refer to, say, Fregean sense or Fregean referent (intension or extension), and so it is not clear whether the "if and only if" expresses an intensional equivalence or extensional equivalence. On the other hand, it would be implausible to interpret Tarski as trying to provide an intensional analysis to the notion of truth, because otherwise, he would not reject that the definition of truth can be given by a definition of provability *simply based on the fact that the extensions of the two notions are not identical* (see also Tarski, 1933, 237-238). Following Davidson (1990, 294-5), Kirkham (2001), section 1.8, Grayling (1990, 159), and Coffa (1991, 293-6), I will interpret this "if and only if" as expressing an extensional equivalence relation, that is, each such biconditional states the necessary and sufficient condition for the sentence named on the left-hand side of the biconditional to be true in the actual world.

by X).

Since each T-biconditional explains in a precise way the usage of the truth predicate relative to a certain sentence, presumably all T-biconditionals would offer a complete definition of the truth predicate. So Tarski requires that an adequate definition of truth, in addition to being formally correct,[76] should imply all T-biconditionals. This is the main point of what Tarski calls convention T (Tarski 1933, 187). Any definition of the notion of truth satisfying this condition (the condition that it implies all T-biconditionals) is called materially adequate.

The strategy Tarski adopts to define the truth predicate that would satisfy the above convention is very simple and direct—to form a logical product (conjunction) of all the T-biconditionals. Of course, the construction of such a logical product would not be as simple as it may first appear to be. But let us put the technical issue aside for a moment. There is a more serious problem in accepting convention T, which is the problem of paradox.

1.1.3 Distinction between metalanguage and object language

If the definition of truth should imply all T-biconditionals of the targeted language, then, if the language contains a liar sentence, like the following:

The first displayed sentence on this page is false

its corresponding T-biconditional would be the following:

---

[76] Tarski does not define precisely what he means by "formally correct", at least, not in his 1933. Commentators have slightly different explanations on this point. For example, in Gémez-Torrente (2019), a definition is formally correct if it makes use of no suspicious vocabulary and follows non-controversial rules for definition. Sher (2006, 147) takes it to be a condition that requires a definition to be free of paradoxes, in addition to the normal procedure for formally correct definition. Taylor (1998, 127) takes the condition of being consistent and non-circular to be the minimal requirement of the condition of being formally correct. Glüer (2011) explains Tarski's formal correctness in terms of explicit definition. A definition is formally correct if it is explicitly defined. A definition is explicitly defined if the term defined does not appear in the definiens. See also Hodges (2018) and Burgess and Burgess (2011), section 2.3 for the notion of explicit definition. In any explanation above, the condition of being formally correct is more like a general requirement of the methodology of definition, rather than some unique or special requirement imposed by an adequate theory of truth. So much for a remark on the formal correctness condition.

The first displayed sentence on this page is true if and only if *the first displayed sentence on this page is false*

which, under classical logic, immediately gives us the contradiction that the first displayed sentence on this page is both true and false. In other word, we can infer, from the definition of truth (with certain assumptions) a pair of contradictory sentences.

In general, if the targeted language for which the truth predicate is defined satisfies the following conditions:

a. It contains, for each of its expressions, a name denoting the expression
b. It contains a semantic term "true" for its sentences
c. All T-biconditionals are assertable (that is, being true, or being provable) in the language

and if classical logic holds for the language, then the language must be inconsistent in the sense that the above paradox is inevitable. Languages satisfying the above conditions (conditions a, b, and c) are called semantically closed languages. And let us call any language to which classical logic holds, *classical language*. Then the above conclusion can be simply put as such:

No semantically closed language can be classical without being inconsistent.[77]

[77] Gupta (1982) shows that this conclusion is not strictly correct. He shows that languages satisfying the following conditions can be classical and semantically closed: first, the language contains quotation-mark names naming the corresponding sentences quoted; second, no name, other than quotation-mark names, refers to sentences. Third, no predicate in the language applies to some sentences but not the others. Fourth, no function whose range is a subset of the set of sentences is in the language. And if the domain of a function is a subset of the sentences, then any input of the function will have the same output. Under these conditions, the only constants in the language that can distinguish among sentences (applying to some sentences but not some others) are quotation-mark names. The upshot of the construction is that, languages satisfying these conditions will be expressively limited. They do not contain enough syntactic recourses for the occurrence of the liar sentences. This kind of languages can be extended to languages with their own truth predicates and preserve classical features and semantical closure. For illustration, it would be sufficient to point out that these conditions guarantee that the only way to refer to a sentence in such a language is by quoting the sentence, and so it is impossible to construct a self-referential sentence like the simple liar sentence (L: L is false). For a complete and technical proof for the claim that languages of the above kind can be both classical and semantically closed, see Gupta (1982), section II. I should also mention that this rejection to Tarski's claim does not by itself constitute a solution to the liar. For it avoids paradoxes only by means of limiting the expressibility of the language, but surely natural languages are far

This conclusion has an important implication. Note that natural languages are universal, in the sense that "it would not be in harmony with the spirit of this language if in some other language a word occurred which could not be translated into it; it could be claimed that 'if we can speak meaningfully about anything at all, we can also speak about it in colloquial language [natural language]'." (Tarski, 1933, 164) So they must be semantically closed. The above conclusion thus implies that natural languages are inconsistent (assuming that classical logic holds for them, of course).

Tarski thus gives up the attempt to define truth predicates for natural languages, and for any language that is semantically closed and chooses to work entirely on formalized languages which are adequate for presenting scientific theories like logic and mathematics.

But if we are to deal a semantically open language (language that is not semantically closed), then given that the language does not contain enough recourses that are needed for describing its own syntax and semantics, the definition would have to be constructed in a language that is different from the target language. This leads us to Tarski's distinction between meta-language and object-language—the language we are talking *in* and the language we are taking *about*. To fulfill its function, the meta-language must contain: first, expressions that are identical to or have the same meaning as the expressions in the object-language; second, names for expressions in the object-language; third, terms that refer to the syntactical relations among sentences in the object-language and also terms from logic. The upshot here is that, for every expression in the object-language, there must be a translation of it as well as a name denoting it in the meta-language and *not vice versa*. This condition

---

richer in expressibility than languages satisfying the above three conditions. The current conclusion that one can draw from the discovery of the above kind of languages is only that some languages with limited syntactic resources can be both classical and semantically closed. The problem is, how rich a language could be, if it wants to preserve both classical features and semantical closure. Gupta admits that he has no answer to this question (Gupta, 1982, 16). So much for Gupta's discovery.

enables the meta-language to be able to talk about the object-language and contain enough syntactic recourses for constructing an adequate definition of truth for the object-language.[78]

Now that we have distinguished two kinds of languages, the T-biconditionals *for* object-languages *in* meta-languages will have a more precise description.

Let us call a sentence, T-biconditional, for an object language *L*, if it has the following form:

X is true if and only if *p*,

where *X* represents a name in the meta-language for a sentence in the object-language, and *p* is a *translation* of that sentence in the meta-language.

The trouble of the liar paradox then does not even arise. On the one hand, it does not bother the object-language, since the troublesome liar sentence cannot be constructed in the object-language at all; on the other hand, it does not bother the meta-language either, for although the liar sentence can be constructed in it, the truth predicate defined in it does not apply to its own sentences, and so the liar sentence will be simply true (or not true, depending on how one reads the negation).

1.2 A simplified Tarskian definition of truth

For simplicity and illustration, I will provide a simplified version of a Tarskian definition of truth for a relatively small formal language. Further, for our purpose,

---

[78] Tarski refers to this difference between the meta-language and the object-language as "essential richness". That is, the meta-language must be essentially richer than the object-language. However, his usage of this term turns out to be somehow ambiguous. In his 1969, he seems to take the "essential richness" condition as simply that the meta-language cannot "coincide or be translatable" into the object language, like we did above. See also the interpretation by Glüer in Glüer (2011), p40. Tarski (1944) specifies the condition as such that the meta-language contains variables of higher order than those in the object-language. Roughly speaking, the order of a variable is determined by the sorts of linguistic entities the variable represents. For example, we may stipulate that variables representing names of individuals to be of the 1st order and those representing names of classes of individuals to be of the 2nd order (in this case, a second order variable can occupy the predicate-place in a sentence) and those representing names of classes of classes of individuals to be of the 3rd order and so on. See Tarski (1969), p68 and Tarski (1944), section 10.

there is no need to specify the meta-language a priori. Any term that we used in defining the truth predicate for the object-language will be automatically admitted as part of the meta-language, if without further indication.

The object-language consists of normal logical connectives for first-order language "$\wedge$", "$\vee$", "$\sim$", where "$\wedge$" represents conjunction, "$\vee$" represents disjunction and "$\sim$" represents negation. It also contains a universal quantifier "$\forall$" and one objectual variable "x" (for simplicity). The set of non-logical constants includes only a one-place predicate sign "**R**" which represents the one-place relation *R*.

The definition of well-formed formula (grammatically correct formula, or sentential function) is the usual one for first-order language.

A Tarskian definition of truth is as follows:
First we define the notion of satisfaction.

An object *a satisfies* a formula θ if and only if
(1) If θ= "**R**(x)", then *a* is *R*;
(2) If θ= "$\sim\tau$", then *a* does not satisfy "τ";
(3) If θ= "$\tau\wedge\nu$", then *a* satisfies formula "τ" and a satisfies "ν";
(4) If θ= "$\tau\vee\nu$", then *a* satisfies "τ" or *a* satisfies "ν";
(5) If θ= "$\forall$(x) τ", then for any object *u* in the domain, *u* satisfies "τ".

What about a closed sentence which has no free variables (variables which are not bounded by any quantifier)? In the present language, since there is no name, but only the variable x and a one-place predicate, the only sort of sentence is of the form of universal quantifications or a compound of universal quantifications. So the satisfaction of them comes down eventually to clause (5), when "τ" is "**R**(x)".

In this case, an object *a* satisfies "∀(x)**R**(x)" if and only if every object satisfies it.[79]

Proof:

Suppose, *for reductio*, that *a* satisfies "∀(x)**R**(x)" but there is an object *b* in the domain which does not satisfy the formula "∀(x)**R**(x)".

By clause (5), *a* satisfies "∀(x)**R**(x)" if and only if *all* objects in the domain satisfy **R**(x).

If *b* does not satisfy "∀(x)**R**(x)", then that means there is an object *c* (which may be anything) in the domain that does not satisfy "**R**(x)", which is impossible.

So it is not the case that *a* satisfies the formula "∀(x)**R**(x)" but there is an object *b* in the domain which does not satisfy the formula "∀(x)**R**(x)".

So the notion of truth is defined as those sentences (closed formula, formula with no free variables) that are satisfied by all objects.

The definition of truth predicate:

D_t: For any *s,* if *s* is a sentence, then *s* is true if and only if s is satisfied by all objects.

Now let us examine whether the above definition satisfies Tarski's convention T.

First, is it formally correct? Well, the definition of the truth predicate for the language appears to be formally correct, but the definition of satisfaction is clearly not. It is a recursive definition and the definiendum appears in the definiens. But

---

[79] If ∀xτ, where τ itself is a closed sentence, for example, it is another universal quantification, "∀(y)**R**(y)", then we can still apply clause (5).
And as to language containing names, like **a, b…**
We need then one clause for the denotation of each name and also one clause to determine the satisfaction condition for each closed sentence formed by using names and predicates.
For example, an object a satisfies "**R**(**a**)" if and only if a is R.

this is not a big problem. For it is said that any recursive definition can be transformed into an explicit definition.[80] Here is how to transform it.

First we introduce a second-order variable which ranges over two-place relations, let it be X. Then we replace each occurrence of the notion under defined (in this case, satisfaction) in the above recursive definition with this second-order variable. What we obtain is an one-place property, let it be $P(X)$:[81]

Now the explicit definition for satisfaction that is equivalent to the above definition can be reconstructed in this way:
An object *a* satisfies a formula θ if and only if for every X, if $P(X)$ then $X(a, \theta)$.[82]

So a complete explicit definition using this second-order variable would be as follows:

An object *a satisfies* a formula θ if and only if $X(a, \theta)$, for any X where the following conditions hold:
(1) If θ= "**R**(x)", then *a* is *R*;
(2) If θ= "~τ", then "X(*a,* "τ")" does not hold;
(3) If θ= "τ∧ν", then X(*a,* "τ") and X(a, "ν");
(4) If θ= "τ∨ν", then X(*a,* "τ") or X(*a*, "ν");
(5) If θ= "∀(x) τ", then for any object *u* in the domain, X(*u*, "τ").

Second, does the definition entails all T-biconditionals for the language?
This seems to be plausible, here is an example (for simplicity, we will be using the recursive definition below rather than the explicit one).

---

[80] This discovery is credited to Frege and Dedekind, and also Whitehead and Russell by Tarski himself (Tarski, 1933, 176, fn1). See Dedekind (1923), pp.33-40; and Whitehead and Russell (1925), pp.550-7 and p.244.
[81] For Tarski's original transformation see Tarski (1933), p177. For a short introduction of this method, see Beall, Glanzberg, and Ripley (2018), p.67.
[82] "$X(a, \theta)$" means that the object a and the formula θ are in the relation X.

Consider the sentence "∀x(**R**(x)∨~**R**(x))". The T-bicondition for it is:

"∀x(**R**(x)∨~**R**(x))" is true if and only if for every object x, either it is *R* or it is not *R*. Here is how we get it.

By the definition of truth we have:

a. "∀x(**R**(x)∨~**R**(x))" is true if and only if "∀x(**R**(x)∨~**R**(x))" is satisfied by all objects;

By the definition of satisfaction we have:

b. "∀x(**R**(x)∨~**R**(x))" is satisfied by all objects if and only if "**R**(x)∨~**R**(x)" is satisfied by all objects;

By clause (4) in the definition of satisfaction we have:

c. an object satisfies "**R**(x)∨~**R**(x)" if and only if it satisfies "**R**(x)" or it satisfies "~**R**(x)"

By clause (2) we have:

d. an object satisfies "**R**(x)∨~**R**(x)" if and only if it satisfies "**R**(x)" or it does not satisfies "**R**(x)"

By clause (1) and d we have:

e. An object satisfies "**R**(x)∨~**R**(x)" if and only if it is *R* or it is not *R*.

By clause a, b, e we have:

"∀x(**R**(x)∨~**R**(x))" is true if and only if every object x is either *R* or not *R*.

1.3 Tarskian hierarchical solution to the liar paradox

Does the above definition of truth avoid semantic paradox? The answer is yes, but it avoids the paradox only because the object-language is expressively limited. The above method can actually apply to quite a lot of other formalized languages, with more logical operators, more non-logical constants like names, predicates, and function symbols. But there is one important thing that the object-language at issue must obey, that is, the object-language must not contain the truth predicate being defined in the meta-language. As has been mentioned in section 1.1, the problem of liar paradox does not even arise.

What about natural language? It is necessary to point out the difference between Tarski's method to avoid semantic paradoxes and applying Tarski's hierarchical definition of truth to solve the liar paradox for natural language. The former is just a technical issue, which concerns how to construct an artificial language in which no paradox occurs, while the latter involves many additional philosophical claims about natural language. In particular, if one wishes to apply Tarski's theory to a natural language, say, English, then presumably the language or at least, our conception of it, must undergo some sort of reform. In particular, it must be or be regarded as having a hierarchical structure. That is, English must be split into a series of languages (L_1, L_2, L_3…). Each language in this sequence serves as a meta-language for the previous language. This means that in addition to names for expressions in the previous language and translations of those expressions, the language at any level, must also contain a truth predicate, which is defined in the language at the current level and applies solely to previous language.

The above condition is equal to say that English must contain not a single, univocal truth predicate (or 'true *simpliciter*'), but a series of truth predicates (T_1, T_2, T_3,…), each one of which has a richer extension than the previous one, or, if there is only one symbol for the truth predicate, it must be regarded as an ambiguous term, expressing multiple different notions of truth practically. Correspondingly, the T-schema must also not be treated as a single schema, but multiple schemas constructed out of the above series of truth predicates respectively.[83]

For T_1, which is a truth predicate defined in L_2 for L_1, we have:

T_1-schema: X is T_1 if and only if *p*,

where X is a name (in L_2) of a sentence in L_1 and *p* is its corresponding translation in L_2.

[83] It is easy to see that a Tarskian hierarchical solution to the liar paradox for natural language should be of this form. One can find relevant description in Priest (2006), chapter 1, section 1.5.

In general, for language L_n, we have the T_{n+1}-schema:

X is T_{n+1} if and only if p[84],

where X is a name (in L_{n+1}) of a sentence in L_n and p is its corresponding translation in L_{n+1}.

Under such a reform, at least, each language in English is now able to avoid the liar paradox. Suppose we have a language, L_i, then it will contains no truth predicate for its own sentence, but a truth predicate T_i for L_{i-1}.

If the language has no syntax resources to talk about its own sentences, then self-referential sentences cannot be constructed at all. If the language has resources to talk about its own sentences, then the liar-like sentence is at best, the following one:

S: S is not T_i

S is a sentence in L_i so we apply the corresponding T_{i+1}-schema

S is T_{i+1} if and only if *S is not* T_i.

If being not T_i means being anything other than T_i, then it is the case that S is not T_i. So S is T_{i+1}. If being not T_i means being $false_i$, then it is not the case that S is not T_i for S is not a sentence at $level_{i-1}$, it is neither T_i nor $false_i$. So S is not T_{i+1}.

In summary, under the Tarskian solution, the liar sentence will be either ungrammatical, that is, it is not expressible in the language at all or it will receive a truth value, but only in a higher level of language.

1.4 Critique

It should be clarified first that critiques to Tarski's theory may be of two different sorts. The first sort criticizes Tarskian definition of truth[85] while the second sort

---

[84] It may be held that in this case, the T-schema has an antecedent: If X is in L_n then X is T_{n+1} if and only if p.

[85] Here I mention two sorts of critiques which are well-known but are less relevant to our topic here. The first kind of argument is based on relativity. Tarski's definition is not for the concept of truth in general, but for many concepts of truth which are relative to some languages. So the definition of truth in German will be totally different from the definition of truth in English. See Blackburn (1984), pp.266-267. The second sort of the objection is that Tarski's definition is trivial in some sense. The definition is a list account, it gives a definition by enumeration

criticizes the Tarskian hierarchical solution. It is the second sort of objections that are the principle concerns of this dissertation. Nevertheless, the critique to Tarskian definition of truth may have some impact on the Tarskian hierarchical solution. This is because the Tarskian hierarchical solution, when viewed as a descriptive project, involves some substantial claims about our language and the concept of truth.

A. Difficulty in level-assignment

One influential critique comes from Kripke (1975). The main idea is that the hierarchical picture is factually wrong (Kripke, 1975, 695). While there is no reason to accept the hierarchical reform, there are reasons to reject it.

In addition to the problem due to the existence of empirical paradoxical sentences, the hierarchical picture actually produces some practical difficulty. If the hierarchical picture is right, then the truth predicate must be taken to be ambiguous, and people when using this predicate will have to, either explicitly or implicitly, attach a subscript (in any form) to the truth predicate they use, and do so without violating the basic principle of the hierarchical picture—truth predicates with lower subscripts cannot be applied to sentences in higher levels.

Kripke points out that no one actually attaches a subscript to the truth predicate, either explicitly or implicitly, for there are situations where they just have no ability to do that. The first sort of situation is where blind truth ascription is involved. For example, it is quite common for one to make a comment like "Everything John says is true" without actually knowing each utterance John makes. It seems that if the hierarchical restriction is imposed, then the speaker will have to decide which

---

rather than by enclosing the underlying property that would explain relevant phenomenon. One famous argument comes from Field (1972). He gives an analogous definition by enumeration for a chemical notion valence Field (1972, 380):

> For any x, x has valence n if and only if:
> if x is potassium then n=+1…if x is sulfur then n=-2…

This definition is informative for it tells us which chemical element has which index as their valence, but it does not tell us the nature of the valence of an element and so does not tell us what it is for an element to have valence n. Similar arguments can be found in Black (1948).

predicate he or she is going to use before he or she can talk meaningfully by the above comment (Kripke, 1975, 695). The defender may reply that there is a difference between a strict analysis to the language and the language we use practically. In everyday life, we always speak ungrammatically, but that does not mean that the correct grammar does not exist. The hierarchical picture provides a correct description of the semantic grammar of our language, but that does not mean that we cannot violate it practically. This may be the case, but another sort of difficulty seems to show that there is some intrinsic problem with this picture. Consider the following two sentences:

Dean: Everything Nixon says about Watergate is false.

Nixon: Everything Dean says about Watergate is false.

The trouble is that there seems to be no acceptable level that can be consistently assigned to both sentences. For while Dean needs to choose a level for the truth predicate that is higher than those for all the truth predicates Nixon uses, Nixon should also do the same, which is impossible. But it seems quite intuitive that in some situations it is very easy and intuitive to say that both sentences are false. For example, when Dean and Nixon have said at least one thing right about Watergate (Kripke, 1975, 696).

Now some may argue that the above examples are indeed problematic, *exactly because the hierarchical picture is right*. Here we come back to the most difficult philosophical problem. Currently I do not have an idea how to response to this position, but one safe conclusion we can draw from the above argument is that the hierarchical picture discussed so far fails to capture some of our intuitions about the semantics of natural language.

B. Revenge

Priest (2006) provides a different revenge sentence for Tarskian hierarchical solution. He defines the rank of a sentence as the level of the lowest language of which a sentence is a member (Priest, 2006, 19). For example, if a sentence is at

level 1, then its rank is 1.

If "rank" is allowed, then we can further form a complex notion, "being true at its rank". Formally, Priest uses "$T_{rk(x)}$" to represent the notion "being true at its rank", where T represents the truth predicate, rk(x) represents a function from a sentence to its rank, $T_{rk(x)}$ thus represents, given any sentence x, the notion of being true at x's rank. Now we can construct a sentence saying of itself as being not true at its rank:

a: $\sim T_{rk(a)}$ **a**

And we can follow the following steps to get a contradiction:

(1) **a**: $\sim T_{rk(a)}$ **a**

(2) Applying the T-schema for rank i, we have:

T_i**a** if and only if $\sim T_{rk(a)}$ **a**

(3) Since rk(a)=i, we have:

T_i**a** if and only if $\sim T_i$ **a**,

which is a contradiction.

I find that Priest's analysis may have some problems.[86] First, in step (2) it seems that the correct T-schema for rank i is not "T_i*(x) if and only if x*", but should be "T_{i+1}*(x) if and only if x*". Since **a** is at level i, the correct T-schema for it should apply a truth predicate at higher level, not the same level. Correspondingly, what we have is "T_{i+1}**a** if and only if $\sim T_i$ **a**", and there seems to be no problem in this sentence. Second, there seems to be a confusion that is covered by the above formalization. It is one thing to apply the truth predicate which belongs to level i to a sentence. It is another thing to call the sentence true at its rank, where its rank is i. The difference may be explicated by the following fact about the hierarchical theory: Each sentence at level i, is either a $true_{i+1}$ sentence or a $false_{i+1}$ sentence at

[86] I do not want to insist on this point. If Priest's analysis is good, then the Tarskian hierarchical approach does face a revenge problem. If Priest's analysis fails, then my further analysis shows that the Tarskian hierarchical approach still has a revenge problem anyway. In any case, I think I have shown enough that the approach does have a revenge problem.

rank i.

In other words, any sentence is true at its rank, but by saying so, we are not yet specifying in what sense it is true (that is, which truth predicate is at play). If this distinction is correct, then a sentence like **a** above is incomplete, for it does not specify exactly which truth predicate is in use.

Suppose the truth predicate in use is of level i, then we can formalize **a** in this way:

a: $\sim T_{i,\, rk(a)}$**a**

Now, since **a** is a sentence at level i, we can put i for rk(a) and so we have:

a: $\sim T_{i,\, i}$ **a**

What the sentence says is something like, this sentence is not $true_i$ at rank i. If read the sentence as saying that **a** is anything but $true_i$ at rank i, then **a** is $true_{i+1}$; if read the sentence as saying that **a** is $false_i$ at rank i, then the sentence is $false_{i+1}$. The reason for both results is that no sentence at rank i is either $true_i$ or $false_i$.

The above revenge may fail, but if Priest's second pattern for revenge paradox is correct,[87] then a new revenge sentence seems to be inevitable. Instead of talking about the rank of a sentence, one may quantify directly into the index of the truth predicate. So we may form a sentence like:

R: R is not $true_i$, for any i.

Or, we can define a concept, which is a logical sum of all truth predicates defined in the hierarchy:

For any sentence x, $T_s(x)=T_1(x) \vee T_2(x) \vee T_3(x)\ldots$

And the revenge sentence for the hierarchical theory is:

R': R' is not $T_s(x)$.

This is actually the type P revenge we have mentioned in chapter 1. Suppose the above sentences, either R, or R', are expressible in natural language, then according

[87] See chapter 1, section 2.

to the hierarchical picture, they should be somewhere in the hierarchy. Let it be either $true_j$ or $false_j$, for some level j. Applying the corresponding T_j-schema for R, we have:

R is $true_j$ if and only if R is not $true_i$, for any i.

But if R is not $true_i$, for any i, then R is not $true_j$, contradicting with our assumption. So R cannot be $true_j$. Since level j is arbitrary, we conclude that for any j, R cannot be $true_j$. But this is exactly what R says, so it must be $true_i$, for some i. A similar reasoning works for R'.

One way to get around the above problem is to argue that the relevant notions, like $T_s(x)$ or quantification into the levels of the hierarchy, is not intelligible or not coherent. I haven't seen any precise definition on these two terms, but it seems that there is no convincing reason to reject the legitimacy of the notion of $T_s(x)$ or quantification into levels. As Priest (2006, 20) points out, even to explain the hierarchical theory, these concepts are needed. For example, it seems that according to the theory itself, every sentence is T_s, that is, every sentence is at least true *in some sense*. Also, using quantification, we may be able to assert some general claims like "In any level i, every sentence is either $true_{i+1}$ or $false_{i+1}$". This statement is just the hierarchical version of the principle of bivalence (the original univocal version is, every sentence is either true or false). The hierarchal theorists may argue that we need not to apply quantification in order to express the above principle. For example, they may suggest that one can express the hierarchical version of the principle of bivalence by what is sometimes called typical ambiguity (Priest, 2006, 20). That is, we take the statement "Every sentence is either $true_x$ or $false_x$" to be typically ambiguous in the predicate place so that the indices attached to the truth and falsity predicates are not specified, and the whole sentence will be true no matter what the subscripts are. However, I think that having an alternative way to express something does not mean that the original way (using quantification) to express it is not legitimate. And moreover, as Priest (2006) questions, exactly what do we grasp by such a typically ambiguous assertion? It seems that the answer can

only be: "For what such typically ambiguous assertion means, what we are supposed to understand by it, is just what is expressed by a single sentence which quantifies universally over i [the variable representing subscripts of truth predicate]" (Priest, 2006, 20).

Of course, the defenders of the hierarchical theory may feel that the argument is question begging, for if their theory is correct, Priest's suggestion is simply wrong. But at the first place, it is the hierarchical theorists that try to convince us that we should give up some way of expression which we think to be highly natural. And so far we do not see any convincing justification for this, so the move seems to be ad hoc and it achieves its intended result only by limiting what we can express in natural language.

The conflict here can be summarized like this: The hierarchical theory suggests that our natural language is constituted by an infinite series of languages which form a hierarchical structure, each language at the higher level can talk about language at lower level, and all of our talk come at some level in the hierarchy. The trouble comes when we are trying to talk about the semantics of the *whole* hierarchy, either by forming a concept which is said to exhaust all levels, or by universally quantifying into levels. For if these talks are allowed, then they must appear at some level in the hierarchy, but they cannot, for then they will be talking about something at their own level (which includes themselves) or even higher levels. In the end, whichever level we posit the original talk, we find that it cannot appear at that level, and eventually, we exhaust all possible levels. In the end, we have to conclude, either we cannot talk about the hierarchy as a whole, or the talk of it is not in the hierarchy. In the first case, it seems to undermine our expressibility, in the latter case, the theory fails to give a semantic theory for natural language. Thus the *Dialetheist's Conjecture* that we mentioned in chapter one seems to hold here.

2. Kripke's paracomplete approach

After criticizing the Tarskian solution to the liar paradox, Kripke proposes his own

approach which is based on truth value gap. While most of the sentences in a language are either true or false, some sentences are neither. Among those that are neither are sentences which Kripke called ungrounded. Liar sentences are ungrounded sentences so they are neither true nor false (at least, in the sense that they do not have a classical truth value). One of the premises in the liar paradox thus is blocked and so no contradiction would follow.

Kripke's approach has two merits. First, from the technical side, he shows us that under certain three-valued semantic schema, some certain languages can indeed contain their own truth predicates, despite the fact that they are semantically closed (Tarski's theorem does not work here because the truth value gap is present). Second, from the philosophical side, he defines a semantic phenomenon—ungroundedness, in a precise way, which does capture some of our intuition about the difference between sentence like "snow is white" and sentences like "This sentence is not true". Below I will first introduce these two aspects and after that we will discuss the revenge problem for Kripke's solution.

2.1 Three-valued semantics and jump operator[88]

In the methodology chapter we have said that a language can be formally represented by a triple L=<L, M, σ>, where L provides syntax, M provides interpretation of the syntax, and σ is semantic schema which gives us rules for determining the truth value of compound sentences from the truth value of the sub-sentences that constitute it.

A classical semantics for a language is a semantics in which M makes use of the classical truth values {true, false} in interpreting relevant constants in L and that σ actually defines a truth-function from {true, false} to {true, false}. A three-valued

[88] The kind of construction of a three-valued language model in this section (section 2.1) is quite standard. For Kripke's own construction (which is very simplified), see Kripke (1975), p.700 and p702. For a more detailed construction, see Beall (2007), pp. 23-26.

semantics, at least in this dissertation, means that the truth-value set is changed from {true, false} to a triple {true, false, n}, where n represents a third truth value, in both M and σ.[89]

Kripke's construction is based on a three-valued semantics which is normally called strong Kleene semantics. The truth-value set is {true, false, undefined}, the semantic schema is normally denoted by the sign κ, and can be represented by the following rules, each one of which can be regarded as defining a truth function, function from truth values to truth values:[90]

Rules for (choice) negation:

~A is true if A is false, and is false if A is true, and is undefined if otherwise.

Rules for disjunction:

A∨B is true if at least one of them is true, and is false if both of them are false, and is undefined if otherwise.

Rules for quantification:

∀xA(x) is true if A(x) is true for all objects in the domain, and is false if A(x) is false for some object in the domain, and is undefined if otherwise;

∃xA(x) is true if A(x) is true for at least one object in the domain, and is false if it is false for all objects in the domain, and is undefined if otherwise.

I use L_g to represent a language that is exactly like L except that it contains an extra one-place predicate T and the language is interpreted by M+g, where M+g is a model exactly like M excepts that it interprets the predicate T with g.[91] Following Kripke, we will treat an interpretation on an one-place predicate as an ordered pair

---

[89] Kripke does not make a distinction between classical truth value and the value "undefined", for it seems that he does not take "undefined" to be another truth value at all. But for our purposes, adopting this terminology will be much more convenient.

[90] The notion of truth here and below is not in the object-language we are to work on. For our current purpose, it can be regarded as a notion of truth that is defined via Tarski's method (see chapter 3, section 1) for the object-language below.

[91] The language so presented is just a form or it represents a group of language. It tells us that every time we see a sign formed by an italic "*L*" and a subscript, it means that it is a language with an extra predicate sign T, and which is interpreted by an interpretation denoted by the subscript.

$\{S^+, S^-\}$, the members of which provide the extension and anti-extension[92] of the predicate being interpreted. So L_g=<L+T, M+g, κ>, where g= $\{S^+, S^-\}$.[93] And we assume that *L* contains no semantic predicate and that the language has enough syntactic resources to talk about its own syntax.

Our job is to see whether T can be interpreted as a truth predicate for L_g. But in what situation can we say that this can be done? One criterion is that, for every sentence p in the language, p and that "p" is T must have the same truth value. In other words, every time we can assert p, we can also assert that "p" is T, for any sentence. This feature is normally taken to be the characteristic feature of the truth predicate. So if an interpretation g for T gives us the extensional equivalence between p and the sentence "p" is T, for any sentence in the language, then we can say that T is the truth predicate for the language, at least, from an extensional perspective.

But then the question becomes how do we know whether there can be an interpretation that satisfies the above requirement? For this, Kripke introduces an operator, which is normally called the jump operator. We shall use "δ" to represent this function.

The jump operator can be roughly understood as an operator to take all true (false)[94]

---

[92] Extension: the set of objects of which the predicate is true; Anti-extension: the set of objects of which the predicate is false. A methodology note: here we use the notion of truth to define the notion of extension and anti-extension, some may think that this may involve some sort of circularity for it seems that our final aim is to define truth. But this is not the case. We already know what truth is and there can be a truth predicate defined in a strict way via Tarskian method in a meta-language for the object language we are to work on. Our current project is rather to find out whether the object-language can contain its own truth predicate—a predicate whose extension is the set of all true sentences of the language and which respects some characteristic feature of the truth predicate in the language.

[93] Here g=$\{S^+, S^-\}$ is also just a form. It means that an interpretation to a one-place predicate is an ordered pair.

[94] As mentioned, the notion of truth I use here is not in the object-language. For simplicity, I do not give a formal definition of it here. For our current purpose, the following guideline would be useful:

(1) An atomic sentence of the form "a is P" is true just in case the object denoted by the name a is in the extension of P and is false if it is in the anti-extension of P and is undefined if otherwise.

sentences in a given language interpreted by a model, and so its product is just another interpretation for the candidate truth predicate T, {S^{+}', S^{-}'}.

Formally, it is defined as a function from interpretations of one-place predicates to interpretations of one-place predicates according to the following rules:
For any interpretation g for the one-place predicate T:

$\delta(g) = g'$

where g'= {S^{+}', S^{-}'}, and S^{+}' is the set of all true sentences in L_g and S^{-}' is the set of all false sentences in L_g.
When δ reaches a fixed point, that is, when δ(g) = g, this means that everything that can be true or false in the language L_g has been recorded by the interpretation g (and δ(g)). Now given any sentence p, suppose that "p" is true in L_g, then "p" $\in S^{+}$'. Since δ(g) is the interpretation of T in L_g, the sentence "'p' is T" will also be true, since "p" is now in the extension of T. And since everything that is true or false has been recorded by the fixed point, it means that the sentence "'p' is T" is also in the extension of T. This holds for all sentences in the language. So a fixed point ensures the extensional equivalence between "p" and "'p' is T", for any sentence p.[95]

In short, given a language, if the jump operator defined for the language has a fixed point, then it is always possible to interpret an one-place predicate as the truth predicate for the language. The fixed point is one such interpretation.

2.2 Kripke's construction of a fixed point

Kripke also shows us how to construct a fixed point.[96] The technical details are less

(2) The truth status of a compound sentence is determined according to the semantic schema.
(3) The information of extension and anti-extension of relevant predicates and names are given by the ground model M for the language.

[95] There may be some problem for this quantification. I do not mean to be strictly correct in technical details. If one finds this an unacceptable problem, then one may simply take the statement to be schematic, that is, any sentences that can have truth value at all and is of the form p if and only if x is T, where p is the sentence and x is its name, will be true when the fixed point is reached.

[96] The construction below is for a fixed point which is normally called the minimal fixed point. Our description below and later are all based on the minimal fixed point, if without

important for our purpose so I will simply sketch some important steps.

Let us use L_g defined above as the object language. Let $g_0=\{S_0^+, S_0^-\}$ be the primary interpretation on the predicate T, where S_0^+, and S_0^- are both empty. At this stage, we can regard T as being totally undefined, nothing in its extension or its anti-extension. Let the language being interpreted by $M+g_0$ be L_{g0}.

To expand the interpretation, we apply the jump operator δ:

$\delta(g_0)= g_1$;

where $g_1=\{ S_1^+, S_1^-\}$.

According to the rule of the jump operator, S_1^+ will be the set of all true sentences in L_{g0}, and S_1^- will be the set of all false sentences in L_{g0}. Since T is totally undefined, there will be no sentence which involves the predicate T to be collected. We take g_1 to be the new interpretation of the predicate T, and so form a new language L_{g1}.

Now g_1 is not empty. S_1^+ and S_1^- contain those true and those false sentences in L_{g0} respectively. This increases the number of true sentences and false sentences in L_{g1} in contrast to L_{g0}. We apply once again the jump operator to collect those sentences and so form a new interpretation $g_2=\{S_2^+, S_2^-\}$ and a new language L_{g2}. The above process can go on into infinite stages, generating more and more interpretations for the predicate T, g_3, g_4,…g_n…and the corresponding languages L_{g3}, L_{g4}…L_{gn}…

There is one important property of the jump operator in the above process, which is normally called *monotonicity*.[97] Roughly speaking, this property ensures that

---

further indication.

[97] The "monotonicity" of a function means that the function preserves the order of relevant arguments. For example, the number function y=x+1 will preserve the order of the argument whatever we put into the variable x. This means that if $x_1 \leq x_2$, then the corresponding output, y_1 and y_2 will be in the same order relation. That is, $y_1 \leq y_2$. In the case of the jump operator here, Kripke chooses to define the partial order by the set-theoretical relation "is a subset of", which is normally represented by the symbol "⊆". Set A ≤ Set B if and only if Set A ⊆ Set B. And Set A ⊆ Set B if and only if every member belongs to A is a member of B. If the jump operator is

beginning with the empty interpretation g_0, the above process will only expand previous interpretations in the sense that sentences that are put into S_i^+ or into S_i^- for some stage i, they will never come out again. This means that once a sentence is evaluated as being true or false, at some stage, it will never change its truth value in later stages. The process only gets more and more previously undefined sentences to be defined (evaluated as being true or false). So by iteratively applying the jump operator, we keep expanding the interpretation of the predicate T.

Whether this process will have a fixed point eventually is purely a mathematical issue[98] which can be proved by using some set-theoretic apparatus, and I will not go into there. But the process does have a fixed point. Actually the conclusion can be more general. The jump operator will always have a fixed point, as long as the relevant three-valued semantic schema is monotonic.[99]

2.3 The truth status of the liar sentence in this process

Now in a fixed point, what is the truth status of the liar sentence? Notice that in the above process, once a sentence is evaluated as being true or false at some stage, it will never change its truth status in later stages. But how does a sentence get its truth value?

If a sentence is of the form:

---

monotonic, then that means, for example, if $g_i \subseteq g_j$, (in the sense that $S_i^+ \subseteq S_j^+$ and $S_i^- \subseteq S_j^-$), then $\delta(g_i) \subseteq \delta(g_j)$ (Kripke, 1975, 703).

[98] Note that the process will be carried into transfinite stages, the discussion of which is beyond the scope of this dissertation. For a complete construction, see Kripke (1975), p.704.

[99] The monotonic of a semantic schema can be defined in terms of some partial order relation among the truth set {t, f, u}. One of the partial relation is defined as follows:
$u \leq t$
$u \leq f$
it is not the case that either $f \leq t$ or $t \leq f$,
where $\leq$ is a partial order relation.
For a proof of the point that the jump operator defined based on any three-valued monotonic schema always has a fixed point on the space of interpretation of one-place predicates based on three-valued semantics, see Gupta and Belnap (1993), chapter 2.

"A" is T,[100]

then reflecting on the above process, we know whether it will have a truth value at the fixed point depends on whether the sentence "A" has a truth value in the process.

If A is still a sentence of the above form, then we keep tracing the source of its truth value by removing the predicate T, disquoting the sentence. Eventually, we reach a sentence which does not contain the predicate T, for example, a sentence of the following form:

a is P

then whether it will have a truth value depends on whether P is defined for the object denoted by the name a. That is, it depends on whether the extension of P or the anti-extension of P contains the object denoted by a.

What about sentences like the liar sentence? Note that in such a language, if there is a liar sentence, it will make use of the predicate T:

S: S is not T.

But the problem is, we cannot trace the source of the truth value of S in the same way we trace the above sentences. For in this case, the predicate T is indispensable. And so whether S has a truth value in this process depends on whether the extension or anti-extension of T contains S at the beginning of the interpretation process. But g_0 is empty. So a sentence like S never gets into the evaluation process, it will never be collected by the jump operator in any stage. So S does not have a classical truth value at the fixed point. Its truth status remains undefined throughout the whole process.

2.4 Ungroundedness

Normally, if we find a sentence of the following form:

A is true

[100] There can be sentences of other forms, like those involving logical connectives. For the sake of simplicity, I will only work on the simplest form.

then if we are to determine whether it is true, we will be led to determine whether A holds. Suppose that A is just another truth ascription, like:

B is true

Then following the same thought, we will need to determine whether B holds. Now eventually we will need to land on some ground where we need not appeal to the truth value of other sentences in other to determine the truth value of the current sentence.

For example, suppose the above tracing process eventually leads us to the sentence "Snow is white", then the truth value of the original sentence depends not on other sentence, but on the empirical world—whether snow is white in actual world.

In this sense it seems that we can say, the truth value of the first sentence, "A is true" is eventually grounded by the sentence "snow is white", or more directly, it is grounded by empirical fact.

Now consider the liar sentence:

S: S is false

The distinctive feature of S is that, every time we want to determine its truth value, we are not led to other sentence, but back to the very sentence itself. But the falsity predicate is not dispensable in the sentence, so we are entering into an endless loop, never landing on a ground that can determine the truth value of it, at least, not the same kind of ground we have for determining sentence like "'snow is white' is true".[101]

Kripke's construction actually reflects this determination process. Whether the truth value of a sentence can be determined at the fixed point depends eventually on whether we can trace back to a sentence whose truth value has a determinate ground. Based on this distinction, the notion of groundedness can be precisely defined via

[101] For Kripke's original description of this intuition, see Kripke (1975), pp.693-694.

the above interpretation process.

In particular, grounded sentences are defined as those that have a classical truth value at the fixed point that we have seen in section 2.2[102] and ungrounded sentences are defined as those that receive no classical truth value at the fixed point. (Kripke, 1975, 706)

One virtue of this approach is that, whether a sentence is ungrounded may not be something can be fixed a priori by its syntax or semantics. Where empirical liar sentence is involved, whether the liar sentence is ungrounded will be a matter that links up with empirical fact. Consider the following empirical sentence:

The first displayed sentence on this page is false.

If the sentence itself happens to be the first displayed sentence on this page, then it will receive no truth value in the above interpretation process, and so has no classical truth value at the fixed point.

On the other hand, if it happens to be not the first displayed sentence on this page, then which truth value the sentence may have will depend on the truth status of the first displayed sentence on this page. The problem of empirical paradoxical sentences that disrupts Tarskian hierarchical solution is resolved here.

2.5 Revenge

It is not at all clear whether Kripke would like to call ungrounded sentences neither true nor false, for it seems that he does not take "undefined" to be a third truth value (Kripke, 1975, fn18). So we will not use the strengthened liar sentence which makes use of the concept "neither true nor false" here. But it is still possible, and which I

[102] Note that there can be multiple fixed points. For example, the truth-teller, "This sentence is true", can be put into either the extension or the anti-extension of the predicate T at the initial stage, g_0. If it is in the extension/anti-extension, then it will always receive a classical truth value at later stages and keep it at the fixed point. But such a fixed point is different from what we have seen above, for its extension and anti-extension contain much more sentence than the above one.

think is recognized by Kripke, to construct a revenge sentence for his theory. In particular, one can make use of the notion of ungroundedness or the notion of exclusion negation to construct a revenge. Below I will introduce them in turn.[103]

2.5.1 Ungrounded revenge sentence

What if we have a sentence saying of itself as false or ungrounded? For example:

R_g: R_g is not T or ungrounded.

Kripke admits that the meta-language he uses is classical (Kripke, 1975, fn18), so presumably the ungrounded predicate would also be classically constructed in the sense that its extension and anti-extension are mutually exclusive and jointly exhaustive. So for every sentence in the language, it is either ungrounded or not ungrounded.

If Kripke's theory is going to provide a uniform solution to the liar paradox, then one option is to treat the revenge sentence, just like the liar, as ungrounded.

Suppose R_g is in the extension of the ungroundedness predicate, then it will enter the interpretation process at the initial stage. At the second stage, it will be collected by the jump operator. In particular, it will be taken into the extension of the predicate T in all later stages. This means that it will have a classical truth value at the fixed point, which makes it a grounded sentence. Contradiction. Suppose on the other hand, that it is in the anti-extension of the ungroundedness predicate. This means that it is grounded and so the second disjunct is not satisfied. So its semantic status will depend on the semantic status of the pristine liar sentence. And since the pristine liar sentence is ungrounded, it is, ungrounded as well. Contradiction.

[103] That these two notions cause trouble for Kripke's theory is widely discussed and accepted. The same sorts of discussion (though may appear in different forms) can be found in, Simmons (1993), chapter 3, section 3.2; Kirkham (2001), chapter 9, section 9.5; Sher (2006), p.158.

2.5.2 Exclusion negation[104]

This problem can be viewed from another perspective. There are two kinds of negation—choice negation and exclusion negation. Choice negation takes truth into falsity and vice versa while exclusion negation is more like "everything other than". So if you say "snow is white" is not true, where "not" here is choice negation, then it means only that it is false. But if you say the same thing with an exclusion negation, then what you mean is more like that it is anything but truth, which, in a three-valued language may be regarded as saying that it is either false or neither true nor false. Let not-$_{ex}$ and not-$_{ch}$ represent exclusion negation and choice negation respectively. In classical two-valued semantics these two make no difference, but in three-valued semantics, while being not-$_{ch}$ true is equal to being false, being not-$_{ex}$ true is equal to being false or *n*, where *n* is used to express the third truth value in the three-valued semantics—either it is undefined, or neither true nor false or something else.

If we expand strong Kleene semantics to a semantic schema with an exclusion negation, then first of all, the semantics of it is no longer strong Kleene semantics; second, the notion of truth we have employed outside of the object-language will be modified correspondingly (by adding a new clause for the newly introduced connective). Third, the corresponding jump operator will also be equipped with the new semantics. The semantics for exclusion negation is as follows:

Rules for exclusion negation[105]

The exclusion-negation of a sentence A is true if A is either false or undefined;

The exclusion-negation of a sentence A is false if A is true;

Now consider the following revenge sentence:

---

[104] My knowledge on the expressibility of the logical connectives in a three-valued language like Kripke's is mainly from Gupta and Belnap (1993). For more technical discussion on which logical connectives cannot be expressed in three-valued language, see Gupta and Belnap (1993), chapter 3.

[105] See Gupta and Belnap (1993), p.94, for a truth-table version of the semantics of exclusion negation.

R_g': R_g' is not-$_{ex}$ T.

How does this sentence behave in Kripke's construction? In Kripke's construction, g_0 is empty, nothing in the extension of T nor in the anti-extension of T. R_g' is true in L_{g0}, so it will be collected by the jump operator and put into the extension of T in the next stage, forming a new interpretation, g_1.

The trouble is, at the second stage, since R_g' is in the extension of T in L_{g1}, it will then become false and be put into the anti-extension of T, forming a third interpretation g_2. The process can keep going on, and presumably the truth value of R_g' will keep vibrating. The jump operator in this case does not have a fixed point at all.

Note that the conclusion we have here does not violate Kripke's result above. For his original result assumes first, the language at issue has no semantic predicate other than T; second, the language is a three-valued language based on strong Kleene semantics. In other words, his conclusion that a language can contain its own truth predicate if it admits the truth value gap applies only for a relatively small group of languages, which although, can be semantically closed, are short in expressing certain semantic notions and logical notions.

The introduction of the groundedness predicate as well as the expansion of the semantic schema (by introducing exclusion negation) gives us a language that is different from what Kripke was working on. In Kripke's language model, the revenge sentence like R_g and R_g' are not expressible at all. Nevertheless, Kripke does not deny the expressibility of these notions in natural language, and so there is a gap in expressibility between his language model and natural language. As a result of the revenge problem, Kripke is forced to ascend to a richer language,[106]

[106] Kripke admits that his language model does not provide a model for a language model, that the terms like "grounded" are not expressible in the object-language. See Kripke (1975), p. 714, n.34.

which contains the groundedness predicate for the object-language. But if natural language, as he argues, does not come in a hierarchy, then there is no way to ascend. For this problem he seems to suggest that if the fixed point is taken to be an interpretation for natural language, then our claim about the ungroundedness of a sentence must be viewed as something we get after reflecting on the interpretation process that leads to the fixed point rather than something completed in the same interpretation process. This idea can be manifested as follows.

The interpretation process is constituted by a sequence of interpreted languages:

$L_{g0} \rightarrow L_{g1} \rightarrow L_{g2} \ldots L_{gn} \ldots \rightarrow$ fixed point $\rightarrow$ interpretation on groundedness

The suggestion is that, in this sequence, the extension/anti-extension of ungroundedness predicate is not (cannot be) taken into consideration before reaching the fixed point. Only after we reflect on the whole interpretation process leading to the fixed point do we realize that some sentences are ground in the whole process and some others are not and so reach an even later stage where the groundedness predicate gets interpreted based on our reflection on the "whole" interpretation process. The situation is similar to what we have seen in the Tarskian hierarchical solution. There a claim about the hierarchy as a whole must not be in the hierarchy, and here a claim about the interpretation process as a whole must not itself be in the process (in the sense that the relevant notions required for describing the process must not be interpreted in the process).[107]

In summary, the conflict inside the approach is this: The groundedness of a sentence is defined in terms of an interpretation process. Before reflecting on the interpretation process, a sentence talking about its ungroundedness is ungrounded (if we try to solve the revenge in the same way we solve the pristine liar). But once

[107] Note that at this moment I do not have a final position on this strategy. But none of what I say here and below depends on it. If one can avoid the contradiction by excluding the groundedness predicate from being interpreted in the above interpretation process, then it only shows that the language to which the interpretation process belongs is incomplete. If, on the other hand, it does not work, then the revenge problem remains and so there is no way to escape inconsistency. Either way, the intended conclusion that Kripke's theory cannot escape *Dialetheist Conjecture* still holds.

we have completed the interpretation process, the fact of how the sentence behaves in this process seems to provide a ground for evaluating the sentence and so making the sentence grounded. To avoid this awkward situation we need to somehow remove the evaluation of the revenge sentence from the interpretation process. One way is to eliminate it once and for all, say, to deny the legitimacy of relevant concepts. Another way is to put them into a richer meta-language, which, for natural language, means that the interpretation process is not a complete interpretation process for the whole language, not at least include the groundedness predicate (and other predicates that may be defined based on it).

In the first case, the theory requires some independent reasons for rejecting those notions, on pain of being ad hoc, but so far I do not see any such reason. In the second case, the theory fails to be a model for natural language, for there is a gap in expressibility between the object-language and natural language. In any case, the *Dialetheist Conjecture* seems to follow.

2.6 Absolute revenge?

I need to mention one more possible attempt to avoid the above revenge problem. The above argument is based on the assumption that the groundedness predicate is classical in the sense that every sentence is either grounded or ungrounded. But it may be suggested that this needs not be the case (recall Beall's too-easy revenge in chapter 1). Some sentence may be neither grounded nor ungrounded, and the revenge sentence R_g perhaps is one of them.

In addition to the problem of motivation (why the revenge sentence is neither grounded nor ungrounded), the approach seems to have its own revenge sentence:

R_g*: R_g* is not T, or ungrounded or neither ungrounded nor grounded.

For easy reference, let us call the original ungroundedness predicate ungrounded_0 and the new introduced predicate ungrounded_1, where being ungrouded_1 is to be neither grounded_0 nor ungrounded_0. One may use the same kind of method to define

more and more ungroundedness predicate, ungrounded$_2$, ungrounded$_3$,…

However, Priest's pattern for *type S* revenge[108] seems to rule out this attempt once and for all. For we may simply construct the following absolute revenge sentence, using Priest's pattern as follows:

$R_{g\text{-}ab}$: $R_{g\text{-}ab}$ is not T, or ungrounded$_0$, or ungrounded$_1$, or ungrounded$_2$,…

If $R_{g\text{-}ab}$ is classified eventually as being ungrounded$_i$ for some i, it will then be true and so be grounded$_0$, and so be grounded$_1$, grounded$_2$,…until grounded$_i$. So it cannot be ungrounded$_i$, for any i. But if so, its truth status will then be just like the pristine liar sentence, which, according to the original theory, is ungrounded$_0$. Contradiction.

Now of course, the construction of the ungroundedness predicates sequence here is just one out of many possible constructions. There may be some construction that may avoid the above absolute revenge problem. For other possibilities, I leave to further study.

3. Gupta's revision theory of truth

Gupta and Belnep's revision theory is first a theory of definition and then a solution to the liar paradox. In their words, the key to solve the liar paradox consists in a proper understanding of definition. Traditionally, only non-circular definition is admitted. Circular definitions and circular concepts are normally ruled out as not legitimate. However, Gupta and Belnap suggest that we can understand most of the behaviors of the concept of truth (including the liar paradox), once we take truth to be a circular concept (and the truth predicate to be circularly defined). The strange behavior of the truth predicate in the liar paradox is nothing but a special case of the behavior of circularly defined predicates.

[108] See chapter 1, section 2.

In the following, I will first introduce Gupta and Belnep's theory[109] of definition based on their (1993) and then explain how this theory helps us understand the behavior of the truth predicate in the liar paradox. And finally, we will see its own revenge problem.

3.1 Circular definition: its semantics and logic

Circular definition is a kind of definition in which the definiendum (the term being defined) appears in its definiens (the terms defining the definiendum). Normally we reject circular definition. The traditional view is that, the purpose of a definition is to tell/fix the *meaning* of the definiendum via the meaning of the definiens. But if the definiendum appears in the definiens, then one could not know the meaning of the definiens before he or she knows the meaning of the definiendum. The function of the definition fails. Another problem of circular definitions is that, they are, in Gupta's word, *creative*. That is, it is possible to infer something that is absurd from a circular definition (Gupta and Belnap, 1993, 113). If Gupta wants to legitimize circular definitions, he will need to provide an acceptable semantic theory and an acceptable logic. Below I will sketch both theories briefly.

3.1.1 Semantics for circular definition

Traditionally, we believe, among the other things, that in a definition, it is the definiens that fixes the meaning[110] of the definiendum. In an extensional analysis, we may say that the definiens provides a rule that determines the extension and/or

---

[109] For simplicity, from now on and until the end of this dissertation, I will refer to their theory as "Gupta's theory". And whenever I say "According to Gupta", "Gupta suggests", what I mean is "according to Gupta and Belnap", "Gupta and Benlap suggest", if there is no further indication.

[110] The term "meaning" is indeed ambiguous. Here and below I use it to refer to what Gupta calls *signification* of terms: "Let the (*extensional*) signification of an expression (or a concept) in a world w be an abstract something that carries all the information about the expression's extensional relations in w. The signification of a *classical* predicate can be represented by its extension or by a function that determines of each object whether the predicate is true or false of it."(Gupta and Belnap, 1993, 30). For example, in classical semantics, the signification of a predicate is its extension. In three-valued semantics, the signification of a predicate is its extension and anti-extension. We will see that in Gupta's new semantic theory, the signification of a circularly defined term will be something different.

anti-extension (and in general, the signification) of the definiendum. So given the meaning of the definiens, which allows us to know the signification of each term in the definiens, we can calculate the signification of the definiendum. Given a model (an interpretation to the terms in definiens), we may say, a legitimate definition of a term determines its signification and so fixes its meaning in the model. So in most of the cases (here we ignore cases where something like vagueness and context occur), if a legitimate definition is given, then the definiendum has a *definite* signification. In an extensionl analysis, we normally identify this signification as the *meaning* of the definiendum.[111]

In *classical semantics*, we say that the signification of a term is its extension (the set of objects to which the term applies), and so a meaningful term of this sort has a definite extension. However, in the case where circular definitions are involved, we have trouble regarding the meaning of the definiendum. For in this case, there may be no such thing (not at least in general) as the *definite* extension of the definiendum because the definiens may fail to provide a rule that determines the extension of the definiendum categorically in a model.

For example, consider the following definition:

Definition I: For any a, a is G $=_{Df}$ a is not H or a is F and not G[112]

Here "$=_{Df}$" may be read as "is defined as" and H and F are two one-place properties. G is the definiendum and since it occurs in the right-hand side of the definition, the definition is circular.

Now suppose in model M_1, object a_1 is H and is F, then whether a_1 is G will depends on the last clause, namely "a_1 is not G". But since M_1 does not interpret G, we cannot know whether it is the case that a_1 is not G, and so we cannot know that

[111] See Gupta and Belnap (1993), p.118, for the main idea of the traditional account of meaning summarized here.

[112] Definition of this kind is easily obtained. For Gupta's own example, see Gupta and Belnap (1993), p.113.

whether a_1 is G.

So the traditional account of meaning cannot cover, at least in general, the meaning of the defniendum that is circularly defined. Gupta therefore suggests to think of meaning in a different way.

Gupta's suggestion is that, at least for circular definition, the meaning of the definiendum is not a definite set of objects but a function which yields a rule of revision of the extension of the definiendum (Gupta and Belnap, 1993, 119). Given a hypothetical extension of the definiendum *on the right-hand side*, the function will produce a resulting extension. Given a circular concept, we cannot say that some definite set is its extension, but we can, according to its definition, say that under the hypothesis that its extension is set A, its extension will be set B.

In ***Definition I***, we say that the extension of predicate G cannot be settled by the definition, for in order to know whether a given object, say, a_1, is G, we need to know whether a_1 is not G. Now suppose that the hypothetical extension says that a_1 is not G, then the right-hand side of the definition, which is "a_1 is not H or a_1 is F and not G" will hold. Thus according to the definition, "a_1 is G" holds. Similarly, given any object in the domain, we can calculate whether it is G *in a model with a hypothetical extension of G* and so generate a new extension of G.

So we can think of ***Definition I*** (and indeed, any circular definition) as a kind of definition that yields not a definite set but a function of the above kind as the signification of the definiendum, which gives us "extension under various hypotheses" (Gupta and Belnap, 1993,120).

With the new understanding of the meaning of a circular definition, it is possible to

define some semantic concepts and develop a different semantic theory.[113]

Consider the revision function determined by ***Definition I***. For easy reference, let us call it $\delta_{D, M}$. The subscript "D" indicates that the function is relative to a certain definition (which in this case is ***Definition I***), while "M" indicates that it is relative to a model.

When $\delta_{D, M}$ applies to a hypothetical extension of G, say, h_1, it yields a revised extension $h_2=\delta_{D, M}(h_1)$. By iteratively applying the same function, we attain a sequence of revised extensions of G, $<h_1, h_2, h_3,\ldots>$. There are some key semantic concepts that can be defined by means of this sequence.

Suppose that there is an object b, which, in model M_2, is not H. By ***Definition I***, it is G, regardless how we hypothesize the extension of G. This is because one of the disjuncts in the definition is satisfied by b, so whether or not it satisfies the rest of the clause does not affect its being G. This fact is significant because it shows that, even when it comes to circular concepts, which have no definite extension in general, we can still make categorical claims in certain cases to assert *definitely* whether a given object falls under the extension/anti-extension of the predicate. Projecting this fact into the revision sequence, we discover that this sort of sentences are sentences whose truth values do not change *after some point* in the sequence, regardless of which hypothetical extension we choose to begin the revision sequence. If a sentence stabilizes as being true after some point, then it is a *valid sentence*.[114] Gupta calls a sentence *categorical* if it or its negation is *valid*

---

[113] The following construction, including the definitions of several key semantic concepts, is what Gutpa calls system S_0. It is not the final system Gupta adopts, but it is used as an example to show the main idea of the semantic theory he has in mind. For our current purpose, we need not discuss other systems. For those systems, see Gupta and Belnap (1993), chapter 5.

[114] A technical definition of "valid sentence" is as follows: (Gupta and Belnap, 1993, 123):

> A sentence *A* is *valid in* M (*relative to* the definition D in the system S_0) iff there is a natural number p such that, for all $q \geq p$ and all subsets X of the universe, *A* is true in $M+\delta^q_{D,M}(X)$.

The definition is an accurate characterization of the fact that the truth value of the sentence stabilizes after some point in the sequence. An interesting equivalence is that, when the sentence

in the above sense (Gupta and Belnap, 1993,123). Categorical sentences are therefore sentences which can be categorically asserted in a given model, even if they involve circular concepts.

Other groups of sentences can also be defined by means of revision sequence:

Paradoxical sentences are identified as sentences whose truth value oscillates[115] in the revision sequence generated by *each* initial hypothesis of the extension of the definiendum.

Pathological sentences (which include paradoxical sentences) are defined as sentences which are not categorical. One quick example for a pathological sentence that is not paradoxical is the truth-teller. Its truth value in the revision sequence is a constant line: sometimes it is true throughout the revision sequence, sometimes it is false, depending on the initial hypothesis of the extension of the definiendum, but it does not oscillate.

3.1.2 And its logic

Even though we can make sense of circular definition, it remains a big problem for allowing such kind of definition in doing reasoning. For circular definition is *creative*. It allows us, with classical logic, to infer something that is substantial and possibly absurd.

According to Gupta, we have two inferential rules concerning definition. Let us consider a correct definition.

Definition II: For any object a, a is T $=_{Df}$ a is R and not E.

Now given an object a, one can infer "a is T" from "a is R and not E".Conversely,

---

A is *valid* in the above sense, there will be a point after which the sentence will be true, regardless how we hypothesize the initial extension of the definiendum. Also note that a *valid* sentence in this sense is different from valid sentences in logic. The latter is a logical truth. I make such a distinction later by using the italic form of the term "valid" for the former and the normal form of "valid" for the latter.

[115] For example, being true in one stage and being false in the next and so on. Note that this is just one kind of oscillation and so it only defines one kind of paradoxicality.

one can infer "a is R and not E" from "a is T". In short, given a definition, one can infer its definiens from its definiendum and vice versa. These two inferential rules are called DfI (Definiendum introduction) and DfE (Definiendum elimination) by Gupta and Belnap (1993, 114).

With these two inferential rules and classical logic plus a circular definition, one can actually infer quite a lot of things, things that are substantial but are irrelevant to the definiendum.

Consider again ***Definition I***: For any a, a is G $=_{Df}$ a is not H or a is F and not G. With a few assumptions of an object a and G, we can infer that everything that is F will also not be H.[116]

Proof:

Suppose the otherwise:

(1) there is an object a such that a is H and is F

And now suppose that

(2) a is G

By DfE, we can infer:

(3) a is not H or a is F and not G

with (1) and classical logic, we can infer:

(4) a is not G

So from the assumption that a is G, we infer that a is not G. Contradiction. By reductio, we conclude:

(5) a is not G

But if a is not G, then with (1) and DfI, we can infer:

(6) a is G

Another contradiction again, so a cannot be not G.

[116] Below I use Gupta's method to do the proof. For Gupta's original proof, see Gupta and Belnap (1993), pp.113-114.

Since for any a, either a is G or a is not G (law of excluded middle), we come to the conclusion that (1) is false (we assume (1) and infer from this assumption that a is neither G nor not-G. By reductio, we conclude that (1) is false).
But if (1) is false, that means, there is no such an object that is both F and H. This implies the following:
(7) Any object that is F will be not-H.

Now "F" and "H" are two arbitrarily chosen properties. They can be anything without affecting the validity of the above inference. That means, we can prove a lot of things, simply by ***Definition I***. For example, let F be "is a man", and let H be "is married", then we prove from ***Definition I*** that every man is unmarried.
However, this does not mean that any circular definition in any situation will only give us some problematic but valid inference. In some cases, we could have inferences that are valid and unproblematic, at least, not absurd. For example, suppose that under a certain model, there is an object such that it is not H. In this case, we can apply DfI to deduce that it is G by definition, which is unproblematic.[117]

So if we want to use circular concepts in reasoning, then the next question is whether or not we can have a logical system which preserves correct reasoning with circular concept and at the same time invalidates those that are problematic.

Gupta's answer is yes. The key modification of the traditional rules of inference consists in the modification of DfI and DfE. Traditionally, one can infer a definiendum from the definiens and vice versa without restriction. But as we have seen, under these unrestricted rules, reasoning with circular concepts may come out with something absurd. So some restriction is desired. Gupta's semantics for circular definition implies one possible solution.

[117] For more on this, see Gupta and Belnap (1993), pp120-123.

Observe that in a revision sequence of a circular definition, say <h_0,h_1, h_2, h_3,…>, when the definiens holds at stage n, the definiendum will hold at stage n+1. And if the definiendum holds at stage n+1, then the definiens will hold at stage n. This semantic suggests two modified rules of inferences.

For illustration, let us see a concrete but very simplified example. Suppose we have a classical model M=<D, I>, where D={a} and I is an interpretation function which assigns extension to the following predicates:

I(H)= {a}

I(F)={a}

Now consider ***Definition I***: For any a, a is G $=_{Df}$ a is not H or a is F and not G

Let the revision sequence of the predicate G begins with empty set, that is, let us first assume that the initial extension of G, h_0, is ∅, which contains nothing.

Now we consider the truth value of the definiendum and definiens at each stage:

h_0=∅: "a is G" is false;

"a is not H or a is F and not G" is true.

$h_1=\delta_{D, M}(h_0)$={a}: "a is G" is true;

"a is not H or a is F and not G" is false.

$h_2=\delta_{D, M}(h_1)$= ∅: "a is G" is false;

"a is not H or a is F and not G" is true.

…

$h_{2n+1}=\delta_{D, M}(h_{2n})$={a}: "a is G" is true;

"a is not H or a is F and not G" is false.[118]

One can easily see that the pattern: at the odd stages of revision, the definiens will be false, while the definiendum will be true; at the even stages, the definiens will

[118] Recall that h_{n+1} is obtained by collecting everything that satisfies the whole definiens in stage n, where G is assumed to have the extension h_n.

be true while the definiendum will be false. Throughout the sequence, the truth value of the definiens at stage n will always be the same as the truth value of the definiendum at stage n+1. This gives us Gupta's suggestions: "…when applying DfI and DfE we should keep track of the stages of revision." (Gupta and Belnap, 1993,126). The new rules of inference replacing the unrestricted DfI and DfE are as follows:[119]

DfI_r: [a is not H or a is F and not G" is true]$^{i} \rightarrow$[a is G]$^{i+1}$

DfE_r: [a is G]$^{i} \rightarrow$[a is not H or a is F and not G" is true]$^{i-1}$

The superscripts indicate the stages of revision in the corresponding revision sequence. And the subscript "r" in "DfI_r" and "DfE_r" is just for distinguishing the new rules from the old ones.

There are some exceptions where we can simply ignore the indices. For example, where the relevant definition is non-circular, one can work with unmodified DfI and DfE. This is because of a special rule of inference, which Gupta calls "index shift" (Gupta and Belnap, 1993, 126). It allows the index of a sentence to shift arbitrarily if it does not contain the definiendum.[120]

Moreover, where circular definition is involved, application of classical logic should also track the indices of relevant sentences. In short, application of a classical rule of inference is allowed only when the indices of relevant sentences are the same. So for example, if we are to apply modus ponens to deduce the sentence "A" from the sentence "If B then A" and "B", then we should ensure that the three sentences here have the same index, if circular definition is involved in

---

[119] For Gupta's presentation of these two rules, see Gupta and Belnap (1993), p.126.

[120] Actually, Gupta mentions one more situation where the index can be dispensed with:

> …the indices employed in the calculus are important only in the context of *hypothetical* reasoning. Within categorical contexts, they can be dispensed with; one can work with DfI and DfE unmodified. This is a consequence of the fact that a uniform shifting of indices preserves the correctness of derivations. (Gupta and Belnap, 1993,128)

However, Gupta does not define the key term "hypothetical reasoning" and "categorical context". It is therefore unclear exactly what situation he refers to in which the indices can be dispensed with.

those sentences.[121] So classical logic works freely at each stage.

So the new logical system in Gupta's theory is a combination of the following three:

DfI_r and DfE_r + Index Shift+ Classical Logic (in each stage).

Let us see how the new logical system invalidates *creative* inference. Consider the creative inference with ***Definition I*** at the beginning of section 3 above: We prove that everything that is F will not be H by supposing the otherwise, and deduce a contradiction. The reasoning requires the application of the unrestricted DfI and DfE at step 2 and step 5, which are now invalidated for using unacceptable rules of inference.

Now under the new system, the valid reasoning will be the following:

Definition I: For any a, a is G $=_{Df}$ a is not H or a is F and not G

(1) Suppose that there is an object a such that a is not H and is F;

Suppose that,

(2) $(a \text{ is } G)^i$

By DfE_i, we can infer:

(3) $(a \text{ is not H or a is F and not } G)^{i-1}$;

with (1) and classical logic, we can infer:

(4) $(a \text{ is not } G)^{i-1}$.

So from the assumption that $(a \text{ is } G)^i$, we can infer that $(a \text{ is not } G)^{i-1}$

Since $(a \text{ is } G)^i$ and $(a \text{ is not } G)^{i-1}$ are of different stages, they fail to constitute a contradiction.

Now some may question: why don't they constitute a contradiction? Simply attaching different indices to a pair of contradictory statements A and not-A does not dispel our puzzle. Otherwise it is quite a cheap thing to eliminate contradiction. The attachment of indices must have justification so that we can be convinced that

[121] There is a question concerning with the scope of the index. For example, consider the conjunction, "A and B", should the index be attached to the whole conjunction, so that we write it as: $(A \text{ and } B)^i$ or it should be assigned to each conjunct respectively, so that we have: A^i and B^i? Gutpa does not clarify this and for our purpose, I need not to resolve this issue here.

the two seemingly contradictory statements are not really contradictory.

The ground in Gupta's theory, of course, is from his semantic theory which views the above kind of reasoning as only a small part of a revision sequence. In a revision sequence, it is not only possible, but also natural to accept both "a is G" and "a is not G", because we do not need to accept them as holding *at the same time*. But does the logic really reflect the actual way of reasoning, when we are thinking with a circular definition? Note that, an index in Gupta's theory is used to indicate stage of revision. If we are to apply them in actually reasoning, then we can only regard ourselves as thinking about a revision process. But before Gupta, perhaps nobody will regard himself as beginning a revision sequence defined technically in Gupta's theory, when he or she is making an assumption like "if a is G". Moreover, Gupta's logic does not directly distinguish circular definition and non-circular definition—it is supposed to apply to any predicate in general. To maintain the normal reasoning with non-circular definition (in which case, we use DfI and DfE, directly and deduce that both the definiens and definiendum hold *at the same time*), he introduces the special rule "Index Shift", so that he can keep the indices of both sides of the definition to be the same. But that means after using DfEr and DfIr, one still have to employ one more rule of inference to get his intended result. This seems to be something that does not happen in actual reasoning. Thus my current view on Gupta's logical system is that it works to block the creative, unacceptable inference with circular definition, but the system fails to be descriptive. But perhaps when it comes to logic, Gupta does not aim at a descriptive project, but a normative one—the actual reasoning is defective, so we need to change it a little bit. If this is what Gupta has in mind, then his logical theory is indeed successful, at least, in the sense that it blocks some unintended inferences involving circular definition.[122]

[122] A similar question can be asked about semantics. Is it really the case that when we are making a statement like "if a is G…", what we mean is "if we assume that a is in the extension of G in a revision sequence…"? Certainly not, at least, not with our intention. But again, what Gupta is doing is to provide a new semantics that can accommodate the meaning of circular concepts. In other words, he provides us with a new understanding of the meaning of some terms, which, before him, are ruled out as illegitimate.

3.2 Truth as a circular concept

Let us assume that Gupta's theory so far (which includes a semantic theory for circular definition and a logical theory for the corresponding logic) is good enough to legitimize circular definitions and circular concepts. The next job would be to examine whether or not truth is indeed a circular concept and if it is, whether or not Gupta's theory can help to solve the liar paradox.

3.2.1 Preliminary note: partial definition and T-biconditionals

Gupta accepts what is normally called *partial definition*. This kind of definition for a term does not specify in general the application condition of the term for any object, but it will specify at least the application condition for some objects. The extreme case is that it specifies the application condition for *one* object.

So for Gupta, the following form of partial definition is acceptable:

For the object b, b is G $=_{Df}$ A(b);

Here "A(b)" represents a sentence of the form "b is A". If the sentence contains G, that is, if it contains the definiendum, then it can be represented, like Gupta suggests, as A(b, G), so we have the following form of partial circular definition:

For the object b, b is G $=_{Df}$ A(b, G).

Now an example of partial circular definition, according to Gupta, is the set of T-biconditionals. Tarski once described each T-biconditional as a partial definition of truth which specifies what it is to call a particular sentence true.[123] Gupta agrees with this view and actually takes it to be the basis for his theory of truth (Gupta and Belnap, 1993, 132). In Gupta's theory, all T-biconditionals when read as definitional will be taken to be of the following form:

"A" is true $=_{Df}$ A

And we take Tarski's T-biconditionals to be of the following form:

"A" is true$\leftrightarrow$A.

For distinction I will label all definitional T-biconditionals as T_{df}-biconditional. The

[123] See chapter 3, section 1.

difference between T_{df}-biconditional and T-biconditional consists in the connective "$=_{Df}$" and "$\leftrightarrow$". The former is a connective which may be interpreted as "is defined as" while the latter one is a material biconditional (the logical connective in classical logic).

Though Gupta does not specify a semantics for "$=_{Df}$" and "$\leftrightarrow$", it is obvious that the logic for the connectives is quite different. One can infer that the definiendum holds at stage n, from the assumption that the definiens holds at stage n-1, and vice versa. But one cannot therefore infer the corresponding material biconditionals, "for it requires that the premises and the conclusions have the same indices" (Gupta and Belnap, 1993,138).

If my understanding is correct, then the correct form of T-biconditionals, and T_{df}-biconditional (and actually for any definitional biconditional) should be the following:

T_{Df}-biconditionals: ("A" is true)$^n =_{df} A^{n-1}$

T-biconditionals: ("A" is true $\leftrightarrow$ A)^i

where n and i here indicate the stage of revision.

One can infer a T-biconditional from the corresponding T_{Df}-biconditional when one can infer ("A" is true)^i from A^i and vice versa and eventually using propositional logic to infer ("A" is true $\leftrightarrow$ A)^i. This situation is possible when, for example, A is categorical[124] (for now index does not matter).

So while Gupta's original theory accepts preliminarily all T_{Df}-biconditionals, and does not say a word on T-biconditionals, the theory accepts most of the T-

[124] For example, in a ground model where snow is white, it is easily seen that the sentence "'snow is white' is true" is categorically assertable. Its truth value therefore will be a constant line after some point at the revision sequence regardless how we begin the sequence. And it can be easily seen that after some point, the truth value of the sentence " ' 'snow is white' is true' is true" and " 'snow is white' is true" will always be the same. The index here thus is not important.

biconditionals. An interesting thing is, this distinction allows Gupta to accept all T_{Df}-biconditionals and at the same time to reject some T-biconditionals, like the T-biconditional for the liar sentences. In Gupta's words:

> …the distinction between the two kinds of biconditionals enables us to accommodate tendencies that otherwise conflict with one another: (i) to accept the Tarski bicondtional [T-biconditional] because it is definitional of truth and (ii) to reject the same biconditional because it is a contradiction [false]. (Gupta and Belnap, 1993, 139)

3.2.2 Circularity of truth

We can see at least two justifications in Gupta's theory for the claim that truth is a circular concept. The first one is just a comparison between the behavior of circular concepts in general and the behavior of the concept of truth. The second one is an argument that originates from Tarski's view that each T-biconditional can be regarded as a partial definition of the concept of truth. When we treat T-biconditionals as definitional of the concept of truth, we will discover that the semantics for circular definition works for the concept of truth—the set of T-biconditionals determines a revision function.

A. *Similarity between circular concepts and the concept of truth*

Let us see again ***Definition I***: For any a, a is G $=_{Df}$ a is not H or a is F and not G

Now suppose that in a model, there is an object b such that b is H and b is F, then by definition, the question about whether b is G will come down to the question about whether b is not G.

The situation here is, in order to determine whether or not b is G, we need to determine whether b is not G. On the other hand, if we are to know whether or not it is not the case that b is G, then what we need to know is whether or not it is not the case that *b is not H or b is F and not G*, which in the model we assume, can be simplified as whether or not it is not the case that b is not G, or more simply, whether

or not b is G. Following Gupta,[125] we can picture the process in diagrams:

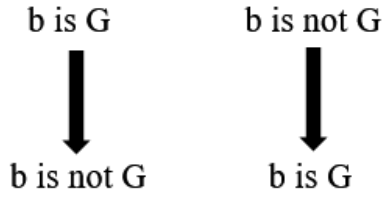

Figure 3-1

The arrows here represent the direction of the above determination process. So we are entering into an endless loop: To determine whether b is G, we need to first determine whether b is not G. But to determine whether b is not G, we need to first determine whether b is G.

Gupta thinks that this strange and endless loop is exactly what we see in the case of the liar paradox. To ask whether the liar sentence is true, we are directed to the question of whether the liar sentence is not true; and to ask whether the liar sentence is not true, we are directed to the question of whether the liar sentence is true. The loop for the liar sentence can be pictured as the following diagram:

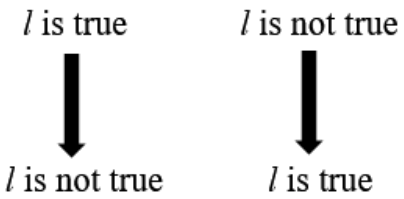

Figure 3-2

The name *l* here is the name of the sentence "*l* is not true". The similarity between the loop for ***Definition I*** and the loop for the liar sentence may suggest that there are indeed some connections between circular concepts and the concept of truth and perhaps the strange behavior of the truth predicate in the liar sentence is just one of the instances of the general phenomenon of circular definition.

More connections between circular concepts and truth can be observed. Let us modify ***Definition I*** a little bit and get ***Definition II***:

For any a, a is G $=_{Df}$ a is not H or a is F and is G

[125] For Gupta's original diagrams, see Gupta and Belnap (1993, 115).

The difference between ***Definition I*** and ***Definition II*** consists in the last clause, where in the former definition it is "is not G", in the latter one it is "is G".

With a similar reflection, the loop for determining whether the object b (above) is G now turns into the following one (on next page):

Figure 3-3

To determine whether b is G, we are directed back to the very same question—whether b is G. This phenomenon is the same as the one that we observe in another pathological sentence, truth-teller.

t: *t* is true

To determine whether *t* is true, we are directed back to the very same question—whether *t* is true. The loop now can be pictured as such:

Figure 3-4

Besides the similarity between circular concepts (in our case, the concept of G in two definitions) and the pathologicality of the concept of truth, there is one more aspect (at least this is the only one other than the above two that Gupta mentions) that they behave similarly. That is, both can be applied unproblematically over a

great range of objects (Gupta and Belnap, 1993, 117). For circular concepts in general, we have seen that one can still use them to make categorical statements. In ***Definition I***, the definition for G, in many situations, can still be used to unproblematically determine whether an object is G. For example, any object that is not H, by ***Definition I***, would count as G. As to the truth predicate, at least for sentences that involve no truth predicate or sentences that are not constructed via some vicious references, one can determine whether the sentence at issue is true in a given model unproblematically.

These similarities do not settle the issue of whether truth is indeed a circular concept, and whether the behavior of the truth predicate *is* the behavior of circular concepts, but it does show that there are some connections here. At least, you can certainly use Gupta's method to construct infinitely many "like-liar" contradictions with the help of circular concepts.

B. T-biconditionals as circular partial definition of truth

As mentioned, Gupta's theory distinguishes two sorts of T-biconditionals. Normally, T-biconditionals are read as material biconditionals, but Gupta's theory reads them as T_{Df}-biconditionals, which he thinks is the correct reading on Tarski's view. If we read all T-biconditionals as T_{Df}-biconditionals, then one thing results—there will be some partial definitions that are circular definitions.

This would happen when the partial definition is a T_{Df}-biconditional for a sentence that contains the truth predicate. Consider the following two T_{Df}-biconditionals:

(i) ""Snow is white" is true" is true$=_{Df}$ "snow is white" is true

(ii) "Everything Jones says is true" is true$=_{Df}$ Everthing John says is true[126]

Now in both partial definitions, the definiendum appears in the definiens, then it might be possible that the truth predicate defined by these partial definitions will be

[126] See Gupta and Belnap (1993), p.132, for the original presentations of these two examples. Note that there may be ways to eliminate the circularity of both examples, but in some situations, especially the second one, elimination is difficult. Gupta admits that the circularity in the first one can be eliminated but denies that we can do this for the second one.

circular. Gupta then applies the previous theory of circular definition (including the semantic theory and the logic) to the truth predicate. In that case, if the truth predicate is so circularly defined then the set of T_{Df}-biconditionals should determine a revision function. With a little reflection, it seems that this is indeed the case.

For simplicity, let us consider a relatively simple language which contains only three sentences:[127]

(i) Snow is white

(ii) "snow is white" is true

(iii) " 'snow is white' is true" is true

The corresponding T_{Df}-biconditional for each sentence above can be easily constructed.

Let the model interpreting the language be $M+g_i$, where M is a model in which snow is indeed white and g_i represents the hypothetical interpretation of the truth predicate. Let us assume that $g_0=\varnothing$ (that is, empty), in this case the extension of the truth predicate does not contain anything. So according to each T_{Df}-biconditional for (i), (ii) and (iii), we have:

(i) is true in $M+g_0$, and (ii) and (iii) are not true in $M+g_0$.

We collect all sentences that are true in $M+g_0$, which is just (i), and so the set of T_{Df}-biconditionals determines a new extension of the truth predicate $g_1=\{$"snow is white"$\}$.

Now we enter the second stage of revision, where in this case we take g_1 to be the hypothetical interpretation of the truth predicate.

According to each T_{Df}-biconditional, we have:

(i) is true in $M+g_1$, (ii) is true in $M+g_1$, (iii) is not true in $M+g_1$.

[127] For Gupta's own example, see Gupta and Belnap (1993, 133-134).

This gives us a new extension of the truth predicate g_2={"snow is white", ""snow is white" is true"}.

In the revised model M+g_2, we have all the sentences true, and so we get once more a new revised interpretation of the truth predicate g_4={"snow is white", " 'snow is white' is true", " ' 'snow is white' is true' is true"}.

With a little reflection on the procedure, one can easily see that after this stage, the extension of the truth predicate will not change any more, however we repeat the same operation. Thus we reach a fixed point of the revision function.

The example shows that the set of T_{Df}-biconditionals does determine a revision function for the truth predicate, and so by the semantic theory for circular definition, the set of T_{Df}-biconditionals fixes the signification of the truth predicate. Gupta calls the function, *Tarski jump*, or *classical jump* (Gupta and Belnap, 1993,133).

The signification of truth determined by the T_{Df}-biconditionals has several desired features, which Gupta thinks can be used to explain several behaviors of the truth predicate.

We have seen that different sorts of sentences involving circular concepts may behave differently in a revision sequence but still, the behavior of each sort of sentences will follow some sort of pattern respectively. Gupta shows us that the same thing happens when it comes to the truth predicate.[128]

There are categorical sentences, sentences whose truth value will stabilize after some point in the revision sequence. In the little language we discuss above, all three sentences are categorical. In Gupta's view, as long as the language at issue does not have any *vicious* reference (like those generating the liar sentence or the

[128] The following is a summary of some simple patterns of revision sequences of the truth predicate. For Gupta's original and complete presentation of these patterns, see Gupta and Belnap (1993), pp. 134-137.

truth-teller sentence), the revision sequence for the truth predicate (or say, the Tarski jump) will have one and only one fixed point. In other words, each sentence in the language will be categorically assertable. If, on the other hand, the language contains some vicious reference, then the revision sequence for the truth predicate will either be unstable, in the sense that it does not have a fixed point under any initial hypothesis of the extension of the truth predicate, if the only sort of vicious reference is the one that generates liar sentences; or it will be stable but with more than one fixed point, if the only sort of vicious reference is the one that generates truth-teller sentences; or it will be something that is a mixture of the two patterns, if it contains both sorts of vicious reference. Gupta believes therefore that his theory explains the pathologicality of a language.

3.3 The liar paradox and its revenge

3.3.1 The liar paradox

How does revision theory solve the liar paradox? First of all, the argument of the liar paradox, due to the modification of logic, is now ruled out. The unrestricted application of DfI and DfE in the original liar paradox is invalid and should be replaced by the restricted ones, DfI_r and DfE_r. In this case, the conclusion is not that the liar sentence is both true and false (which is a contradiction) but that the liar sentence is true at some stage of revision and is false at another. So if we accept the idea that circular definition is legitimate and at the same time accept Gupta's logic, then we need to acknowledge that the correct reasoning does not lead us to contradiction.

As to the semantic side, the theory acknowledges that there is no possible definite truth-value that can be assigned to the liar sentence. Its truth value has only a hypothetical character. We cannot say whether the liar sentence is true. We can only say that if it is true/false (at one stage of revision),[129] then it is false/true (at the next

[129] Being true/false at one stage of revision is just equivalent to saying that it is true/false under some hypothesis of the extension of the truth predicate.

stage). And this is just one special case of the big phenomenon—the semantics of circular definition. The signification of a circular concept is not a definite set, but only a revision function. We cannot say that some set is *the* extension of the concept. What we can say is just that, if some set is hypothesized as the extension of the concept, then some other (which could be the hypothetical one) will be the extension of the concept. For those sentences whose truth values behave stably in the revision sequence, Gupta says that we can indeed assert categorically what their truth values are (since their truth values in this case do not depend on any hypothesis of the extension of the circular concept). But for the liar sentence, this is not the case since its truth value never settles down in the sequence.[130]

3.3.2 And its revenge

Now Gupta's theory introduces a semantic concept, *categoricalness* and the original liar sentence is classified as being not categorical. By identifying the liar sentence as being not categorical, it does not mean that the liar sentence is not true, it only means that its truth value is not stable in the revision sequence of truth. And so there is no problem in calling the liar sentence as being not categorical. But then, as Gupta himself foresees (Gupta and Belnap, 1993, 253) one can construct the following revenge sentence, using exactly the concept we use to describe the original liar sentence:

R_1: R_1 is either not categorical or not true.

Suppose that R_1 is not categorical, then its first disjunct is satisfied, so its truth status will be the same after some point in the revision sequence of the truth predicate, regardless what the initial hypothesis of the extension of the truth predicate is. Thus by definition, R_1 is categorical. Suppose on the other hand, that R_1 is categorical, then since the first disjunct is not true and so contributes nothing to the truth status of the whole disjunction anymore, its truth status in the revision sequence will then be the same as the original liar sentence. Since the liar sentence is not categorical,

[130] For Gupta's own description of his solution to the liar paradox, see Gupta and Belnap (1993), p.254.

R_1 will be not categorical. By classical logic, we conclude that R_1 is categorical and not categorical. The theory of categoricalness seems to be inconsistent in this sense. This is the revenge paradox for the revision theory of truth.

Gupta has several replies to this issue. In his first reply, he points that the revenge argument uses the T-biconditionals (in its rule form) in a wrong way and what's more, it relies on a principle that is problematic:

> We believe, in the first place, that the argument in the strengthened paradox is not sound. Not only does it use the T-biconditionals in an uncritical manner (a manner that is called into question by our theory); it also relies on a principle, namely
>
> (2)All truths are categorical,
>
> that is clearly problematic. (Gupta and Belnap, 1993, 255)

It is confusing because it seems that, in the revenge argument we discussed above, there is no obvious need to apply the T-biconditionals nor there is any need to apply principle (2) in Gupta's description. When assuming that it is not categorical, we have the conclusion that it is categorical *because* its first disjunct is satisfied regardless of what the initial hypothesis of the extension of the truth predicate is. In Gupta's theory, it seems that this is sufficient to establish the claim that the sentence is categorical. Of course, one can choose to enrich Gupta's own description so that the argument can make use of the T-biconditonals. For example, we can reason in the following way:

If the revenge sentence is not categorical, then by classical logic, we can infer that the revenge sentence holds. And so by T-biconditional, we can infer that the revenge sentence is true. And since its truth does not depend on any initial hypothesis of the extension of truth, it follows that it will always be true after some point in the revision sequence. We could also attach an index to each step of reasoning:

(The revenge sentence is not categorical) ^{i}

By classical logic we have:

(The revenge sentence is not categorical or is not true)^{i}

By T-biconditional (in its rule form):

("The revenge sentence is not categorical or is not true" is true)$^{i-1}$

Since the index "i" is arbitrary, we come to the conclusion that there will be a stage after which the revenge sentence is true at any stage after *that* and it is so regardless of what the initial hypothesis of the extension of the truth predicate is. So by definition it is categorical. As to the principle (2), I see no application of it in the above reasoning.

Perhaps, what Gupta has in mind is another attack from the revenge sentence. That is, it may be suggested that if we consider the truth value of the revenge sentence we may come to the conclusion that it is both true and not true, provided that we apply T-biconditionals in a wrong way and presuppose principle (2). Here is how it goes.

Suppose that R_1 is true, then by the unrestricted T-biconditional, it will be either not categorical or not true. If it is not true, then we infer from its truth to its non-truth. Contradiction. If it is not categorical, here is where principle (2) comes in: by principle (2) we conclude that R_1 is not true. Either case, we infer that it is not true from its being true. Suppose, on the other hand, that R_1 is not true, then by the unrestricted T-biconditional, it is true after all. Thus as usual, we come to the conclusion that R_1 is both true and not true.

By denying principle (2) and the application of the unrestricted T-biconditional, Gupta can block the above reasoning. But this does not block the revenge argument in his original description. The revenge sentence still shows that it is both categorical and not categorical, which is a contradiction.

Gupta's second reply is that the above argument presupposes that the language at issue contains the semantic concept "is categorical in *L*", where *L* is the object-language whose semantics is at issue.

If the object language does not contain its own categoricalness predicate, then the revenge sentence cannot be constructed in the language, and so at least for this object language, there is no revenge problem. However, this also means that the

language cannot talk about its own semantics, at least, not all of it.

Now whether or not we can accept this result depends on whether or not natural languages are semantically self-sufficient. That is, whether they can talk about their own semantics *within themselves*. If natural languages are semantically self-sufficient, then one may argue that what Gupta provides is a theory that fails to explain the semantics of natural languages, for it fails to provide a linguistic model that can contain all its own semantic concepts. Tarski has a famous description of the nature of natural language, that natural language is universal.

> A characteristic feature of colloquial language (in contrast to various scientific languages) is its universality. It would not be in harmony with the spirit of this language if in some other language a word occurred which could not be translated into it; it could be claimed that "if we can speak meaningfully about anything at all, we can also speak about it in colloquial language". (Tarski, 1933, 164)

How to understand this universality is not without controversy. The strongest version is that, according to Gupta (Gupta and Belnap, 1993, 257), natural language is universal in the sense that every meaningful concept is expressible in it. This is not very plausible. A counterexample that I can think of now is real numbers. The set of real numbers is uncountable (which can be roughly regarded as meaning that there are more real numbers than natural numbers), and the set of terms that can be generated in English (through 26 English letters) is countable (which can be roughly regarded as meaning that there are at most as many English terms as natural numbers). So there must be some real numbers that cannot be denoted by English terms.[131] A weaker and more plausible version of "universality" is that whatever concept, as long as it can be expressed in some language, it can be expressed in natural language. Under this understanding, natural language must be a language that has the greatest expressive power. Any weakening of its expresibility would be unacceptable. Since Gupta does not deny that the concept "categorical in L" is expressible (in at least some language), he cannot deny that the concept can be

[131] For a proof on this point, see Boolos, Burgess, and Jeffrey (2007), p.10, example 1.13 and p.14 for its proof.

expressed in natural language.

Actually, Gupta does not insist that the concept is not in his object-language.[132] His main objection is that the concept "categorical in *L*" should not be treated as having an ordinary semantics and ordinary logic (Gupta and Belnap, 1993, 255). In his view, the concept just like the concept of truth, is circular. And so its semantics and logic should be treated in the same way as we treat the concept of truth. Let's for the sake of argument, accept this point, that the concept of categoricalness is circular. If so, then surely the above argument fails. For in this case, the index matters. According to Gupta, what the first half of the revenge argument establishes is nothing but that the revenge sentence is categorical at one stage and it is not at the next and a similar situation holds for the second half (Gutap and Belnap, 1993, 256). Thus we cannot derive the conclusion that the revenge sentence is both categorical and not categorical just like we cannot derive the conclusion that the original liar sentence is both true and not true.

So far so good. But I think there will still be a problem when we consider the semantic status of the revenge sentence. We all know that the truth value of the liar sentence is not stable in the revision sequence of the extension of the truth predicate. But it seems that the truth value of *this* statement itself is indeed stable in the revision sequence. So it is categorical to assert that the liar sentence is not categorical in the above sense. It seems that we just have the same certainty and desire to assert that the categoricalness of the revenge sentence is not stable in the revision sequence of the categoricalness predicate. But if we assert that the revenge sentence is not categorical, then we are once again entering into the revision sequence, and so what we are actually saying becomes that the revenge sentence is not categorical at some stage in the revision sequence of the categoricalness predicate.[133] But this is not what we wish to say, and so there is some categorical

[132] Gupta says that he can accept this point. See Gupta and Belnap (1993), p.255.

[133] Actually, Gupta does not specify exactly what we are saying, when we are calling a paradoxical sentence true/false. It may turn out that his original intension is to rule out completely the meaningfulness of this sort of truth ascription. Similarly, it is not clear what

fact that we cannot categorically assert in the current picture.[134]

To this, Gupta actually has a theoretical response. The categoricalness concept that is appropriate for describing the semantic status of the original liar sentence is not appropriate for describing the revenge sentence. To describe the pathologicality of the revenge sentence we need what he calls a higher-order categoricalness concept (Gupta and Belnap, 1993, 256), which, if my understanding is correct, is defined not in terms of the revision sequence of the truth predicate, but in terms of the revision sequence of the *original* categoricalness predicate. Let us call the original categoricalness predicate K_0, and the one that is appropriate for the revenge sentence R_1, K_1. Now it seems that we can categorically assert that the revenge sentence is not K_1 and do so without falling into the revision sequence of K_0.[135] The introduction of K_1 solves the problem of describing the semantic status of the revenge sentence just like the introduction of K_0 solves the problem of describing the semantic status of the original liar sentence.

However, if we follow Gupta's theory, then since we have now introduced a new semantic concept, it seems that a new revenge sentence can be constructed with the help of the new concept.

Thus we have:

R_2: R_2 is not K_1, or not K_0, or not true.

The new revenge sentence, once again, is pathological in the sense that its K_1ness is not stable in the revision sequence of K_1,[136] and the description of this fact

---

Gupta would say about calling a revenge sentence not categorical. But none of these matter at all. For however we interpret Gupta, the only thing that matters is that we cannot use these semantic concepts to correctly describe the semantic status of relevant paradoxical sentences.

[134] We can only assert it in a hypothetical way.

[135] Gupta does not provide a definition for higher-order categoricalness predicates, but following his method, we may defined a categoricalness predicate at order i as follows:
where i=0, then it is K_0
where $i \geq 1$, then a sentence is K_i if its K_{i-1}ness status always be the same after some point at the revision sequence of K_{i-1}ness predicate irrespective to the initial hypothesis of the extension of K_{i-1}ness, and is not K_i if otherwise.

[136] To see this point, consider the following reasoning: Suppose that R_2 is not K_1, then its first

requires the introduction of a new semantic concept defined in the parallel way we define K_1 and K_0. Let us call it K_2. Now we can correctly assert that the new revenge sentence is not K_2. This procedure can be carried out infinitely and produce a sequence of revenge sentences R_1, R_2, R_3…and a sequence of categoricalness predicates: K_0, K_1, K_2,… With a parallel reasoning, we can see that for any i, if R_i is the revenge sentence:

R_i: R_i is not K_{i-1} or not K_{i-2} or…or not K_1 or not K_0 or not true,

then K_i is the appropriate categoricalness concept for *describing the unstability of K_{i-1}ness of R_i*. And each time we introduce a categoricalness concept K_i, we can produce another revenge sentence:

R_{i+1}: R_{i+1} is not K_i or not K_{i-1}…or not K_1 or not K_0 or not true,

whose semantic status can be described by the categoricalness predicate K_{i+1}.

I have serval observations (not all of them count as an objection) to this solution. First, although the theory has to produce a hierarchy of categoricalness predicates, it is different from the Tarskian hierarchical solution to the liar paradox. The crucial difference is that, in a Tarskian hierarchical solution,[137] language at one level cannot contain its own truth predicate. Its truth predicate can only be found in a richer meta-language. But the above infinite procedure in Gupta's theory does not mean that the object-language cannot contain some semantic concepts in the sequence. Rather, as Gupta points out (Gupta and Belnap, 1993, 258), exactly

---

disjunct is satisfied regardless of what the hypothesis of the extension of K_0 and of the truth predicate are (since they are irrelevant in this case), so its truth value will be the same (and is true) after some point in the revision sequence of the truth predicate regardless of what the initial hypothesis of the extension of the truth predicate is. So it is K_0. And since it is K_0 regardless of what the hypothesis of the extension of the K_0ness predicate is, its K_0ness will be the same after some point in the revision sequence of K_0 (since we infer that it is K_0 regardless of what the extension of K_0 is), it is K_1. Contradiction. Suppose that R_2 is K_1, then its truth status will just like R_1. And since R_1 is not K_1, R_2 is not K_1. So we come to the conclusion that R_2 should be both K_1 and not K_1. So to resolve the contradiction, one can take K_2 to be another circular concept.

[137] Note that there may be some Tarskian hierarchical theory that may allow language of a level to contain the truth predicate for the language itself. But a study on this issue is clearly beyond the scope of this dissertation, so I will leave it here.

because of his theory, we can know that there is no problem in bringing in any one of the semantic concepts in the sequence into the object language. Or put it in this way, the addition of a truth predicate for a language into the language itself in Tarskian hierarchy will create inconsistency, but the addition of a categoricalness predicate into the language itself will not.

Second, the theory may not promise a language that is semantically self-sufficient. In the hierarchical picture above, given any stage of development, we see that the language will always lack an appropriate categoricalness predicate that is necessary for describing some revenge sentence in the language. In other words, it does not provide a linguistic model that shows us how a language can talk about its own semantics *in its full scale*. There will always be some sentences whose semantic status cannot be described in the language at that stage of development. Now this will not be a problem for those that reject the very idea that natural language is semantically self-sufficient. But there seems to be no obvious objection to this idea. Of course, as Gupta points out (Gupta and Belnap, 1993, 258), there is no obvious argument supporting this idea either. The thought that natural language is universal does not by itself imply semantical self-sufficiency because there is no obvious consensus on how to read this "universality" as mentiond. The kind of reading on "universality" that Gupta accepts is "Universal languages are possible in the sense that for any semantic concept C there is a language *L* that contains its own C-concept." (Gupta and Belnap, 1993, 258). Natural languages, under this reading are universal because they can always be extended to include more and more semantic concepts that are needed to describe their own semantics. But the process never completes.[138]

[138] This position assumes that natural language is finite. That is, natural language, however extended, has only a limited number of syntax. For example, at some particular time, it contains a finite number of terms. If this assumption is denied, if an infinite language is acceptable, then it may be suggested that perhaps we could have a natural language that contains all those categoricalness predicates, and so it would be semantically self-sufficient. However, my next two observations seems to show that this is problematic as well.

Third, the theory is in some sense *redundant*. It is redundant not in the sense that it contains some terms that are not necessary for describing the semantics of the whole language, but that it contains too many terms that are only necessary for describing the semantics of the set of pathological sentences. The original categoricalness predicate, which we label as K_0 is appropriate for describing the semantic status of sentences that involve no K_0. Actually, if it is not because of the revenge problem, we may think of Gupta's theory as dividing sentences into three groups: categorically true, categorically not true, and not categorical, which corresponds roughly to what we normally think of as assertable, not assertable and pathological. But now, in order to cover the semantic status of the infinite set of revenge sentences, we have to introduce infinitely many semantic concepts: K_1, K_2, K_3,…K_n… which can be understood only via Gupta's theory and each one of the categoricalness predicates at some point in the sequence is designed only for describing the pathological sentences which seem to be the by-products of the introduction of previous categoricalness predicates.

Fourth, it seems that there is some semantic concept that will still be a problem for Gupta's theory. We can as usual, following Priest's pattern for type S revenge (see chapter 1, section 2) to construct the concept, "*being non-categorical in some sense*". For a sentence to be *non-categorical in some sense* is for it to be either not K_0, or not K_1, or not K_2,… or not K_n,…, the sequence goes on until we exhaust all possible categoricalness concepts in Gupta's theory. Let us, for easy reference, call this concept, *absolute non-categoricalness*, and label it as non-K_c.

The technical way to achieve the definition of this concept is not important here. What is important is that, we do have a conception of this concept. Actually, the concept can be used to talk in a general way about the categoricalness status of all sentences in the language. One can use it to say, any sentence is either non-Kc or it is not non-Kc, that is, *any sentence is either non-categorical in some sense, or it is categorical in any sense.* If this concept is meaningful, then it seems that we can form the following revenge sentence (or proposition), which we may call the

absolute revenge sentence:

R_c: R_c is non-K_c or not true.

The problem occurs when we consider the semantic status of R_c.

R_c, when unfolded, is:

R_c: R_c is....not K_n or not K_{n-1} or not K_{n-2}...or not K_3 or not K_2 or not K_1 or not K_0 or not true.

which is just another revenge sentence like those we have constructed before. So presumably non-Kcness is just another circular concept and its truth status requires a higher-order categoricalness predicate to describe.[139] But we cannot find an appropriate categoricalness concept to describe the semantic status of R_c like what we did for all those normal revenge sentences before, for any such concept has been included as part of the concept non-K_c itself[140] and is appropriate only for describing the semantic status of revenge sentence formed via previous categoricalness predicates. We thus find ourselves in an awkward situation: we are in desperate need of an even higher-order categoricalness concept (since we believe that it is non-categorical in some sense), but there is simply no such thing, since

---

[139] One can try the following reasoning to see this point:
Suppose that R_c is not K_0, and it is non-K_c and so is true. Since its first disjunct is satisfied regardless what the hypothesis of the extension of truth is, its truth value will be the same after some point in the revision sequence of truth regardless what the initial hypothesis of the extension of the truth predicate in the revision sequence is, so it is K_0. Contradiction.

Suppose that Rc is not K_1, and so it is non-K_c and so is true. Since its first disjunct is satisfied regardless what the hypothesis of the extension of truth and the extension of K_0 are, its truth value will be the same after some point in the revision sequence of truth regardless what the initial hypothesis of the extension of the truth predicate is, so it is K_0. And since it is K_0 regardless what the hypothesis of K_0 is, its K_0ness will be the same after some point in the revision sequence of K_0, so it is K_1. Contradiction.

By similar reasoning, for any i, we can infer that if R_c is not K_i, then it is. So we come to the conclusion that R_c cannot be not K_0, not K_1, not K_2,…not Ki,… *so Rc is not non-Kc*. Suppose, on the other hand, that R_c is not non-Kc, then its truth status will be the same as that of the ordinary liar. And so it is not K_0, and so it will be non-K_c. So the non-$K_c$ness of R_c is not stable in the revision sequence of non-Kc.

[140] Any categoricalness concept is appropriate only for describing the semantic status of a revenge sentence constructed via previous categoricalness predicate. So if a concept, K_i is part of K_c, then it cannot be used to describe the semantic status of R_c. A revenge sentence cannot be described by the categoricalness concept that constitutes it.

non-K_c has exhausted all of them (so it is not *non-categorical in some sense*).

But what is the consequence of this fact? Recall what we said about the ordinary revenge sentence, R_1. We said that K_0 is not appropriate for describing its semantic status. Its semantic status is that its K_0ness is not stable in the revision sequence of the extension of the K_0ness predicate. Without the concept of K_1ness, we cannot describe this fact. Similarly, the non-$K_c$ness of R_c is not stable in the revision sequence of non-$K_c$ness. Without the appropriate concept of relevant categoricalness, we cannot describe this fact.[141]

To get around this problem, Gupta may suggest that either the absolute concept is not intelligible (whatever this mean) and should be rejected once and for all, or we need to acknowledge that there is some semantic fact about natural language that cannot be described by natural language on pain of being inconsistent. In the former case, we still require an independent reason to reject the legitimacy of the absolute concept, without which the theory is still ad hoc. In the latter case, it seems to fail to provide a semantic theory for natural language. In either case, the Dialetheist's Conjecture seems to hold here

4. Summary

In this chapter, we have seen three prominent approaches to the liar paradox. Each one of the approaches provides a different semantic-logical picture of our language. In the case of the Tarskian hierarchical approach, we are told that our language comes in a hierarchical structure, that the truth predicate in use is actually systematically ambiguous. In the case of Kripke's paracomplete theory, we are told that our language is a three-valued one—it is interpreted by strong Kleene

[141] A mysterious and fascinating thing is, it seems that by uttering "the non-$K_c$ness of Rc is not stable in the revision sequence of non-$K_c$ness", we are saying what is said to be unsayable. This consequence is strange but it appears to have an affinity with what I call the silence approach in chapter 1, section 3. It will be an important phenomenon in my future development of the silence approach. But given that my dissertation here is only an initial study on the silence approach, I will not go any further on this issue.

semantics and that the liar sentence has no classical truth value at all. In the case of the revision theory by Gupta and Belnap, we are told that if we accept the idea that notion of truth is actually a circular concept, then we can explain and understand its strange behavior in the liar paradox easily. These theories though provide profound insights in understanding our language, are not perfect. In addition to some special theoretical difficulties they face, the revenge problem disrupts them a lot. By using Priest's two patterns for generating the revenge paradox, one can easily construct an absolute revenge sentence for each theory which is extremely hard to deal with. So currently, we may say that the Dialetheist's Conjecture is not yet disproved.

Chapter 4 Silence approach

1. Silence position and its three main worries

In Beall (2007), we find that there is a possible solution which I will call the silence approach. The main idea is that, given that most of the attempts to solve the liar paradox fail, it seems we should give up the attempt to classify the semantic status (or truth category, in Beall's term) of liar-like sentences, and "whereof one cannot truly classify, thereof one must…be silent" (Beall, 2007, 4). Beall thinks that since this kind of approach does not engage in classification, that is, it remains quiet as to what is the truth value/semantic status of liar-like sentences, it does not have a chance to evoke the revenge problem, but "it offers no clear account of truth or the paradoxes at all", and thus little can be said about it. Given that Beall does not cite any work holding this position, it is very unlikely to characterize fully the nature of the approach that Beall has in mind here. But at least on the face of it, the two patterns we sketch in chapter 1 simply do not apply. If one stops doing the classification, or, is totally silent about it, then no new semantic status will be introduced, no new liar sentence will be formed and so no revenge problem occurs. It seems that here we have a promising solution to the revenge problem and the liar paradox.

But is it? Here are some worries.

First of all, one can clearly observe that while a silence position *may* be a sufficient solution to avoid the revenge problem, it is not by itself a solution to the pristine liar paradox. Normally, when we are searching for a solution to a paradox, what we are expecting is to find out which premise is unsound or which inferential step is invalid, or at the very least, to find out why the seemingly unacceptable conclusion is in fact acceptable, as what we have seen in the prominent approaches.[142] But clearly, the silence approach is not of this kind. Based on the above limited

[142] More on this see chapter 2.

description, it seems like an approach saying that, if you are doing a valid reasoning but it leads to a contradiction, then you give up doing that valid reasoning so as to *avoid* that contradiction. But this does not *eliminate* the contradiction, if there is one. Whether the truth status of the liar sentence is classified or not seems to have no direct connection to whether the pristine liar paradox can be blocked.

Secondly, the promise of avoiding the revenge problem is not very promising. In the Tarskian hierarchical solution, if we are allowed to talk about the hierarchy as a whole, we have a dilemma of having an inconsistent theory or an incomplete semantic theory for natural language; in Kripke's interpretation process, if we are allowed to talk about the interpretation process itself in the language to which the process belongs, then a similar dilemma follows. In the revision theory of truth, though we can avoid the problem by iteratively applying the theory, the absolute concept of categoricalness, which is the concept we need to talk about the categoricalness status of the whole language in general, remains a trouble. The parallel problem for the silence approach is that, if we should be silent about the semantic status of the liar sentence, can we talk about *this* fact? If we are allowed to have relevant concepts to describe the fact regarding whether the semantic status of a sentence can be classified, and if any sentence in a language is either *classifiable* or not, then we have a new revenge sentence. Consider the following one:

R-$_{\text{silence}}$: R-$_{\text{silence}}$ is either not true, or is not classifiable.

Suppose that R-$_{\text{silence}}$ is either true or false, then we have four situations:

(1) If it is classifiable

then its semantic status will be just like the pristine liar sentence which is not classifiable. Contradiction.

(2) If it is not classifiable,

then it is true, and it seems that we are classifying it as being true now. Contradiction.

We can also begin with the principle of bivalence and suppose that:

(3) If it is true, then we have:

it is either not true, or is not classifiable.

if it is not true, we have a contradiction,

if it is not classifiable, then it is true and not classifiable. And so, we classify it as being true, it seems that it is classifiable after all.

(4) If it is not true, then we have:

it is true. Contradiction.

Each option seems to give us a contradiction. One may try to fix the problem by iteratively applying the theory so as to produce more and more higher-order concepts of classification. For example, $R\text{-}_{silence}$ may be said to be *not classifiably classifiable*, in the sense that it cannot be classified whether its truth status can be classified. But any such iterative approach seems to fall in the pattern of type S revenge and so won't help. Another reply may be that if a sentence is *metaphysically* not classifiable (that is, it is such and so, as a matter of fact), then that means even to call it not classifiable should not be allowed. This may be a way out, but it requires a proper definition of "classification" which is not yet seen. In particular, it should take into consideration the fact that the above reasoning only needs to assume that the revenge sentence is either classifiable or it is not classifiable, which does not involve a firm, definite assertion like "it is classifiable" or "it is not classifiable".

Finally, the silence approach perhaps has more trouble in providing a non-ad hoc motivation for itself. Why the liar sentence is so mysterious that we should be silent about it? Merely saying that all other attempts fail is not a very good answer, for first, it still does not explain why the liar sentence is so mysterious and second, there seems to be other given-up options. For example, one may simply admit that the concept of truth is inconsistent. It is governed by inconsistent principles of use. Or if the revision theory is right about the circularity of the concept of truth, then why not take the absolute revenge sentence to be unclassifiable instead of the pristine liar sentence? These division cannot be settled if the defenders of the silence

approach cannot come up with a strong and independent justification for their own position.

My study on the silence approach is still on an initial stage, there remains quite a lot of questions to be answered. In this final chapter, I do not intend to provide a detailed, well-developed silence approach which can solve all the above problems. Instead, I wish to provide a primary characterization of this kind of approach, defining several key concepts and questions regarding it, and examine some current theories that may satisfy at least part of the characteristic features of it and see whether there is any hope for further development.

2. What is silence?

A characterization of silence approach is that the semantic status of the liar sentences are not classifiable, and so we should be silent about it. The central concept lies in the concept of *classification*.

Now to make the position clearer, we should ask, what exactly is *the classification of the semantic status* of a sentence here? And precisely what does it mean for the truth status of a sentence to be unclassifiable? What I have in mind about what it is to classify the semantic status of a sentence is simply to *assert* the semantic status of a sentence *according* to a certain classification theory, where a classification theory is a set of principles that we need in order to do the classification, which normally, is just the semantic theory for a language.

For example, in order to classify the semantic status of a sentence, there must be some categories of semantic status[143] *for* it. And we also need some truth condition

[143] The concept of semantic status is intuitively clear but hard to be precise. Here I can only provide some examples. In a classical setting, we hold the principle of bivalence, so that the category of semantic status contains only "true" and "false". In paracomplete theory, in addition to "true" and "false", there is a third value, "undefined". In the semantics for circular concepts, the semantic status is not a fixed value but a revision sequence, and the notion of categoricalness is the semantic concept describing the truth status (the truth value) of a sentence involving a

schema like "the proposition that p is true if and only if p",[144] so that when snow is white, we can, based on our acceptance of the instances of the above schema, classify the proposition "snow is white" as being true.[145] Of course, a complete classification theory can be more complicated. In the previous chapter we have seen several approaches to the liar paradox, each one of which can be regarded as offering a distinctive classification theory.[146] For example, under Tarski's hierarchy system (at least under the version of interpretation I give), the liar sentence is classified as simply false; under Kripke's paracomplete system, the liar sentence is, in effect, classified as neither true nor false;[147] under some revision theories, it is classified as not categorical in the sense that its truth status is never settled in the revision sequence of truth.

Three further points deserve clarification here. First, classifying the semantic status of a sentence is not the same thing as *assuming* it. The difference is manifested by the fact that one can assume the semantic status of a sentence without obeying the classification theory. For example, one can assume that "snow is white" is true, even though the required instance of the T-schema, " 'snow is white' is true if and only if snow is white" is not accepted. If this instance of the schema is the necessary principle for classification, then the classification cannot be done if it is rejected. One more example, one can certainly assume that a sentence is both true and false, which gives us a contradiction, but one cannot classify a sentence as both truth and false, if the underlying classification theory embraces classical logic and

---

circular concept in the revision sequence of that concept. Semantic status, perhaps can be characterized as some concepts that describing whether a sentence hold or not, and in which way it holds.

[144] Note that although here I use proposition as the truth bearer, it can be modified to fit other truth bearers as well.

[145] Or, if snow is not white, we can classify the sentence as being false accordingly, by means of the schema.

[146] I have to confess, at this moment, an accurate characterization of a complete classification theory is not available here, but the current description would be sufficient to show that there are indeed some principles that are necessary to do the classification and based on different principles, there could be different classifications on the same sentence, as those theories show.

[147] Strictly speaking, Kripke's theory does not classify the liar sentence as being *neither true nor false*, but *undefined*.

semantics—which in this case does not allow classifying a sentence as both true and false. Second, classifying the semantic status of a sentence is different from *knowing* the semantic status of the sentence, although in some sense they coincide. The former is more like a process in which we are reasoning according to certain rules of the classification theory and *deduce* the semantic status of a sentence, while knowing the semantic status of a sentence, in some sense, is just capturing the result of such a classification. Of course, if we take the semantic status of a sentence as an objective matter regarding the sentence itself, then, whether we can indeed know the semantic status depends on whether the classification theory is a *correct* one. Third, classification, one could say, comes in degrees, in the sense that some of them may be definite, some may be less definite or even very indefinite. If a sentence is true, then, the correct classification of it can be *either true or false*, or, *either true or false or neither true nor false*, and so on. In the most indefinite case, one can, of course, form a logical sum of every possible semantic status, say C, so that, as long as in his or her classification theory the targeted sentence is not without semantic status, he can classify it as C.

So much for the notion of classification. Now we turn to the question of what it is for a sentence to be unclassifiable. Given that a classification of the semantic status of a sentence is always based on a certain classification theory, the semantic status of a sentence is unclassifiable if the *accepted* classification theory is not applicable to the targeted sentence. That means, although we can use the classification theory to classify most of the sentences in our language, the same theory cannot be used to classify some sentences for some reason. Of course, given that classification comes in degree, we may find that although the theory cannot classify the status in a definite degree, it may still be able to classify it in an indefinite degree.

There could be different reasons/situations where an accepted classification theory is not applicable on a certain sentence. As we will see in the following section, the two candidate theories of the silence approach offer their distinctive rationales. For

now, we can talk about it generally. Since each classification theory can be regarded as constituted by a set of principles for classification, and a set of semantic statuses to which sentences are classified, there may be two ways to block the classification: Either the principles that are necessary for the classification simply do not apply to the sentences at issue or the category of semantic statuses simply does not fit the targeted sentence, and there is no way to expand the category. In these cases, we have no ability to do the classification.[148] In the first case, the set of category of semantic statuses are expressible in languages, the problem is just that we don't know which one in the set fits the targeted sentence; in the second case, the correct status of the sentence at issue is not expressible in languages at all.[149]

One last remark for the silence approach. It is both cognitive and metaphysical. If one cannot classify the semantic status of a sentence, then first, one cannot know the definite semantic status of the sentence, either it is one of the value in the semantic category but we cannot know which, or that it is some value, but which is beyond our cognitive capacity. Second, in both cases, it will still involve a claim about the metaphysical status of the truth value of the sentence. It does not say that the value is indeterminate, or that it has no truth value, or that its value is in some sense unstable. Rather, it acknowledges that the sentence has a definite semantic status, but it just gives up the attempt to locate it, since it believes that we just have no such capacity.

But if one classification theory fails to accommodate a sentence, does it mean that the theory is defective, or that we may simply choose to improve the classification theory by, for example, adding more principles or more members to the category of semantic status? To this, I have no definite answer at the moment, but an intuitive

[148] Recall that in Gupta's revision theory (see chapter 3, section 3), when it comes to the absolute revenge sentence, we eventually encounter the situation that there is no proper semantic notion that can be used to properly classify the absolute revenge sentence for the theory.

[149] I need to thank Graham Priest for pointing out this distinction between two sorts of "silence".

response is that, if we require for another classification theory, then what we are requiring is another conception/understanding of the semantics of our language. For example, if we are to change the classification theory from a Tarskian hierarchical picture to a Kripkean paracomplete picture, then our conception of natural language will change quite dramatically. In the former case, natural language comes in a hierarchy, and the truth predicate is a systematically ambiguous term. In the latter case, natural language comes in one piece, but it is gappy. The change of classification theory is therefore not random. It can only be done if one has reason to believe that the improvement can make the theory more descriptive than the previous one. The silence approach holds that, the *correct descriptive* semantic theory for natural language simply does not work for certain sentences. In other words, there is some intrinsic incapacity in our cognition that prevents us from knowing everything about the semantics of our language. At this moment, this is just a theoretical hypothesis, we may call it, *Silence Conjecture.*

The current characterization of the silence approach shows some advantages in responding to the three problems that we listed at the beginning of this chapter. For the revenge problem, a sentence saying of itself as being unclassifiable means nothing but that its semantic status cannot be determined by the accepted classification theory. A consistent way to treat the sentence is that it is unclassifiable, and nothing more. Since if it is unclassifiable, then the classification theory fails to classify it, we simply do not have the ability to further classify it as being true. As to the pristine liar sentence, first, we cannot classify it and so its semantic status is unknown in some degree; second, given the loss of relevant semantic principles that are necessary for classifying the sentence, the derivation from the liar sentence to the contradiction may be blocked (we will see how this is done later), so the concern about the metaphysical status of the liar sentence is also dispelled. Our previous ground for establishing the contradictory status of the liar sentence is undermined, and there is no new ground for establishing it. As a result, with the silence conjecture, the solution leads us to a silence position that we cannot classify the

semantic status of some sentences and so should be silent about it, (not that the silence position is the solution). As to the rationale, this is still a hard part, but in general, the theory suggests that we should be silent because we have no capacity to classify it due to the loss of relevant principles or relevant semantic statuses. As to why we have no capacity, different theories may offer different justifications.

In the rest of this chapter, I will make an attempt to interpret two current theories and see how these theories solve the liar paradox and lead us to a silence position. In particular, I will first examine the so called *semantic epistemicism* provided by Horwich (2005) and discussed by Armour-Garb and Beall (2005) and Asay (2015), and second, an approach can be properly called *exceptional theory*, which is advocated by Hofweber (2007) and Hofweber (2010), and see whether their theories indeed can give a satisfactory answer.

3. Horwich: Semantic Epistemicism

Strictly speaking, Horwich does not himself fully formulate the so-called *Semantic Epistemicism* as an approach to the liar paradox. His initial attempt is just to formulate a use theory of meaning (Horwich 1998a) and a corresponding theory of truth (Howrich 1998b). But given that his theory of truth embraces all the classical settings (for example, law of excluded middle, principle of bivalence and so on) and instances of the equivalence schema[150] and does not reject any form of self-reference, the theory thus seems to satisfy all the conditions one needs to generate the liar paradox. So he nevertheless needs to reply to the liar paradox.

One difficulty in presenting his solution is that, among his two theories, the solution to the liar paradox is only mentioned and is never worked out in any detail by himself. In his use theory of meaning (Horwich, 1998a), he identifies the liar sentence as one of the examples of *indeterminacy of everyday properties*, which is

[150] Equivalence schema: the proposition that p is true if and only if p.

the situation where it is indeterminate whether a predicate is applied to some object. More specifically, it is indeterminate whether "is true" is applied to the liar sentence for in this case we have a conflicting inclination--both to apply and withhold it. In his minimal theory of truth (Horwich, 1998b), one specific solution is mentioned, "…the only acceptable solution is…only certain instances of the equivalence schema [the proposition that p is true if and only if p] are correct" (Horwich, 1998b, 41). But does it mean that in Horwich's mind, those instances of the equivalence schema are not correct *because of* the indeterminacy they generate? May be, but in some other places, Horwich seems to give a completely different answer. "The intuitive idea is that an instance of the equivalence schema will be acceptable…as long as that proposition (or its negation) is *grounded*…" (Horwich, 2005, 81). There Horwich seems to hold the idea that an instance of the equivalence schema is not accepted because the sentence at issue is ungrounded. From these divergences in the text there are two interpretations of Horwich's theory to the liar. Armour-Garb and Beall (2005) develops Horwich's solution based on the first kind of position while Asay (2015) criticizes it based on the other. I have no intention to figure out which interpretation on Horwich is the correct one, so I will simply assume that they are both correct and present Horwich's solution based on their two interpretations.

3.1 Preliminary: the use theory of meaning and the minimalist theory of truth

In order to understand Horwich's semantic epistemicism, we will need some preliminary knowledge of his use theory of meaning and his minimalist theory of truth. The whole theories are highly comprehensive and sophisticated. For our current purpose, I will only introduce several points that are relevant to our discussion later.

A. Use property as meaning-constituting property

One of the main theses of the use theory of meaning is that, a term means what it means by possessing a certain *explanatorily basic* use property. By "explanatorily

basic", Horwich means that from these properties, every other use property of the term can be explained. These use properties, which are sometimes called regularities governing the overall use of a term, stem from what Horwich calls acceptance property. An acceptance property of a term can be roughly presented in the following way:[151]

Let x be a term, let A(x) be its acceptance property, then

> A(x)= our disposition to accept such and so sentences containing x in such and so situation.

For example, the acceptance governing our use of the term "and", as Horwich (1998a, 45) suggests (roughly), is:

> *We inline to accept the sentences "p and q" if and only we incline to accept the sentence p and accept the sentence q.*

Two terms have the same meaning property (that they mean the same thing), if they have the same use property (that their general use is governed by the same acceptance property).

The basic idea here is that, an acceptance property of a term, which is composed of a relatively small group of principles (certain sentences specifying our inclination to accept some sentences containing the term), will be able to explain everything (with other relevant facts) regarding the use of the term. It is the use property that has this explanatory power that Horwich would call explanatorily basic and it is the explanatorily basic use property that Horwich would take them to be the meaning-constituting property of a term (Horiwch, 1998a, 60). And since each word *may* have a unique use property (though it may be relevant to others), which is constituted by a unique acceptance property, there is no uniform analysis of meaning relation "x means y-ness". What we have is for each word, an acceptance property that can account for its meaning what it means.

[151] For Horwich's original characterization, see Horiwch (1998a), p.45.

Note that, by saying that the use property of a term is its meaning-constituting property, Horwich does not mean that meaning *is* use. What he means is just that the use property constitutes the meaning property. This constitution relation will hold between two properties, according to Horwich, when the following two conditions are satisfied (Horwich, 1998a, 25):

> Let a(x) be the property being analyzed and let u(x) be the property that is said to constitute a(x), then u(x) constitutes a(x) if and only if:
>
> First, "a(x)" and "u(x)" apply to the same thing, and;
>
> Second, facts about "a(x)" can be explained by the first fact.

One example Horwich provides is the property of being water and the property of being composed of H_2O molecules (Horwich, 1998a, 25). Every fact regarding things that are water will then be explained by the fact that they are composed of H_2O. For Horwich, the relation between meaning property and use property are the same. Every fact regarding a term's meaning what it does can be explained by the fact that it has such and so explanatorily basic use property. And it seems that Horwich insists that this constitution property does not require the identity of the property being constituted and the property that constitutes it.

B. Minimal theory of truth

Normally, there is a distinction in theory of an existing entity and theory of the linguistic term expressing that entity. For example, a theory of the substance water may specify some basic fact about water, say, it is a kind of substance constituted by H_2O, from which all facts regarding water can be explained. For example, it may explain why water is liquid at room-temperature, or that it is transparent based on the fact that it is composed of H_2O molecules. A theory on the linguistic term "water" specifies the basic fact about the English word, from which all its linguistic features are to be explained. For example, it is noun referring to some such substance in the world. And it seems to have no necessary connection (though it could have) between the nature of the substance water, and the way we use the term "water"—we have been using the term long before we discover that water is composed of

H_2O molecules.

However, when it comes to the property of truth and the truth predicate, Horwich suggests that they somehow coincide on the set of instances of the *equivalence schema*:

The proposition that p is true if and only if p.

The instances of the equivalence schema (which is just a version of the T-schema we encountered in Tarski's theory) constitute the basic facts of the property of truth. And correspondingly, what we mean by the truth predicate, that is, the meaning property of the truth predicate, is constituted by the acceptance property that we incline to accept instances of the equivalence schema. Moreover, the central view of his theory of truth, and which is why his theory has been called minimalism, is that, the set of instances of the equivalence schema (together with facts of relevant objects other than truth) is sufficient to account for *every* fact regarding truth and the truth predicate. Nothing more needs to be assumed.

So much for Horwich's theory of meaning and truth. In next section we will see Horwich's solution to the liar paradox.

3.2 The solution

As mentioned, Horwich's theory of truth implies that we should accept instances of the equivalence schema, which, in conjunction with classical logic and the referential apparatus of natural language, is sufficient to generate the liar paradox. Horwich himself suggests a solution from the point of view of the Use theory of meaning and the Minimal theory of truth. That is, while the meaning regarding the truth predicate is constituted by (the acceptance of) the instances of the equivalence schema, not all of them are acceptable. More specifically, given a liar sentence:

l: *l is not true*,

the following instance is not acceptable:

T-*l*: *l* is true if and only if *l* is not true

Recall that in the argument of the liar paradox, one key step is to apply the T-schema (or in this case, the equivalence schema) to the liar sentence and so to produce something like T-$_l$. Thus if T-$_l$ is unaccepted, then the argument of the liar paradox is undermined. No contradiction can be derived from Horwch's theory of truth.

Note that the liar sentence in this case can still be either true or false, as Horwich argues "…there is no contradiction in supposing that D [the liar sentence] is true (or in supposing that it is false): the problems arise only if the equivalence schema were to be applied" (Horwich, 2005, 82, fn10). We may say in another way, that given the universal validity of the law of excluded middle and the principle of bivalence, one should assume that the liar sentence is either true or false (one can say, that this is a kind of classification, but in an indefinite degree), but one can never know which one it is, for the very key principle, the equivalence schema in this case is not applicable.[152] Here, as Armour-Garb and Beall (2005, 91) suggests, Horwich holds the principle that if it is impossible to deduce a sentence to be true through the equivalence schema, then it is impossible to know that sentence is true.[153] In other words, the equivalence schema is one of the necessary principles for truth-status classification, at least for Horwich. So restricting the schema also prevents one from being able to classify the truth status of the liar sentence.

Interpreting the theory in this way it perfectly fits the general feature of the silence approach—since we lack the necessary principle to do the classification of the truth status of the liar sentence, it is unknowable. But what about the three main problems

[152] A short illustration: in most of the cases, the equivalence schema is applicable—from what is the case, we get what is true. For example, from the fact that snow is white, we get that "snow is white" is true. But when the schema is not applicable, then from what is the case, we cannot simply infer what is true. In the case of the liar sentence, the instance of the schema, this sentence is false is true if and only if this sentence is false would be unacceptable. So one cannot get this sentence is true, from this sentence is false. Of course, it would be questionable whether, for all classification theory, the equivalence schema is necessary, but at least in Horwich's theory, it is necessary. And if we take his theory as a correct one, then under this framework, there is no way to do the classification.

[153] One justification for this interpretation to Horwich is that his minimalist theory of truth requires that every truth-related fact will require only the instances of the equivalence schema to explain. Anything more than those instances would be truth beyond his minimalist truth.

regarding the silence approach that we are worried about at the beginning of this chapter?

The first problem is how the silence approach prevents the original liar paradox. The answer is, we can assume that it is either true or false, but since the key instance of the equivalence schema for the liar sentence is not accepted, we can never know which truth value it is. But in either case, since the instance of the equivalence schema is lost, contradiction cannot be derived.

The second question is whether it really can block the revenge problem in the same way it blocks the pristine liar paradox. In particular, how can a silence position deal with the following revenge sentence:

$R\text{-}_{silence}$: $R\text{-}_{silence}$ is either not true, or is not classifiable.

One suggestion is, $R\text{-}_{silence}$ is indeed not classifiable. In Horwich's semantic empistemicism, being not classifiable means nothing more than that we cannot classify the semantic status of the sentence given the loss of the instance of the equivalence schema for the sentence. If the sentence is not classifiable in this sense, then that means the following instance of the equivalence schema is not accepted:

$R\text{-}_{silence}$ is true if and only if $R\text{-}_{silence}$ is either not true or not classifiable.

Again, if this principle is lost, one cannot derive that $R\text{-}_{silence}$ is true from its being not classifiable. Thus the possible contradictory situations that we list at the beginning of this chapter are resolved. Now of course, if we can assert that $R\text{-}_{silence}$ is not classifiable, then surely we can assert that $R\text{-}_{silence}$ is either not true or not classifiable. This means that we can assert $R\text{-}_{silence}$ while at the same time reject to classify it as being true. Given the closed affinity between truth and assertability, one may think that this rejection is self-refuting—if one can assert a sentence, then he must know that the sentence is true, and therefore be able to classify it as being true. This is indeed a crucial issue, and I think the answer of which requires a theory both of truth and assertability which surely is not appropriate to be carried out here. My current reply is that, this divergence between truth and assertbaility may be

exactly what the liar paradox teaches us, just like Gödel's theorem teaches us that in a strong enough axiomatic system of arithmetic, the notion of truth and the notion of provability may not coincide—there is some truth that is not provable or that there is some falsity that is provable. But of course, I do not take this to be the final answer, since it is just an initial theory that requires further development.

The only problem remains is the rationale for such a solution. Exactly based on what reason can Horwich preclude some instances of the equivalence schema from his minimalist theory of truth? But his attitude toward why we should/can reject some instances of the schema is far from clear. On this point, Armour-Garb and Beall (2005) and Asay (2015) have different views. And we will discuss their expositions on Horwich's rationale in turn.

A. Indeterminacy of meaning

In Horwich (1998a), Horwich distinguishes several kinds of indeterminacy in meaning. One special kind of indeterminacy concerns with vagueness and paradox. Theoretically, since our use of a term is governed by some sort of regularity, then it is possible that in some situation given an object and a predicate, first, the regularity does not determine whether it applies to that object and second, whether it does not apply to that object. Furthermore, none of the outcome can be determined no matter what further discoveries are made (Horwich, 1998a, 64).

Let us call the above three conditions, *indeterminacy conditions*. There are two situations when the indeterminacy conditions are satisfied. First, when the regularity yield *no* inclination either to apply or not to apply the predicate. Horwich believes that this is what we meet when vagueness occurs. Second, when the regularities of some terms yield *conflicting* inclinations, that is, inclinations both to apply and not to apply the predicate. Note that, when some regularities yield conflicting inclinations, this does not mean that those regularities will always conflict with each other. It happens, in Horwich's word "…when the predicate is

normally used within a certain restricted domain and when the simplest regularity that would accommodate that practice, if it were extended beyond that domain, would conflict with other use regularities" (Horwich, 1998a, 64). Horwich believes that semantic paradox, especially the liar paradox, is the consequence of this situation.[154]

As far as I know, Horwich does not specify in details exactly which regularities are conflicting with each other in the case of the liar sentence. All Horwich says is that, as is quoted above, they will conflict with other regularities when they are extended beyond their normal domain. Armour-Garb and Beall (2005) develops Horwich's idea by specifying the other party that creates this conflicting inclination: "…the regularity underlying our use of 'not' is incompatible with that underlying our use of 'true' "(Armour-Garb and Beall, 2005, 92). This development is supported by Horwich's description on the regularity governing[155] the notion of "not", which can be roughly presented by the following three principles (Horwich, 1998a, 72):

(1) If p then not-not-p.

(2) If not p then not p.[156]

(3) p or not p.

The relevant schema (together with T-$_l$) that gives rise to conflicting inclination is

---

154 One thing must be clarified. Though Armour-Garb and Beall do not mention this, it seems that the superficial phenomenon of indeterminacy here is that, there will be some situation where we just don't know whether a term applies or not applies to a given object. And the underlying phenomenon explaining this situation is either we have *no* inclination to apply or not to apply the term (in this case, vagueness occurs) or we have *conflicting* inclinations both to apply and not to apply the term (in this case, contradiction occurs). So for Horwich, the consequence of both phenomena is that, it is indeterminate whether a term is applicable or not applicable to the given object. That is why Horwich emphasizes that "…just as it is indeterminate' whether a certain vague predicate applies, or does not apply, to a certain borderline case (although certainly it does or does not), so (and for the same sort of reason) it is indeterminate whether D [the liar sentence] is true or whether it is false." Horwich, (2005, 82, fn10).

155 Note that what Horwich says about these schemas is that the term "not" can be, or at least can partly be *implicitly defined* by the three schema. And it seems that this means that the meaning property of "not" is at least partially constituted by the acceptance property that we incline to accept instances of the three schemas above.

156 If my understanding of Horwich's theory is correct (though I cannot be sure), one can get (2) by eliminating the truth predicate from the semantic principle for negation: if "~A" is true then it is not true that A. By eliminating the truth predicate from the conditional, we have that if ~A then ~A.

perhaps (3). The corresponding instance for the liar sentence is:

N-l: it is the case that l or it is not the case that l, which is:

l is not true or l is true.

With T-l: l is true if and only if l is not true, we can derive that l is true and is not true.

When this conflicting inclination occurs, that is, when the regularity of the truth predicate gives rise to contradiction, Armour-Garb and Beall claim that our inclination to accept the relevant instance of the equivalence schema will be *overridden* and so won't be accepted. More specifically, T-l is *precluded* from the meaning-constituting regularity of the truth predicate.

B. Ungroundedness

Horwich(2005) mentions another way to rule out the problematic instances of the equivalence schema:

> The intuitive idea is that an instance of the equivalence schema will be acceptable, even if it governs a proposition concerning truth…as long as that proposition (or its negation) is *grounded*—i.e. is entailed either by the non-truth-theoretic facts, or by those facts together with whichever truth-theoretic facts are 'immediately' entailed by them (via the already legitimized instances of the equivalence schema), or…and so on (Horwich, 2005, 82).

The notion of groundedness here is the one that Kripke (1975) defines. By "non-truth-theoretic facts", Horwich means roughly that facts that involve no truth (or other semantic notions). A grounded sentence thus is a sentence whose truth can be traced back to non-truth-theoretic facts.

Asay therefore identifies Horwich's rationale for ruling out some instances as that the relevant sentences at issue are ungrounded:[157] "This distinction between grounded and ungrounded propositions is the basis of Horwich's explanation of

[157] For example, if p is a sentence at issue, and the instance of the equivalence schema for it is "p is true if and only if p", then under this rationale, the instance will be precluded from the meaning of truth if the sentence p is ungrounded.

why ($T_{(L)}$) may be dismissed, and so is an essential component of the epistemicist's solution to the paradox." (Asay, 2015, 685).

In summary, Horwich's semantic epistemicism, at least, when exposited in the above ways, does provide a solution that fits at least part of the feature of the silence approach. It does not deny that the liar sentence has a determinate truth value. It denies that its truth status can be classified for it rejects the principles that are necessary for making that classification. As a result, although one can assume consistently that the liar sentence is either true or false, one cannot know which one it is. Furthermore, at least on the face of it, it can answer the three problems that we listed at the beginning of this chapter. It definitely blocks the pristine liar paradox; it can treat the revenge sentence formed via the notion of classification in the same way it treats the pristine liar sentence; and at least two sorts of rationale are justifying its rejection of the instances of the equivalence schema for the liar sentences.

However, just like many other solutions to the liar paradox, it faces its own problems as we shall see in next section.

3.3 **Critique**

Critiques to Horwich's semantic epistemicism can be roughly divided into two kinds. The first kind of critique questions whether it is compatible with Horwich's minimalist theory of truth. This is important because it affects the tenability of some of the positions Horwich holds on the liar sentence. The second kind of critique questions more about the solution itself. That is, whether simply blocking the relevant instances of the equivalence schema can really avoid the revenge problem and whether the rationales for these rejections (of the instances of the equivalence schema) are acceptable. Below we will discuss them in turns.

A. Can one be a semantic epistemicist on the liar paradox and at the same time a minimalist on truth?

If one is a semantic epistemicist on the liar paradox, then he or she will hold the following positions:

A-1 The liar sentence is either true or false.

A-2 T-l (l is true if and only if l is false) is not acceptable.

A-3 There is no way to know which truth value (in the two) that the liar sentence possesses.

If one is a minimalist on truth, then he or she will hold the following positions:

A-4 The meaning of the truth predicate is exhausted by the instances of the equivalence schema:

The proposition that p is true if and only if p

A-5 Every truth-related fact can be explained by the instances of the equivalence schema together with facts of other relevant entities.

Asay (2015, 682) points out that A-2 is cagey description. T-l is not only unacceptable, but should be false. For if it is not false, then it is true (principle of bivalence). And if so, we can apply it freely in any reasoning, including the liar paradox. So T-l must be false.

If so, then the sentences "l is true" and "l is false" will have different truth values (given the semantics for biconditional). Given A-1, l is either true or false, if "l is false" is true, then "l is true" will be false, and vice versa. Moreover, iteratively applying the truth predicate on l will keep reversing the truth value of the resulting sentences.

Thus according to Asay,[158] we have:

(1) l is false

[158] See Asay (2015), pp.683-684, for a more complete description of the behavior of the truth predicate in these and the following groups of sentences.

(2) "*l* is false" is true

(3) " '*l* is false' is true" is true

(4) " ' ' '*l* is false' is true' ' is true" is true

…

The pattern is that (1) will always have the same truth value as the following form of sentences:

"*l* is false" is true…is true…

(with 2n occurrences of the truth predicate, for any $n \geq 0$)

and will have reverse truth value as the following form of sentences:

"l is false" is true…is true…

(with 2n+1 occurrences of the truth predicate, for any $n \geq 0$)

Another relevant phenomenon is that sentences in the following series have the same truth value:

(1) *l* is false

(2) "*l* is false" is false

(3) " '*l* is false' is false" is false

(4) " ' '*l* is false' is false' is false" is false

…

One important observation is that, the pattern of behavior of the truth predicate for the liar sentence is the pattern of behavior of the falsity predicate for normal sentences; and the pattern of behavior of the falsity predicate for the liar sentence is the pattern of behavior of the truth predicate for normal sentences.

For example, when it comes to normal sentences, say, "snow is white", we will keep reversing the truth value of the resulting sentence by iteratively applying the falsity predicate on the sentence "snow is white":

(1) snow is white

(2) "snow is white" is false

(3) " 'snow is white' is false" is false

(4) " ' 'snow is white' is false' is false" is false

…

and similarly, by iteratively applying the truth predicate on the sentence "snow is white", we will get more and more sentences with the same truth value:

(1) snow is white

(2) "snow is white" is true

(3) " 'snow is white' is true" is true

(4) " ' 'snow is white' is true' is true" is true

…

The conclusion is that, the truth predicate and the falsity predicate *behave in the reverse way when it comes to the liar sentence, in contrast to their behaviors in normal sentences.*

Even though they behave strangely when it comes to the liar sentence, Asay points out that they nevertheless *behave* and this is a distinctive *truth-related fact* that requires explanation but which cannot be explained simply by the instances of the equivalence schema. These truth-related facts, according to Asay, can be explained by the instances of the following schema, which for easy reference, I will call the F-schema:[159]

F-schema: *The proposition that p is false if and only if p*

The behavior of the falsity predicate in the case of the liar sentence is obviously obeying the F-schema. As to the truth predicate, note that being true in this case is to be not false (given that Horwich accepts all classical setting). So, to say that *l* is true is to say that it is not the case that *l* is false, this accounts for why "*l* is true" has reverse truth value as "*l* is false".

Thus in Asay's analysis, A-1 and A-2 are not compatible with A-4 and A-5. Since the above truth-related facts are derived from A-1 and A-2, but they will require

[159] Asay's original terminology is "F-biconditionals" which refer to the instance of the F-schema here. See Asay (2015), p.690.

something more than A-4 promises in order to get explained, and this surely violates A-5.

Does this argument work? It all depends on whether the new truth-related facts that Asay points out can be explained with the resources that are *available* to a minimalist. And I think so.

Horwich may reply simply by pointing out that the strange behaviors of the truth predicate and falsity predicate in the cases of the liar sentences can actually be explained by something else other than a theory of truth.

The identity in truth value of sentences obtained by iteratively applying the falsity predicate on the liar sentence can be simply explained by law of identity. That is, if object A and B are identical, then they will have the same property.

With a little reflection, one can easily see that the following sentences are all identical to the liar sentence:

(1) *l* is false

(2) "*l* is false" is false

(3) " '*l* is false' is false" is false

(4) " ' '*l* is false' is false' is false" is false

…

for the name of the liar sentence (1) is "*l*", that is *l*= "*l* is false". So we can replace every occurrence of "*l* is false" with *l*. Thus (2) is nothing but "*l* is false"; (3) can first be transformed into (2) by replacing the most-inside name "*l* is false" with *l* and then be transformed into (1); and similarly for the rest of the sentence in the sequence.

Thus it can be shown that by iteratively applying the falsity predicate on the pristine liar sentence, one only gets more and more sentences that are identical to pristine

liar sentence. By law of identity, they have the same truth value.

As to the strange behavior of the truth predicate in the cases of the liar sentence, it can be explained by both the law of double-negation and the law of identity.

The Law of double-negation can be roughly characterized as that if you negate the same sentence twice, then you get a sentence that is extensionally equivalently to the original sentence. For example, the sentence that *it is not the case that it is not the case that snow is white* is equivalent to the sentence *snow is white*.

Since in minimalist theory of truth, to call a sentence true is to assert that it is not the case that the sentence is false, we can transform each truth ascription on the liar sentence into the form of negating its falsity.

Thus we have the following series of truth-ascriptions:

(1) *l* is false

(2) It is not the case that "*l* is false" is false

(3) It is not the case that it is not the case that " '*l* is false' is false " is false

(4) It is not the case that it is not the case that it is not the case that " ' '*l* is false' is false' is false" is false

…

By replacing each name "*l* is false" with *l*, we obtain the following series of sentences:

(1) *l* is false

(2) It is not the case that *l* is false

(3) It is not the case that it is not the case that *l* is false

(4) It is not the case that it is not the case that it is not the case that *l* is false

…

By law of double-negation, for example, (1) and (3) are equivalent, (2) and (4) are equivalent. With a little reflection one can easily observe that this pattern goes for the rest of the sentences in the series as well.

So from a minimalist point of view, the strange behavior of the truth predicate and the falsity predicate do not require one to have a theory of truth more than the minimal theory of truth to explain and so do not violate the basic commitment of the minimalist theory of truth. Asay's argument therefore fails.

A better critique to the compatibility of semantic epistemicism and minimalist theory of truth focuses on A-1 itself. Both Armour-Grab and Beall (2005) and Asay (2015) recognize this problem, with silently different descriptions. Below I will summarize the main idea of this problem.[160]

Given A-1, the liar sentence is either true or false. Given A-3, we cannot know which one it is. But even though we cannot know the truth value of the liar sentence, it nevertheless *has* a truth value.

So let us first assume that it is true. Now consider the following assumption:

"*l* is false" is true.

The question is, how can we explain this fact (if it is)? By A-4 and A-5, if this fact can be explained, then it must be explained by the relevant instance of the equivalence schema for *l*. But given A-2, there is no such instance that is acceptable. So if the liar sentence is true, then one cannot be a semantic epistemicist and at the same time a minimalist of truth.

Suppose on the other hand, that:

"*l* is false" is false.

By exactly the same reasoning above we conclude that if the liar sentence is false, then one cannot be a semantic epistemicist and at the same time a minimalist of truth.

Either way, if A-1 holds, then one cannot be a semantic epistemicist and at the same

[160] For their original descriptions of this problem, see Armour-Grab and Beall (2005), pp.92-93, and Asay (2015), pp.688-689.

time be a minimalist of truth. I'm not sure how Horwich would reply to this issue. It may be suggested that A-1 can be given up. But this seems to acknowledge truth value gap, and it seems that we will then know that the liar sentence is determinately neither true nor false, and this truth-related fact seems to be something beyond the explanatory power of the minimalist theory of truth. Or one may suggest on the other hand, that we can give up A-5. That is, although the meaning of the truth predicate is exhausted (A-4) by the accepted instances of the equivalence schema, there is still something that we cannot explain regarding truth, in particular the truth *or* falsity of the liar sentence cannot be explained. But this amounts to giving up the basic commitment of minimalism of truth. It seems that this is indeed an intrinsic inconsistency between semantic epistemicism and minimalist theory of truth.

B. Revenge: the paradox of the knower?

In Horwich's solution, not only the classification of the truth status of the liar sentence cannot be legitimately made, but also the exact truth value of the liar sentence is *unknown* to us. Given that it evokes unknownability, Armour-Garb and Beall (2005) suggests that it may cause another paradox known as the *knower paradox*. Consider the following sentence:

S: Nobody knows S

Now, according to the solution provided by semantic epistemicism, the instance of the equivalence schema for S would be unacceptable, therefore we cannot classify the truth status of S. But S is either true or false after all (based on the general setting). Therefore, the theory is equal to asserting a conjunction: nobody knows S is true and nobody knows S is false. Now, consider the first conjunct. If nobody knows S is true, then we can infer nobody knows S.[161] But given that Horwich (and anyone who believes in his theory) knows *this* fact, Horwich (and anyone who sees

[161] For illustration: if nobody knows snow is white is true, then nobody knows snow is white. Some may think that this is not necessary, for example, to those who does not understand the notion of truth, he or she will know that snow is white, and does not know that "snow is white" is true. But we can always restrict the domain of "nobody" to those who understand the notion of truth. In any case, this inference does not require general validity.

this) knows exactly S. So we have a contradiction again. Because of this, Armour-Garb and Beall argues that semantic epistemicism cannot be free from the revenge problem.

To this, my reply is that, it is sure that semantic epistemicism will lead to the unknowability of the truth status of the liar sentence, but the knower paradox is not *evoked* by the approach.[162] For the notion "unknown" is not introduced by the approach, but one that we have come to know for a long time. One can clearly see, even though we do not hold semantic epistemicism, the knower paradox remains there. But let us, for the sake of argument, assume that the knower paradox is indeed evoked by the approach. Even so, whether this problem will damage the approach still depends on two more factors.[163] First, it depends on whether the liar paradox (which concerns the concept of truth) and the knower paradox (which concerns with the concept of knowledge) share the same or at least a similar structure, so that the knower paradox can be regarded, in some degree, as a liar paradox. Second, it depends on whether the semantic epistemicism offers a uniform solution to both paradoxes. Armour-Garb and Beall pointed out that the liar paradox shares the same structure with the knower paradox based on Priest (2002),[164] and that the specific solution semantic epistemicism offers to the liar paradox (namely, by restricting the equivalence schema) cannot be applied to solve the knower paradox. However, they

---

[162] By "evoked", I mean that the knower paradox is not generated because of our commitment to Horwich's semantic epistemicism. Unlike the previous revenge paradoxes, the knower paradox will not disappear, simply by rejecting the relevant theories that are supposed to cause the revenge by introducing new semantic concepts. Whether one accepts Horwich's theory or not, one will still have to solve the knower paradox, if one wishes to have a consistent theory on knowledge. In this sense, the knower paradox seems to be independent from Horwich's solution. It seems to me that therefore the knower paradox may not require the same kind of solution that Horwich gives to the liar paradox. But of course, this requires further study but this is beyond the scope of this dissertation and I leave it for future study.

[163] As mentioned, the revenge problem is a kind of problem caused by the introduction of the solution to the pristine liar paradox. Since semantic epistemicism may be regarded as classifying the *epistemic* status of the liar sentence as unknown, it in this sense can be regarded as introducing the very paradox. But some more specific condition should be satisfied as the two that we are about to deal with.

[164] See Priest (2002), chapter 10. Since a discussion on the structural similarity is beyond the scope of this dissertation, I will simply assume that Armour-Garb and Beall's assumption is correct.

neglect one important point. That is, while the liar paradox and the knower paradox are structurally similar, they are similar only at a *higher abstract level*, and the solution that we have been discussed is at a relatively lower abstract level. As Smith (2000) argues, "one cannot argue from the fact that two paradoxes have the same structure at a certain level of abstraction, that their solutions should be the same at a lower level of abstraction" (Smith, 2000, 118). What is needed is perhaps only that, while the solution provided by semantic empistemicism solves the liar paradox at a specific level of abstraction, the same *kind* of solution can be applied to solve the knower paradox.

The solution by restricting the equivalence schema, "the proposition p is true if and only if p" cannot solve the knower paradox, for the knower does not rely on this equivalence schema. But the knower paradox has its own similar schema. Reflecting on the knower argument, one can easily get the following inferential rule:

If nobody knows S is true, then nobody knows S

The converse, if nobody knows S, then nobody knows S is true, although is not directly applied in the argument we sketch above,[165] is also very plausible, at least,

[165] Note that though the converse is not applied in the knower paradox we just sketched, but we may try to extend it a little bit:

1. S: nobody knows that S is true. (premise from semantic epistemicism)
2. Nobody knows S. (applying knowledge-schema from right to left)
3. Horwich knows nobody knows S. (a plain fact)
4. Horwich knows S (from S is the name of nobody knows S)
5. Horwich knows S to be true. (applying knowledge-schema from left to right)
6. Somebody knows S is true. (from existential derivation)

Only 6 contradicts with 1, so the application of the knowledge schema from left to right is also applied.

Some may argue that the knowledge-schema may not hold in general. The instances of the equivalence schema may not hold true in a particular situation, because, say, the knower does not understand the truth predicate, or he or she is not intelligent enough to infer from S to S is true. If so, then the knower paradox does not arise at the outset and so there is no problem for semantic epistemicism. But this seems to be not the case. Even if the inference does not generally hold given that some may not understand the notion of truth, we can restrict the domain of "nobody" to those that are capable of understanding the truth predicate to maintain the validity of the inference. In any case, I do not want to (or need to) defend the legitimacy of the knowledge-schema.

in some of the cases. So we have the following schema:

Nobody knows S is true if and only if nobody knows S[166]

One can, call this schema, *knowledge*-schema, while, call what is originally called equivalence schema, *truth*-schema. And call them, in a general way, *equivalence-schema*.[167] Although it is not an easy job to make precise what Smith calls "*levels of abstraction*", it seems quite intuitive that the notion of *knowledge*-schema and that of *truth*-schema are of the same abstraction level and are all less abstract than the general notion, *equivalence-schema*. Rejecting the instance of the truth-schema for the liar sentence only blocks the liar paradox, and rejecting the instance of the *knowledge*-schema only blocks the knower paradox. The solutions described in this way is specified by the particular problems they tackle with. Just like when we describe the liar paradox in a specific way it never be like the knower paradox—the two paradoxes only similar to each other when abstracted in a higher level. Similarly, the two solutions are not similar to each other in a specific description, but they are similar when abstracted at a higher level—one can properly say, the solution to both the liar paradox and the knower paradox is to restrict some instance of the *equivalence-schema*.[168] In this sense, it is indeed possible that semantic epistemicism can offer a general solution to both *kinds* of paradox and avoid the charge of the revenge problem.

Note that, all I have established is only that it is theoretically possible for semantic epistemicism to offer a general solution both to the liar paradox and the knower paradox. I'm not saying that the solution is well-justified or motivated and is free from other charges like being ad hoc. If Horwich would like to adopt this solution, then he will need to provide a theory of knowledge, which may take the *knowledge*-

[166] A more general way to state the schema is possible, for example: S is not known to be true if and only if S is not known, or S is known to be true if and only if S is known.

[167] To distinguish what we previously call "equivalence schema", here I use the italic form and connect the two terms with a hyphen.

[168] It should be quite intuitive that the two schemas are structurally similar to each other, and play a similar role in generating the paradoxes. Of course, further analysis on how they can be subsumed under a category is needed, but I will not go on to discuss it here.

schema to be the meaning-constituting schema for the term "know". And that is quite another project which I will not pursue here.

C. The problem of ad-hocness

The centre of the solution by semantic epistemicism is the rejection of relevant instances of the equivalence schema. To examine whether it is ad hoc we will need to examine its justification. Above we have introduced two rationales given by Armour-Garb and Beall, and Asay respectively. Below I will examine them in turn.

Unacceptable because of conflicting inclination

As has been indicated, according to Armour-Garb and Beall's exposition, the instance of the equivalence schema for the liar sentence, that is:

T-l: "l is false" is true if and only if l is false

is precluded from the meaning of the truth predicate *because* it gives rise to conflicting inclinations both to apply and not to apply the truth predicate. Our inclination to accept T-l is *overridden* because of the contradiction and therefore it fails to constitute the acceptance property of the truth predicate (Armour-Garb and Beall, 2005, 92).

Now on the face of it, it seems that this rationale for rejecting T-l is independent from the liar paradox. It derives from Horwich's use theory of meaning and the minimalist theory of truth. Even if rejecting T-l eventually fails to provide a solution to the liar paradox, as long as one accepts Horwich's two theories and accepts that the acceptance of some sentences will be overridden if it gives rise to contradiction, then he will need to acknowledge that T-l is indeed precluded.

But there is one problem here. Let us just put aside relevant psychological issues, and assume for the sake of argument that it is the case that our inclination of accepting some certain sort of sentences will be overridden once the acceptance of which gives rise to contradictory inclination. It remains a question that exactly which inclination will be overridden. For the conflicting inclination in the

application of the truth predicate is not evoked simply by the instance of the equivalence schema for the liar sentence, rather, as is shown above, it stems from the *incompatibility between the acceptance properties of both "true" and "not"* in the case of the liar sentence. If one of the acceptance properties should be restricted, in the sense that some instances satisfying the regularity indicated in the acceptance property are not acceptable, then the question is, which one? One way to press this problem and make it more obvious is to ask, based on the incompatibility between two acceptance properties here, why it is the acceptance property of "true" that is overridden rather than that of "not".

To be more specific, suppose that one of the basic acceptance property of "not" is our inclination to accept instances of the form "*p* or not *p*", which we may properly call *exhaustion*-schema, then what gives rise to our conflicting inclination (both to apply and not to apply the truth predicate) in the case of the liar sentence is the following two instances:

T-*l*: "*l* is false" is true if and only if *l* is false.

E-*l*: Either *l* is false or *l* is true.

If our acceptance of them gives rise to conflicting inclination, and therefore something *is* overridden, then my question is, which one, T-$_l$, or E-$_l$?

Some may argue that, perhaps we have reason not to reject E-$_l$, for example, since it appears that rejecting E-$_l$ will give rise to the truth-value gap, which we may have reason to deny, so the remaining option is to reject T-$_l$. However, the fact that we have reason to maintain E-$_l$ does not mean that our inclination to accept E-$_l$ won't be overridden. It, at best, gives rise to one more conflicting inclination, between rejecting E-$_l$ and maintaining it. On the other hand, the fact that we have reason to believe that our inclination to accept E-$_l$ won't be overridden does not mean that our inclination to accept T-$_l$, is overridden. The fact that we know all but one answer to a question are incorrect does not mean that we *understand* the remaining one (the only one that we don't know whether it is incorrect) is correct. Of course, in this

case we can still make the *right* choice, since we know that the first three options are incorrect, but that does not mean that we know *why* the last option is correct. Some may suggest that perhaps, given the principle of symmetry, we should reject them both. But the fact that we have to appeal to the principle of symmetry exactly shows that we find no independent reason to accept or reject any one of them. The upshot of my objection is that Horwich does not provide enough justification to *explain* exactly why our inclination to accept T-$_l$ is overridden. There seems to be nothing determining such a result.

Unacceptable because of ungroundedness

In Asay's exposition, the reason why T-$_l$ is not acceptable is because the liar sentence is ungrounded. According to Asay, Horwich's idea is that once a sentence, say, p, is ungrounded, then the relevant instance of the equivalence schema for it will be unacceptable and so be precluded from the meaning of the truth predicate.

Asay finds this rationale unacceptable. In his view, it is not acceptable because it "renders perfectly acceptable T-biconditionals [instances of the equivalence schema] as being unacceptable" (Asay, 2015, 685). Some instances of the equivalence schema for ungrounded sentences are actually true. For example, consider the truth-teller:

t: *t* is true

The sentence is ungrounded, at least in the sense that it is truth-related and is not entailed by non-truth-related sentence. So in Horwich's criterion, the instance of the equivalence schema for it:

T-$_t$: "*t* is true" is true if and only if *t* is true

will be unacceptable. Asay points out that according to semantic epistemicism, T-$_t$ is either true or false, although we cannot know which one, it nevertheless has a truth value. And since the two sides of T-$_t$ are the same sentence, namely, the truth-teller, T-$_t$ is true.

If so, then being ungrounded does not mean that the corresponding instance of the equivalence schema is false and so if T-$_l$ is indeed not acceptable, "the notion of grounding is not useful in accounting for why" (Asay, 2015, 686).

In conclusion, both rationales for rejecting the instance of the equivalence schema for the liar sentence seem to be too weak to pick out exactly the right instance. In the first place, appealing to conflicting inclinations does not by itself constitutes a rejection to T-$_l$; in the second place, being ungrounded does not make the corresponding instance of the equivalence schema false. In any case, unless more restriction is introduced, I see no firm reason why we should reject the relevant instance of the equivalence schema other than it will block the liar paradox. Therefore, semantic epistemicism, at least, under the current two developments, is ad hoc.

3.4 Summary

We have seen that Horwich's semantic epistemicism can indeed be interpreted as a kind of silence approach. For it declares that the liar sentence has a determinate truth value, but it rejects that we can classify it, for the relevant schema that is necessary for doing that classification is not applicable in the case of the liar sentence. As a result, although we can assume that the liar sentence is either true or is false without getting in the paradox, we cannot know exactly which one it is and so fail to have the ability to classify it.

As to the three problems concerning the silence approach, first, given the loss of the equivalence schema, the pristine liar paradox is blocked; second, the theory can treat the revenge sentence R-$_{\text{silence}}$ in the same way it treats the pristine liar sentence. This treatment will result in a divergence between assertability and truth, which requires further exploration. Lastly, and unfortunately, the two rationales for rejecting the relevant instances of the equivalence schema are too weak to do the job. As a result, the solution may work to block the contradiction, but it is not free

from being ad hoc.

In next section, we will turn to another candidate theory for the silence approach and see whether it can provide a satisfactory response to the liar paradox.

4. Hofweber: Exception theory

In this section, we turn to the other possible candidate theory, which I call the *exception theory*. This type of theory claims that unlike our common view, the whole or at least most of the inferential rules, or axioms or principles that we accept for reasoning, despite the fact that they are generally valid, have exceptions. Hofweber (2007) and Hofweber (2010) question the entire logical system, that perhaps what we are so far believing in is not, as what he called "a strictly valid" system, but rather, it is just a "generically valid" system. That is, though in most of the cases those principles, inferential rules, are valid, they have exceptions—that they are not truth preserving in some use.

4.1 Strict validity and generic validity

To understand the solution Hofweber has in mind, we should know his distinction between two kinds of validity, *strictly valid* and *generically valid*. Here is Hofweber's definition:

> …a. Let's call an inference rule strictly valid iff each and every instance is truth preserving.
> b. Let's call an inference rule generically valid iff instances are truth preserving (understood as a generic statement). (Hofweber, 2007, 150)

For an inferential rule to be strictly valid it requires that there is no counterexample to that rule, that is, there is no situation where its application derives something false from something true. While a generically valid rule only requires that in most of the cases, the rule is truth preserving, but it also allows exceptions. Of course, it would be a little bit uncomfortable to call a generic rule "valid". Rules of the latter kind are some rules like those we apply in everyday life. For example, "if t is a bear, then t is dangerous". Certainly in most of the cases a bear is dangerous, but with

further information, say, the bear is dying, the rule is not applicable here, and we say that a dying bear is an exception to that rule.

Now, we always believe that we could have a deductive logic, in the sense that it contains all and only those strictly valid rules, but Hofweber asks us to give up that "dream" (Hofweber, 2007, 147). The system we hold here is not a strict deductive system, but just a generic one which means that the rules it contains are only generically valid.[169] Such a system allows exceptions, in particular, Hofweber says Curry's paradox and the liar paradox are such exceptions.[170]

One may argue that if those rules are only generically valid, from the traditional point of view, they are simply "invalid" rules. For validity is defined as universally truth preserving and does not allow exception, and since they are invalid in this sense, we should give them up. But this conception of "validity" is exactly what Hofweber asks us to give up! If we hold a generic conception of logic and of inferential rules, then reasoning with rules under the generic conception is rational but not strictly valid. In fact, we use this kind of reasoning almost everywhere in our daily life--it seems that we will not say, believing in rules like "if the house is on fire, we should leave" is irrational, even though it is not strictly valid. Put it in this way, we may say, what Hofweber suggests is that the logical rules we hold are rules that are just like what we use in our daily life; though it is possible that they have exception, in the sense that some application of them is not truth preserving, nonetheless, it is rational to reason with to them.

4.2 Exception as solution

Now how does the exception theory solve the liar paradox? As noted, normally,

---

[169] Note that Hofweber never makes it explicit whether all rules we hold have exceptions or just some of them.

[170] Throughout this section, whenever I say something is an exception to some rules, I mean that that thing provides an exception to relevant rules, showing that the rules are not truth-preserving in general.

when we are searching for a solution to a paradox, we are looking for some unsound premises or some invalid inferential rules. Now, it is very easy (and wrong) to identify Hofweber's theory as a solution of this kind, since in this case, it seems what Hofweber suggests is that every rule we employ in the reasoning that leads to the contradiction is invalid and so should be given up. But it is not. Hofweber does not deny the correctness of any rule here, he only denies that they are strictly valid.

> …each step in the reasoning that leads to paradox is correct in the sense that it is based on a rule which is (generically) valid and which is appropriately used in reasoning…but nonetheless…the conclusions drawn with the particular cases of the (generically) valid inference rules are rationally not to be accepted. (Hofweber, 2007, 151).

This is coherent in his theory, for generically valid rules do have exception in which case they are not truth preserving. So we do not need to give up any rule that we have so far been familiar with. For example, we do not need to give up the law of excluded middle, the law of non-contradiction, the introduction rule and elimination rule for the truth predicate (from "p" is true we can infer p, and from p we can infer "p" is true). But since the liar paradox is an exception to those rules, we need not accept the resulting contradiction. Thus the contradictory result still can be rationally rejected.

But there remains a question. If the liar sentence is not both true and false, as the consequence of the paradox, then what would be its truth value? Hofweber seems to suggest that the following two schemas are relevant to the classification of the truth value of the liar sentence:

(1) p∨¬p

(2) True ('p')∨True('¬p')

where "p" is a place-holder for sentences.

The first can be regarded as the law of excluded middle (in its schematic form), while the second one is the principle of bivalence.

Now, according to Hofweber, in his theory, there could be two options for the

classification.[171] First, we read those two schemas as strictly valid. In this case surely the liar sentence is either true or false, and it seems that whichever it is, based on the liar paradox, we have a contradiction. But "any argument towards this contradiction will use some rules which are only generically valid. And the instances of these rules with either the liar sentence or its negation, or the claim that the liar sentence is true, or the claim that its negation is true, will be exception"[172] (Hofweber, 2007, 156). Here Hofweber seems to presuppose that any attempt try to classify the liar sentence will eventually lead to contradiction, if we assume that the above two schemas are strictly valid. And if so, any rules employed in classifying the truth status of the liar sentence will then have the liar sentence as an exception. Thus we just do not have the rules that are necessary for classifying the liar sentence, even though based on the strict validity of the law of excluded middle and principle of bivalence, we know it is either true or false. In this case, "it will be a case of ignorance, although there is an answer to the question." (Hofweber, 2007, 156)

If we adopt this line, then Hofweber's theory can indeed be interpreted as some sort of silence solution. For, like Horwich, it implies that we cannot know the truth status of the liar sentence, given the loss of some relevant rules of inference.

Hofweber also suggests a second way to treat the liar sentence. And that is to read the above two schemas as only generically valid and the liar sentence is an exception to them (Hofweber, 2007, 157). That means, the instance "the liar

---

[171] Actually there could be more. He only discusses two.

[172] To spell it out, consider the following liar argument:

The targeted sentence: *l*: *l* is false.

1. *l* is either true or false (applying the law of excluded middle)
2. If *l* is true, then *l* is false (applying the elimination rule of the truth predicate)
3. If *l* is false, then *l* is true (applying the introduction rule of the truth predicate)
4. *l* is true if and only if *l* is false (applying introduction rule of biconditional on 2 and 3)
5. Either *l* is both true and false, or *l* is neither true nor false. (applying the elimination rule of biconditional on 4)
6. *l* is both true and false. (applying disjuntive syllogism on 1 and 5)

Hofweber's idea may be understood as that, the applications of some certain inferential rules above on the liar sentence and its related sentences are not truth preserving. And so, even though the conclusion derived is a contradiction, we can rationally reject it.

sentence is either true or false" is simply false, so the liar sentence is neither true nor false.[173] In this case, the truth status of the liar sentence is classified and so it is not a silence approach.

Now since my primary interest is to see whether Hofweber's solution, when interpreted as a silence approach, can provide a satisfactory response to the liar paradox, below I will simply take into consideration the first sort of treatment on the liar sentence. That is, to read the law of excluded middle and principle of bivalence as strictly valid.

Now under this situation, how can Hofweber's theory resolve the three problems concerning the silence solution?

First, the pristine liar paradox is certainly blocked. For in Hofweber's theory, the liar paradox is exactly an exception to the rules of inference that we employ in constructing the liar paradox. So we can rationally accept those rules of inference as generically valid while at the same time rationally reject its consequence that the liar sentence is both true and false. And as to the truth status of the liar sentence, it is unknown. For any attempt to classify it will lead us to a contradiction, which only shows that the rules of inference we use in the classification have the liar sentence as an exception. Although we can know that the liar sentence is either true or false (based on the strict reading on principle of bivalence), we cannot further classify it *as true* or *as false*.

Second, the revenge problem. Hofweber believes that one of the advantages of his theory is that it is free from the revenge problem. The revenge sentence:

$R_{\text{-silence}}$: $R_{\text{-silence}}$ is not true or not classifiable,

if it produces another contradiction, for Hofweber, that is just a symptom for another

[173] This is different from semantic epistemicism, for the reason why a rule is not applicable in this case is that it is not truth preserving, which means, it deduces something false from true premise. The above schema can be regarded as an inferential rule with no premise. So when it has an exception, this means that instance of the application of the rule is simply false.

exception to the rules of inference we employ.

This leaves us with the last problem, about the ad-hocness of the solution, which will be discussed in next section.

4.3 Critique

I find that Hofweber's theory is attractive for two reasons. First, as Hofweber himself points out, his theory satisfies our intuition about the liar paradox. We find no problem with each one of the premises and the rules of inference employed in the liar paradox, and under the generic conceptions of them, we find it natural as well to reject the consequence. And furthermore, even though the liar paradox is puzzling, it does not bother our normal reasoning and everyday life. In Hofweber's theory, these intuitions are explained by reading those premises and rules of inference as merely generically valid. Second, it can handle the revenge problem without much difficulty. For any further contradiction is nothing but further exceptions to the rules of inferences employed. So I think that the theory does block the liar paradox and the related revenge problem.

The problem lies in the third, its ad-hocness. What the exception theory tells us is that our conception of our logical system should be a generic one rather than a strict deductive one. If one questions whether Hofweber has any independent reason for this claim, then Hofweber may probably reply by questioning whether there is any independent reason for the claim that the logical system is a strictly valid one. Debate at this point is very easy to be question begging, for both parties. So I do not want to argue on this point. Rather, I will simply accept his main position, that our logical system is only generically valid and argue that still, it cannot be free from ad-hocness.

Now without telling us a priori in what kinds of situation which kinds of reasoning are exceptions to which generically valid rules, it seems that whenever one holds

some principles but those principles lead to some problematic conclusion (conclusion that appears to be false), he or she can avoid being accused of holding some unsound principle by appealing to, as Hofweber does here, that those conclusions are cases where we have exceptions to those principles. He can always solve the predicament he encounters by making a special pardon by claiming that the conclusion in question is an exception (this is how he handles the revenge problem). In this sense it suffers severely from being ad hoc. Indeed, if Hofweber wants his theory to be theoretically plausible and practically possible, he needs to draw a priori, the realm where rules are applicable without exception and the realm which is exception to those rules.

To this, the most direct answer given by Hofweber is that, "The exceptions to the generically valid rules are simply the instances that don't preserve truth" (Hofweber, 2007, 152). Hofweber gives an example in everyday reasoning: To the question of what the exception is to the rule "if t is a bear, then t is dangerous", a proper answer is that every t that is a bear but not dangerous. However, this argument does not avoid the problem of being ad hoc.

Frist, one can immediately see that Hofweber's answer is a tautology. Since exception to a valid rule is when a particular application of the valid rule does not preserve truth, Hofweber's answer is like saying, the exception to a given rule is the exception to the given rule. That does not provide us with any useful characterization of the situations where the accepted rules are not truth preserving. Now it may be the case that, as Hofweber suggests, if we are to look for an exception to the principle "if t is a bear, then t is dangerous", then we should go into empirical world to see whether a given bear is dangerous. But in the case of the liar sentence, there is no such help that can cancel the triviality of this suggestion, since there is just no empirical ground for determining whether the liar paradox is an exception to some rules of inference.

Second, even if for the sake of argument, we ignore the above problem, there is no clear reason as to why, in the case of the liar paradox, if we apply those rules in reasoning and they lead to a contradictory consequence, we should conclude that the liar paradox is an exception to the rules we employ in the *construction* of the argument. According to Hofweber, an inferential rule has an exception in a certain case only if in that case it is not truth preserving—it derives something false from something true. In the case of the liar paradox, the conclusion is that the liar sentence is both true and false. So if Hofweber wishes to maintain that the liar argument is an exception to whatever rules we employ in constructing the paradox, he will need to insist that the conclusion it derives (that the liar sentence is both true and false) is *false,* and this presupposes law of non-contradiction, which says that no contradiction is true. But what if, in this case, the law of non-contradiction is itself not truth preserving? After all, in Hofweber's own light, our logical system is only generically valid, not strictly valid. What if the contradictory conclusion that the liar sentence is both true and false is indeed correct, as what *Dialetheist* claims? In this case, the inferential rules that we employ in constructing the argument are still truth preserving and so the liar does not constitute an exception to them. Rather, as the *Dialetheist* would argue, the liar paradox shows precisely that there is true contradiction—it is an exception to law of non-contradiction.

Hofweber may avoid the above problem by insisting that law of non-contradiction is exceptionless and so is strictly valid. But this choice only aggravates its ad-hocness, if there is *no independent reason* to explain why law of non-contradiction should only be read as strictly valid. Especially, given that the main position of Hofweber's thesis is that the ideal of a strictly valid deductive logical system is just a dream, in his essay there seems to be no good reason not to insist that law of non-contradiction is only generically valid.

4.4 Summary

In this section I have interpreted Hofweber's exception theory as a kind of silence approach. It declares that the liar sentence is either true or false, but it concedes that

we cannot further classify it for the relevant rules of inference that are necessary for doing that classification are only generically valid and the liar sentence is just an exception to them. As a result, although we can assume that the liar sentence is either true or is false without getting in the paradox, we cannot know exactly which one it is.

As to the three problems concerning the silence approach, it can handles the pristine liar and the revenge problem. It can always declare that the relevant contradictions are nothing but exceptions to those rules of inference we employ in classifying the paradoxical sentences. But this feature makes the whole theory highly ad hoc. More importantly, the theory has to make the most ad-hoc move, to declare that the law of non-contradiction is strictly valid, without exception. This won't be a problem for other theories, but it will be for Hofweber, given that the main thesis of Hofweber's theory is that, a strictly valid logical system is nothing but a dream.

5. Summary and future development

In this chapter I have sketched some core features of what I have been called the silence approach to the liar paradox. In particular, it involves the positions that first, the liar sentence has a truth value; second, we cannot classify it and so we cannot know its truth value. And the main reason for this unknowablity is that while the correct (or the best) semantic theory for our language works to explain most of the sentences, it fails to classify the liar-like sentences given either that the relevant principles that are necessary for the classification are not applicable in the case of the liar-like sentences or that no semantic status in the theory can fit those sentences.

I have tried to interpret Horwich's semantic epistemicism and Hofweber's exceptional theory as two possible silence approaches and find that there are some theoretical advantages of this kind of approach. One is that the way they treat the pristine liar sentence can be easily extended to the revenge sentence without much trouble. In Horwich's solution, the relevant instances of the equivalence schema are

precluded from the meaning of truth. And given the loss of this principle, one cannot classify the truth status of the liar sentences. This treatment can be extended to the revenge sentence easily. As to the knower paradox, while it remains a question whether it requires the same solution that Horwich gives to the liar paradox, I have shown that under Horwich's strategy, it is possible to solve the knower paradox in the same kind of way it solves the liar. But of course, further study on this issue is necessary. The remaining uncomfortable element is that it implies a divergence between assertability and truth, which deserves further exploration. Hofweber's solution asks us to significantly change our conception on our logical system. Instead of believing that logical principles are strictly valid without exception, he asks us to take them as only generically valid which allows situations where they are not truth-preserving. The liar paradox under this conception is just an exception to our inferential system. This treatment can also be easily extended to the revenge paradox—the revenge is just another exception to our inferential system.

My examination also reveals that these two solutions suffer severely from the problem of ad-hocness. Indeed, this is not something unforeseeable. Since both solutions try to undermine our capacity of classifying the semantic status of liar-like sentences by rejecting *some* instances of the general application of relevant semantic or logical principles, which is just one of the two possible strategies for silence approach to maintain the unclassifiablity of the problematic sentences.[174]

Of course, even if we can solve the problem of ad-hocness, that we have independent reason to believe why some principles that are necessary to classify a sentence are not applicable in the case of paradoxical sentences, it remains a question, which I think is the most important one, whether this incompleteness is a symptom revealing that the accepted semantic theory is defective or incomplete since it cannot provide a systematic understanding of our language or it is a

[174] The other is to deny that there is an appropriate semantic status for paradoxical sentences, see the beginning of this chapter.

symptom that shows that there is just no possible semantic theory that can eventually do the job. Perhaps, there is something intrinsically defective in our cognitive capacity that prevents us from having a complete theory for a universal language like natural language. This is what I have called the silence conjecture.

There are indeed some theories that seems to share this conjecture. For example, Palmquist (2000)[175] holds that while most of our ordinary descriptions of realities obey what he calls analytic logic (e.g. classical logic), there are a few usages are governed by what he calls synthetic logic, which follows the direct opposites of the classical laws. So while in analytic logic we have law of excluded middle, law of identity, law of non-contradiction, in synthetic logic, we have law of non-excluded middle, law of non-identity and law of contradiction. The upshot is that, not all realities are appropriate for analytic descriptions, there are at least some realities the descriptions of which are not possible if we do not adopt synthetic logic. If this is the case, then Palmquist's position *may be interpreted as* a kind of silence approach, though it is different from the two theories examined in this dissertation—The reason that the passing semantic theories fail to correctly classify the liar sentence is exactly because they follow analytic logic, a logic which does not apply here (the whole classification theory fails to apply in this case). So if we stick with analytic logic, then surely we should be silent, for we never be able to correctly classify the sentence.[176]

Of course, even so, the silence conjecture is still a conjecture. I have no clear idea what this amounts to or how to understand this phenomenon, if it is indeed a fact. And certainly I have no idea how to prove it, or whether it can be proved at all. This

[175] I should mention first that Palmquist (2000) does not provide a detailed discussion on the liar paradox. The content I present here and below is merely my interpretation of his theory. See also Palmquist (2013) for more discussion on synthetic logic.

[176] Note that, Palmquist may not therefore conclude that there is no hope at all in classifying the liar sentence. As long as we adopt synthetic logic, we can still meaningfully classify the liar sentence. In other words, for Palmquist, the semantic status of the liar sentence belongs to the realm where only synthetic logic applies.

is what I wish to explore more in the future. Here what I can conclude is that, the silence approach does provide a distinctive and possible response to the liar paradox, although it is very vulnerable to the problem of ad-hocness and still lacks any plausible theoretical ground.

References

Aristotle, (1984). Metaphysica, English translation “Metaphysics” by Ross, W.D. in Barnes, J. (ed.) (1984), pp.3343~3317.

Armour-Garb, B. (2005). Wrestling with (and without) dialetheism. *Australasian Journal of Philosophy*, 83(1), 87-102. doi:10.1080/00048400500044306.

Armour-Garb, B. & Beall, J.C. (2005). Minimalism, epistemicism, and paradox. In Armour-Garb, B. and Beall, J.C. (Eds.), *Deflationism and paradox*(pp.85-96). Oxford: Clarendon Press.

Armour-Garb, B. & Beall, J.C. (Eds.) (2005). *Deflationism and paradox*. Oxford: Clarendon Press.

Armour-Garb, B. and Woodbridge, A. J. (2006). Dialetheism, semantic pathology, and the open pair. *Australasian Journal of Philosophy,* 84(3), 395-416. doi:10.1080/00048400600895912

Asay, J. (2015). Epistemicism and the liar. *Synthesis*, 192 (3),679-699. Doi: 10.1007/s11229-014-0596-x

Austin, J.L. (1950). Truth. *Proceedings of the Aristotelian Society*, suppl. vol. 24, pp.111-128.

Bacon, A. (2015). Can the classical logician avoid the revenge paradox? *Philosophical Review*, 124(3), 299-352. doi: 10.1215/00318108-2895327.

Barnes, J. (ed.) (1984). *The complete works of Aristotle: The revised Oxford translation, one volume digital editiion*. US: Princeton University Press.

Barwise, J. & Etchemendy, J. (1989). *The liar: An essay on truth and circularity*. New York: Oxford University Press.

Beall, J.C. (2001). A neglected deflationist approach to the liar, *Analysis*, 61(2), 126-129. doi: https://doi.org/10.1093/analys/61.2.126.

Beall, J.C. (2006). Truth and paradox: a philosophical sketch. In Dale Jacquette (ed.), (2006). *Philosophy of logic*. pp.187-272.

Beall, J.C. (Ed.). (2007). *Revenge of the liar: New essays on the paradox*. New York: Oxford University Press.

Beall, J.C. (2009). *Spandrels of truth*. Oxford: Oxford University Press.

Beall, J.C. Glanzberg, M. and Ripley, D. (2018). *Formal theories of truth*. UK: Oxford University.

Black, M. (1948). The semantic definition of truth. *Analysis*, 8(4), 49-63.

Blackburn, S. (1984). *Spreading the word: Goundings in the philosophy of language*. New York: Oxford University Press.

Blackburn, S. (2018). *On truth*. New York: Oxford University Press.

Burge, T. (1979). Semantical paradox. *Journal of Philosophy,* 76(4), 169-198. doi:10.2307/2025724.

Burgess, A.G. and Burgess, J. P. (2011). *Truth*. UK: Princeton University Press.

Boolos, G.S., Burgess, J.P. and Jeffrey, R.C. (2007). *Computability and logic* (5th edition). Cambridge: Cambridge University Press.

Chihara, C. (1979), The semantic paradox: A diagnostic investigation, *Philosophical Review*, 88(4): 590-618.

Coffa, J.A. (1991). *The semantic tradition from Kant to Carnap: To the Vienna station*. Cambridge: Cambridge University Press.

Cook, R.T. (2009). What is a truth value and how many are there? *Studia Logica,* 92(2), 183-201. doi: 10.1007/s11225-009-9194-1

Davidson, D. (1990). The structure and content of truth. *Journal of philosophy*, 87(6), 279~238.

Dedekind, D. (1923), *Was sind und was sollen die Zahlen?* (5th ed.). Braunschweig.

Field, H. (1972). Tarski's theory of truth. *Journal of philosophy*, 69(13), 347~375. The essay is collected in Lynch, P. M. (ed.) (2001), pp.365~396. All page references are to Lynch, P. M. (ed.) (2001).

Glüer, K. (2011). *Donald Davidson: A short introduction.* New York: Oxford University Press.

Goldstein, L. (2009) A consistent way with paradox. *Philosophical Studies*, 144, 377-389. doi: 10.1007/s11098-008-9215-3.

Gómez-Torrente, M. (2019). Alfred Tarski. *The Stanford encyclopedia of philosophy* (Spring 2019 edition), Zalta, N. E. (ed.).

URL = https://plato.stanford.edu/archives/spr2019/entries/tarski-truth/.

Grayling, A. C. (1990). *An introduction to philosophical logic* (new edition). Oxford: Alden Press.

Gupta, A. (1982). Truth and paradox. *Journal of philosophical logic*, 11(1), 1~60.

Gupta, A. & Belnap, N. (1993). *The revision theory of truth.* Cambridge: MIT Press.

Gödel, K. (1931). *Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme* I. *Monatshefte für Mathematik und Physik*, 38: 173–198. Reprinted and translated in Gödel (1986), 144–195.

Gödel, K. (1986). *Collected Works*, vol. 1, Oxford: Oxford University Press.

Haack, S. (1976). Is it true what they say about Tarski? *Philosophy*, 51(197), 323-336.

Hodges, W. (2018). Tarski's Truth Definitions. *The Stanford encyclopedia of philosophy* (Fall 2018 Edition), Zalta, N. E. (ed.).
URL = https://plato.stanford.edu/archives/fall2018/entries/tarski-truth/.

Hofweber, T. (2007). Validity, paradox, and the ideal of deductive logic. In Beall, J.C. (Ed.), *Revenge of the liar: New essays on the paradox* (pp.145-158). New York: Oxford University Press.

Hofweber, T.(2010). Inferential role and the ideal of deductive logic. In *The Baltic International Yearbook of Cognition, Logic and Communication Volume 5: Meaning, Understanding and Knowledge* (pp.1-26). Doi: 10.4148/biyclc.v5i0.283

Horwich, P. (1998a). *Meaning*. New York: Oxford University Press.

Horwich, P. (1998b). *Truth* (2 ed.). New York: Oxford University Press.

Horwich, P. (2005). A minimalist critique of Tarski on truth. In Armour-Garb, B. and Beall, J.C. (Eds.), *Deflationism and paradox* (pp.85-96). Oxford: Clarendon Press.

Horwich, P. (2010). *Truth-Meaning-Reality*. New York: Oxford University Press.

Jacquette, D. (ed.), (2006). Handbook of the philosophy of science Volume 5: Philosophy of logic. Boston: Elsevier/North Holland.

Jacquette, D.(ed.), (2006). *Blackwell companions to philosophy: A companion to philosophical logic*. HK: Blackwell Publishing.

Kirkham, R. L. (2001). *Theories of truth: A critical introduction*. Cambridge: MIT Press.

Kripke, S.A. (1975). Outline of a theory of truth. *Journal of Philosophy*, 72(19), 690-716. doi: 10.2307/2024634.

Littmann, G. and Simmons, K. (2004). A critique of dialetheism. In *The law of non-contradiction*. Priest, G., Beall, J.C. and Armour-Grab, B. (2004) (eds), pp. 355-384.

Lynch, P. M. (ed.) (2001). *The nature of truth: Classic and contemporary perspectives*. US: MIT Press.

Martin, R.L. (1967). Toward a solution to the liar paradox. *The Philosophical Review*, 76(3), 279-311.

Neurath, O. (1931). Physikalismus. *Scientia*, 25 (50): 297-303. All page references are to the English translation in Neurath (1983), pp.52~57.

Neurath, O. (1983). *Philosophical papers*, *1913~1946*. Edited and translated by Robert S. Cohen and Marie Neurath. Dordrecht: D. Reidel.

Palmquist, R. S. (2000). *The tree of philosophy* (4th). Hong Kong: Philopsychy Press.

Palmquist, R. S. (2013). Paradox in Perspective: A Liar's Guide to Humor, in *No More Hanging Around—The Quest for Truth and Meaning*, ed. Ellen Zhang Ying, et al. (Hong Kong: 次文化 [Subculture Limited], 2013), pp.37-44.

Parsons, T. (1990). True contradictions. *Canadian Journal of Philosophy*, 20(3), 335-353. doi: 10.1080/00455091.1990.10716495.

Priest, G. (1979). The logic of paradox. *Journal of Philosophical Logic*, 8(1), 219-241.

Priest, G. (1994). The structure of the paradoxes of self-reference. Mind, 103(409), 25~34.

Priest, G. (1995). Gaps and gluts: Reply to Parsons. Canadian Journal of Philosophy, 25(1), 57-66. doi: 10.1080/00455091.1995.10717404.

Priest, G. (1997). Yablo's paradox. *Analysis*, 57(4):236-242.

Priest, G. (2000). Could everything be true? *Australasian Journal of Philosophy*, 78 (2),189-195.

Priest, G.(2002). *Beyond the limits of thought* (2ed.). New York: Oxford University Press.

Priest, G., Beall, J.C. and Armour-Grab, B. (2004) (eds). *The law of non-contradiction: New philosophical essays.* New York: Oxford University Press.

Priest, G. (2006). *In contradiction: A study of the transconsistent* (2ed). New York: Oxford University Press.

Priest, G. (2008). *An introduction to non-classical logic: From if to is* (2ed.). New York: Cambridge University Press.

Priest, G., Tanaka, K. and Weber, Z. (2018). Paraconsistent Logic, *The Stanford Encyclopedia of Philosophy* (Summer 2018 Edition). Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/sum2018/entries/logic-paraconsistent/.

Ray, G. (2006). Truth, the liar, and Tarskian truth definition. In Jacquette (ed.), (2006). pp. 164-176.

Russell, B. (1912). Truth and Falsehood. Chapter 12 of *The problems of philosophy*. Oxford: Oxford University Press.

Saka, P. (2007). *How to think about meaning*. Dordrecht: Springer.

Scharp, K. (2013). *Replacing truth*. New York: Oxford University Press.

Sher, G. (2006). Truth, the liar, and Tarski's semantics. In Jacquette, D. (ed.) (2006), pp.145~163.

Sider, T. (2010). *Logic for philosophy*. New York: Oxford University Presss.

Simmons, K. (1993). *Universality and the liar: An essay on truth and the diagonal argument.* New York: Cambridge University Press.

Smith, N.J.J. (2000). The principle of uniform solution (of the paradoxes of self-reference). *Mind*, 109(433), 117-122.

Smith, P. (2013). *An introduction to Gödel's theorems* (2ed.). New York: Cambridge University Press.

Soames, S. (1984). What is a theory of truth? *Journal of philosophy,* 81(8), 411~429. The essay is collected in Lynch, P. M. (ed.) (2001), pp.397~418. All page

references are to Lynch, P. M. (ed.) (2001).

Strawson, P.F. (1950). On referring. *Mind*, 59(235), 320-344.

Tarski, A. (1933). *Pojecie prawdy w jezykach nauk dedukcyjnych*. Warsaw. All pages references are to the English translation "The concept of truth in formalized languages" in Tarski (1983), pp. 152~278.

Tarski, A. (1936). O ugruntowaniu naukowej semantyki. *Przeglad Filozoficzny*. 39: 50~57. All page references are to the English translation "The establishment of scientific semantics" in Tarski (1983), pp.401~408.

Tarski, A. (1944). The semantic conception of truth: and the foundations of semantics. *Philosophy and Phenomenological Research*, 4(3), 341-376.

Tarski, A. (1969). Truth and proof. *Scientific American*, 220(June), 63~77.

Tarski, A. (1983). *Logic, semantics, metamathematics: Papers from 1923 to 1983* (2ed). Corcoran, J. (ed.). Indianapolis: Hackett.

Taylor, K. (1998). *Truth and meaning: An introduction to the philosophy of language*. Oxford: Blackwell.

Whitehead, A. N. and Russell, B.A.W. (1925). *Principia Mathematica* (2 ed.), vol. 1 and 3. Cambridge: Cambridge University Press.

Yablo, S. (1993). Paradox without self-reference. *Analysis,* 53(4), 251-2.

CURRICULUM VITAE

Academic qualification of the thesis author, Mr. LI Dilin:

-Received the degree of Bachelor of Management from Shantou University, July 2015.

11.2020