Universidade de São Paulo
Faculdade de Filosofia Letras e Ciências Humanas
Departamento de Filosofia

Liars and Circles

Essays on Truth, Self-Reference and Paradoxes

(O Círculo do Mentiroso: Ensaios sobre Autorreferência, Verdade e Paradoxos)

Dissertação de mestrado apresentada ao Departamento de Filosofia

Candidata: Fernanda Birolli Abrahão
Orientador: Edelcio Gonçalves de Souza

Versão Corrigida

São Paulo, 2023

Fernanda Birolli Abrahão

Liars and Circles

Essays on Truth, Self-Reference and Paradoxes

(O Círculo do Mentiroso: Ensaios sobre Autorreferência, Verdade e Paradoxos)

Dissertação apresentada ao Programa de Pós-Graduação em Filosofia do Departamento de Filosofia da Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo, para obtenção do título de Mestre em Filosofia sob a orientação do Prof. Dr. Edelcio Gonçalves de Souza.

Versão Corrigida

UNIVERSIDADE DE SÃO PAULO
FACULDADE DE FILOSOFIA, LETRAS E CIÊNCIAS HUMANAS

ENTREGA DO EXEMPLAR CORRIGIDO DA DISSERTAÇÃO/TESE

Termo de Anuência do (a) orientador (a)

Nome do (a) aluno (a): Fernanda Birolli Abrahão_

Data da defesa: _26_/_06_/_2023

Nome do Prof. (a) orientador (a): Edelcio Gonçalves de Souza

Nos termos da legislação vigente, declaro **ESTAR CIENTE** do conteúdo deste **EXEMPLAR CORRIGIDO** elaborado em atenção às sugestões dos membros da comissão Julgadora na sessão de defesa do trabalho, manifestando-me **plenamente favorável** ao seu encaminhamento ao Sistema Janus e publicação no **Portal Digital de Teses da USP**.

São Paulo, _25_/_09__/ 2023

Edelcio Gonçalves de Souza
_____
(Assinatura do (a) orientador (a)

Abrahão, F.B. **Liars and Circles: Essays on Truth, Self-Reference and Paradoxes.** Dissertação de Mestrado. Faculdade de Filosofia, Letras e Ciências Humanas. Departamento de Filosofia, Universidade de São Paulo, São Paulo, 2023.

Aprovada em:

Banca examinadora:

Prof[a]. Dra. Itala Maria Loffredo D'Ottaviano. Instituição: Universidade Estadual de Campinas (Unicamp).
Julgamento: Assinatura:

Prof. Dr. Michael Glanzberg Instituição: Rutgers University.
Julgamento: Assinatura:

Dr. Daniel Arvage Nagase Instituição: Universidade de São Paulo (USP).
Julgamento: Assinatura:

To my brother, for all he has taught me.

Acknowledgments

I must begin this acknowledgement section by expressing my earnest thanks to my supervisor, Edelcio Gonçalves de Souza. During my three years of master's studies, I repeatedly witnessed Edelcio's commitment not only to helping his students in their dissertations and thesis, but also to doing the best he can to equip them with a solid foundation in formal logic. Before each semester, he would approach us with the question, "what do you guys want to study this time?," and would embrace our suggestions and requests with the curiosity and excitement of a first-year student. If I have any degree of good training in logic, it is greatly due to him.

During this period, I've also had the immense pleasure of working with one of the most influential figures in shaping my perspective on logic and philosophy: Graham Priest. I've never learned so much in so little time as in the six months I spent in New York having him as my adviser. His impressive erudition did not make him uninterested in his students' research; on the contrary: he listened to our ideas and projects with the interest and rigor that he would listen to any colleague in the area, often with sharp objections that prompted us to rethink part of our research. On a personal level, Graham welcomed me in New York with open arms and ensured that I felt at home at CUNY. For all of this, I owe him my deepest thanks.

Someone whom I must thank as effusively as the first two is Daniel Nagase. I met him in 2018 when he was the assistant in a logic course taught by Edelcio de Souza. At that time, I didn't know anything about logic – set theory, the subject of the course, sounded like Greek to me. In a moment of near desperation, I decided to write to Daniel and ask him for help; I distinctively remember saying "I really don't get what it means to *prove* something." From this point onward, he has been a constant presence in my research: he helped me discover a topic for my MA studies, corrected my logic and mathematics exercises, engaged in discussions about my work, and, above all, he made me feel capable of pursuing such a difficult area as logic, at a time when I myself didn't think I could do it. I also thank him and Olívia (and, more recently, little Oliver) for the friendship.

To the members of my defense committee, Daniel Nagase (once again), Itala D'Ottaviano, and Michael Glanzberg, I extend my heartfelt thanks for accepting the invitation. I am honored to have you all on my committee and hope you enjoy the reading! I must also single out Itala for organizing the remarkable event SPLogIC, which, I assume, required an enormous amount of effort.

Another person who contributed, even though from afar, to my research is Adam Bjorndahl, both in the pleasant conversations about logic during the NASSLLI conference, and by reading the third chapter of my dissertation and making important suggestions. In the same way, I must thank Guilherme Cardoso for inviting me to speak on his 'LLA (Lógica Ladeira Abaixo)' seminar about self-reference and circularity, for his fair objections to my thesis, and for the warm reception from him and the members of the group, to which I wish I could have participated more. Furthermore, I thank Hannes Leitgeb for meeting me over Zoom to discuss my dissertation, and Rafal Urbaniak, for the helpful email exchanges about my ideas and academic life in general.

I also thank everyone from *Areté: Centro de Estudos Helênicos*, particularly the members of the group *Gnomon.* I have gained invaluable knowledge about Greek mathematics and philosophy from their expertise. Moreover, attending the group is always a joy to me, thanks to its rigorous yet light-hearted atmosphere.

My friends from logic were (and continue being) essential for my formation process: Euclides, Taimara, Marco, Paulo, Levi, Pedro Falcão, Caio, Rodrigo Lima, André, Nina (who does not study logic but is usually around), and Luiza. The fun, long meetings we had during the pandemic taught me much of what I know about logic; I'm glad to have such interesting people in my area of study. To Euclides and Caio, may there be much drunk chess in our future.

I'm also grateful to Pedro Nagem for all the support he gave me, which was very important for the completion of this work.

On a personal level, I must thank my dear friend Rodrigo Figueiredo for all those years of friendship and support, and for all our conversations about philosophy, cinema, literature, that helped to shape my worldview in many aspects. The passing of years only makes me more sure of the importance of this friendship to my life.

I cannot forget to thank my New York friends and colleagues: Brian, Esma, Ebubekir, and Filippo had to endure my talks that were often about undercooked ideas in Graham's research group. I also thank Yale for the helpful discussions about logic over a beer. Lastly, I thank Tomasz for the comments on my fourth chapter, as well as for, together with Zuzia and Jeremy, making my stay in New York much more pleasant and fun. Other Brazilian friends which were very important in this period are Marcela and Ana Paula.

My parents, Valéria and Roberto, and my brother, Lucas (to whom I dedicate this dissertation), deserve my immense gratitude. Their unwavering support has

not only enabled me to pursue an academic career, but has also played a major role in shaping my intellectual and cultural development. I couldn't thank them enough for their affection and guidance.

To end this acknowledgement section, I would like to express my gratitude to Thiago Alexandre. Firstly, I thank him for, with his brilliancy and insights, helping me to develop crucial parts of this research. Secondly, I thank him for his endless support and care. Lastly and most importantly, I thank him for all his love, which I treasure most deeply.

Of course, to end the section once and for all, I must give my many thanks to the São Paulo Research Foundation (FAPESP) for the funding that made this research possible. Processes' numbers: 2020/02402-2 and 2021/08273-2.

Each figure seemed to be, somehow, on the borderland of things, just as their theory was on the borderland of thought. He knew that each of these men stood at the extreme end, so to speak, of some wild road of reasoning. He could only fancy, as in some old-world fable, that if a man went westward to the end of the world he would find something – say a tree – that was more or less a tree, a tree possessed by a spirit; and that if he went east to the end of the world he would find something else that was not wholly itself – a tower, perhaps, of which the very shape was wicked. So these figures seemed to stand up, violent, and unaccountable against an ultimate horizon, visions from the verge.

G. K. Chesterton

Resumo

Abrahão, Fernanda Birolli. *Liars and Circles: Essays on Truth, Self-Reference and Paradoxes*. Dissertação (Mestrado) – Universidade de São Paulo. Faculdade de Filosofia, Letras e Ciências Humanas. Departamento de Filosofia. 2023.

Esta dissertação de mestrado consiste em quatro capítulos que investigam aspectos filosóficos e lógicos relacionados aos seguintes temas: paradoxos, autorreferência e circularidade. O primeiro capítulo aborda a questão "O que é um paradoxo?" Inicialmente, são exploradas a etimologia da palavra "paradoxo" e sua possível ligação ao conceito de "nonsense" (não-sentido), com base nas concepções dos filósofos Wittgenstein e Deleuze. O capítulo conclui com uma análise de argumentos lógicos específicos e uma avaliação de sua natureza paradoxal. No segundo capítulo, realizo um estudo formal dos paradoxos lógicos, incluindo o Mentiroso, o Heterológico e o Paradoxo de Yablo. Durante a discussão sobre este último paradoxo, identifico a falta de definições explícitas dos conceitos de autorreferência e circularidade. Esses conceitos são abordados em maior detalhe no terceiro capítulo, no qual são apresentadas definições precisas para os conceitos de referência, autorreferência e circularidade, utilizando interpretações modelo-teóricas. Além disso, são demonstradas propriedades relacionadas a paradoxos bem conhecidos. O quarto capítulo concentra-se nas linguagens e teorias semanticamente fechadas, fornecendo uma definição formal e construção de uma linguagem de primeira ordem semanticamente fechada bissortida. Também discuto se as linguagens cotidianas, como o português ou o inglês, são semanticamente fechadas. Esta dissertação oferece uma exploração aprofundada dos conceitos de verdade, autorreferência e paradoxos, estabelecendo uma conexão entre a filosofia e a lógica formal.

Palavras-chave: Paradoxo, Mentiroso, Autorreferência, Circularidade, Fecho Semântico.

Abstract

Abrahão, Fernanda Birolli. *Liars and Circles: Essays on Truth, Self-Reference and Paradoxes*. Dissertation (master's degree) – Universidade de São Paulo. Faculdade de Filosofia, Letras e Ciências Humanas. Departamento de Filosofia. 2023.

This MA dissertation comprises four chapters that investigate philosophical and logical aspects related to paradoxes, self-reference, and circularity. The first chapter provides possible responses to the question "What is a paradox?." It begins by exploring the etymology of the word "paradox" and its potential connection to nonsense, with the aid of the philosophers Wittgenstein and Deleuze in their conception of nonsense. It ends with the examination of arguments and the judgment whether they are paradoxical or not. In the second chapter, I provide a formal study of logical paradoxes, including the Liar, the Heterological, and the Yablo Paradox; in the discussion about the latter paradox, I identify a lack of explicit definitions for self-reference and circularity. This topic is taken up again in third chapter, where I provide precise definitions for the concepts of reference, self-reference and circularity using model-theoretic interpretations and demonstrate properties about well-known paradoxes. The fourth chapter delves into semantically closed languages and theories, presenting a formal definition and a construction of a first-order bisorted semantically closed language. Moreover, it discusses whether everyday languages such as Portuguese or English are semantically closed or not. This dissertation offers an exploration of truth, self-reference, and paradoxes, bridging philosophy and formal logic.

Keywords: Paradox, Liar, Self-reference, Circularity, Semantic Closure.

Contents

Foreword

This dissertation tells the story of the *evasion* of a certain logical concept; namely, that of *paradox*. The reader will witness how, whenever we try to confine paradoxes within firm boundaries, either by providing a unified definition or a list of necessary attributes, they somehow manage to slip through the cracks and evade each and every wall we lifted in an effort to capture them.

It is a widely held belief among logicians that the families of semantic and set-theoretic paradox – the ones in which I am most interested in this text – bear inviolable ties to the concepts of self-reference and circularity. According to this entrenched conviction, self-reference and circularity are the very *roots* and *causes* of semantic and set-theoretic paradoxes. However, this research led to the unsettling conclusion that antinomies evade these characterizations, for there is at least one semantic paradox, called Yablo's Paradox, which is neither self-referential nor circular. Hence, the root of paradoxicality, if there is any, must lie somewhere else than in circularity or self-reference.

This elusive nature is also manifested in the fact that, to provide a unified definition of the concept of paradox, one needs to rely on slippery notions such as *plausibility* – to say, for instance, that a paradox arises when *plausible* premises lead to contradiction. Defining what one means by "plausible premises" proves to be a task as hard as discovering what exactly a paradox is. Even when we attempt to capture paradoxes through the construction of a semantically closed language (one that can create self-referential sentences), achieving it is a difficult task.

It is thus fitting to the theme of the dissertation that it finishes with such an *aporetic* spirit: there is more to paradoxes than we can grasp with the usual definitions and characterizations – but what is it? This apory is an evanescent yet constant presence throughout the following pages. This, however, is a metalinguistical conclusion: a result that is achieved *about* the dissertation after one reads it whole, but is not contained in any of its chapters.

As for the text's style, I believe there are two types of dissertations: those that aim, from beginning to end, to answer one single question; and those of a more essayist style, a stroll over a bundle of themes and problems that cached the writer's eye. I believe mine belongs to the second kind: it is not inaccurate to describe it as a *patchwork quilt*, an amalgamation of different problems sown together as a structured whole, rather than an in-depth exploration of one single problem. But this does not imply that there are no guiding threads. After

all, even a patchwork quilt is stitched by one single needle. The needle that binds this text together is the very concept of paradox, which is instantiated in the Liar, Yablo's, Russell's and other antinomies that are explored here. Self-reference and circularity, the two satellites that orbit the text, are about as important as the needle itself: even if my primary reason to pursue them was to evaluate self-referential or circular paradoxes, I defined those concepts so that they could be applied to sentences in general, regardless of their paradoxical nature. Semantically closed languages and theories, the theme of the fourth chapter, are the vessels in which all the previous concepts come to life, for languages or theories can *be* self-referential or circular and *contain* paradoxes.

On the text's structure

The first chapter of this dissertation consists of a philosophical exploration of the question "*what is a paradox?*" from many distinct perspectives. It begins by delving into the etymology of the word and how its lexical origins can shape its meaning. During this analysis, a suggestion arises: that the concept of 'paradox' might be connected with that of 'nonsense.' To examine this connection, I turn to two influential philosophers: Wittgenstein and Deleuze. I present each author's conception of nonsense and ask whether they can be related to the notion of paradox. From this, we ascend to another parallel, one that approximates the *role* of paradoxes to that of *apory*: just as the Socratic apories were means to purge one's false beliefs, reaching paradoxes may purge us from the dogmatic image of thought as a faculty of clarity and simplicity. After this, I descend again to the structural realm of the analysis to try to find an adequate formal definition for the concept of paradox. This proves to be a harder task than it appears at first, for the definition must depend on vague concepts such as "plausibility" or "analyticity." The chapter finishes with a stroll through some contradictory arguments under the lenses of the definitions of paradox that were analyzed in the previous sections, in an attempt to determine whether they are truly paradoxes or not.

While the first chapter is of a purely philosophical nature, the second one is nearly completely formal. Together, they set the tone of the text: the third and the fourth chapters are attempts to bind together the philosophical vein of the first with the formal vein of the second. The two last chapters are different from the two first in their *propositive* nature: while the two first chapters are explorations of concepts that are extensively approached, the last two are centered

in propositions of my own.

I begin the second chapter with an exposition of the necessary ingredients to create a Liar type of paradox: I introduce logical principles such as the Law of Excluded Middle, the *Ex Contradictione Quodlibet* and the Tarskian T-schema. I then present numerous proofs of the Liar with comments. This is followed by the exposition of the Heterological Paradox, another semantic paradox that prompts an interesting discussion about *types* in logic. In writing one of the chapter's sections, my attention was grabbed by a topic which I thought to be both quite problematic and less approached in the literature: the Yablo Paradox, which is explored in section 2.4. This paradox consists of an infinite chain of sentences, each of which states that all the other ones up the chain are false. The important question here, as I mentioned earlier, is whether it is circular, self-referential or neither – on the one hand, none of the sentences refer to themselves explicitly, so it seems to escape any sort of circularity; on the other hand, the sentences contain a hidden fixed point, so the paradox might be circular. While becoming acquainted with the state of the art of the literature in the topic, I noticed what I judge to be a deep problem: when considering whether Yablo's Paradox is circular, self-referential or neither, logicians never seemed to define those terms explicitly.

Puzzled about the lack of clarity surrounding these concepts in the logic literature, I plunged on the endeavor that came to be the central theme of the third chapter: defining the concepts of *reference, self-reference* and *circularity* formally. The first conclusion I reached is that the mainstream approach to them in logic, which posits that circularity (or self-reference, undistinguished) is achieved through fixed points, is deeply flawed, and we need to provide better definitions to model these concepts. I then found formal definitions of reference, self-reference and circularity rooted in the model-theoretic notion of *interpretation*. Next, I justified the reasons why my definitions should be adopted, even though I believe (or at least hope) that they possess such a simple and transparent character that they will become immediately clear to any reader who understands the meaning of each term employed.

Through my definitions, I was able to prove that the Liar Paradox is self-referential and the Liar Cycle is circular, which were the expected results. Moreover, I proved that Yablo's Paradox is *neither* self-referential *nor* circular, contrary to the dominant view in logic that there is a hidden circularity in it. After these proofs, I was faced with a difficult issue: Russell's Paradox, whose circularity

is so widely assumed that it is not even questioned by logicians, was found to be non-circular (and non-self-referential) through my definitions. After some thinking, I found that there is a good and simple reason why the paradox does not fit into them: it is *non-well-founded,* rather than circular or self-referential. This is not a new conclusion; the proof that the Russell "set" is non-well-founded is widely known in mathematics. What is different, here, is the distinction between non-well-foundedness and circularity: they are usually believed to be synonyms. This leads me to inquire whether the Yablo Paradox can be considered as non-well-founded as well – a question that I leave unanswered for the time being. Finally, the chapter ends with the consideration of some possible objections to my thesis and ways to respond to them.

The fourth and final chapter explores the concept of semantically closed languages and theories. In general, semantically closed languages (theories) are those capable of expressing semantic facts about themselves – *i.e.*, those that contain semantic predicates that apply to the names of their own sentences. The chapter is divided into two main sections, one which explores the formal counterpart of the concept, and the other its non-formal philosophical counterpart. In the first part, I first define the concept formally, and then use this definition to construct a first-order bisorted semantically closed language. Although there are records of constructions of semantically closed theories, a semantically closed language such as the one I present here (with its necessary bisorted nature, for example) had not been done yet – or at least not to my knowledge. Still in the formal counterpart of the chapter, I present the sketch of a semantic closure operator: a monotonic function which, when applied to theories, does the job of closing them semantically. The construction is only outlined in this dissertation – its completion rests as a future endeavor. In the non-formal counterpart of the chapter, I dive deep into the question "*are everyday languages, such as Portuguese or English, semantically closed*?." In this matter, I side with Graham Priest and defend that they should indeed be regarded as such. I leave it to the reader to decide whether or not my arguments are convincing.

On the text's prerequisites

As I wanted this dissertation to be as self-contained as possible, I tried to define and explain every term, rule or symbol that figures in it, so that everyone could read and understand it, regardless of any background knowledge on any of the topics I explore here. But it is easier said than done: this is, after all, a dissertation

in logic, a subject known for its aridity. It would be presumptuous of me to affirm that literally anyone can understand this text; I will limit myself to saying that no specific prior knowledge is expected of the reader. To provide a firmer historico-philosophical basis on possible solutions to the Liar Paradox, I included an appendix at the end of the text that can be consulted at the reader's will. It bears no direct relation to the other chapters of the text besides standing as an implicit framework to some of what is discussed in them.

Chapter 1

What is a Paradox?

1.1 The Etymology

From Plato to Jorge Luis Borges, G. K. Chesterton or Aristotle, countless thinkers have devoted some thought on the problem of paradoxes. But it was only in the first half of the Twentieth Century, with the flourishing of the field of mathematical logic, that it became clear beyond the shadow of doubt that paradoxes are not simple harmless puzzles or unfair arguments, but rather *linguistic constructions that unveil deep problems on the core of reasoning*. The problem could not achieve this importance if it wasn't for Kurt Gödel and his Incompleteness Theorems, which showed that there is no proof of the absence of paradoxes in the heart of mathematics. But how are paradoxes in literature, philosophy, mathematics and logic related to each other? In other words, what is a paradox? Can we find a broad definition of paradox that merges all its different existing types? I will meditate on this issue before delving into the Liar Paradox itself.

The term "paradox" comes from the Greek word *para* (beyond) and *doxa* (belief). A paradox is, therefore, that which is beyond belief. In its original sense, as explained by Nicholas Rescher, they are incredible contentions, bizarre opinions, arguments against common sense.[1] This does not, however, end the matter, for the meaning of "beyond belief" can be taken in two quite different ways: on the one hand, something that is beyond belief may simply be nonsense, an assertion that is unbelievable precisely because it is false. On the other hand, something beyond *doxa* can be taken as what is beyond *mere* belief, mere opinion; which is beyond what merely *appears*. The etymology of the word seems to project two diametrically opposed meanings, both equally applicable to the word "paradox."

[1] See [44], p. 3.

The second possibility of interpreting the word's etymology finds echo in the first Greek philosophical thought. Parmenides, one of what we call the Pre-Socratic philosophers, is assertive about the bad connotation of *doxas* and how they stand in the way of truth. The following excerpt is taken from his poem *On Nature*, of which only some parts were preserved. The lyrical self starts by narrating his arrival in the domains of the Goddess, and then transcribes her words to him:

> Welcome, O youth, that comest to my abode on the car that bears thee tended by immortal charioteers! It is no ill chance, but right and justice that has sent thee forth to travel on this way. Far, indeed, does it lie from the beaten track of men! Meet it is that thou shouldst learn all things, as well the unshaken heart of well-rounded truth, as the opinions of mortals [*brotōn doxas*] in which is no true belief at all. Yet none the less shalt thou learn these things also,— how passing right through all things one should judge the things that seem to be ([32], p. 128).

The philosopher's usage of the word *doxa* points to a mere belief or false opinion. In this sense, something beyond belief could be characterized as that which lies beyond the bounds of opinion, the false beliefs induced by the senses. Therefore, with this interpretation of *doxa*, paradoxes may lie beyond common sense not because they are merely false assertions, but because they express something that cannot be understood or apprehended by common sense.

I said earlier that one can understand the etymology of the word "paradox" in two different ways: something may be beyond belief because it is simply nonsense, or it may be beyond *doxa*, "in which is no true belief at all" (Idem, ibidem). We've concluded that, through the second alternative, paradoxes may teach us something that is not available to common sense. Now, we must examine the first alternative, according to which paradoxes are *simply nonsense*, assertions that are unbelievable because they are, in some way, false. Well, this description is not yet fully precise: what distinguishes a falsehood from a nonsensical proposition? Most false sentences cannot by any means be said to be nonsensical. "Snow is black," "the earth is flat," "spiders are mammals" are all examples of false, but never nonsensical, sentences. The concept of nonsense demands a more thorough examination, so that we can understand if and how it can disclose the nature of paradoxes.

1.2 Paradox and Nonsense

The historian Alexandre Koyré, in [26], provides an interesting description of the Liar Paradox and an explanation of the reason why it is so difficult for us to comprehend paradoxes. Although he refers only to the Liar, we can ask if his account could be generalized to all paradoxes.

> Or, nous avons l'habitude de parler pour dire quelque chose, d'entendre des phrases qui ont un sens, ou du moins qui veulent en avoir un. Aussi rien ne nous est-il plus difficile que d'appréhender un non-sens: nous mettons un sens partout où il y en a pas. ([26], p. 12.)[2]

On footnote, Koyré completes: "comme nous mettons de l'ordre et de la logique dans nos rêves."[3] The author characterizes the Liar as an excerpt of nonsense – and nonsense, here, is not to be understood as gibberish, but in its literal meaning of something that is devoid of sense; or, better, that is in the opposite direction of sense. Koyré is not arguing that paradoxes are simple falsities, a hypothesis to which I have alluded in the end of section 1.1, but that they *escape* our usual way of reasoning. As it turns out, nonsense has little or nothing to do with falsehood. The question posed by Koyré is the following: if we usually speak to say something, and adorn all things with meaning (even our dreams), how are we to understand sentences that state *no thing*? In what follows, I will investigate the concepts of paradox, nonsense and the relationship between them in Wittgenstein and Deleuze. Then, I will meditate on the relationship between those concepts and that of *apory*.

1.2.1 Wittgenstein

The concept of nonsense is central in Wittgenstein's *Tractatus Logico-Philosophicus* [54], one of his most famous texts. In fact, it is intimately tied to the very aim of the work, as the philosopher writes in the preface:

> Thus the aim of the book is to draw a limit to thought, or rather — not to thought, but to the expression of thoughts: for in order to be able to draw a limit to thought, we should have to find both sides of the limit thinkable (i.e. we should have to be able to think what cannot be thought).

[2] My translation: "Now, we are used to speaking in order to say something, to hearing sentences which have a meaning, or at least which want to have one. So nothing is more difficult for us than to apprehend a nonsense: we put meaning wherever there is none."

[3] Idem, ibidem. My translation: "As we put order and logic in our dreams".

> It will therefore only be in language that the limit can be drawn, and what lies on the other side of the limit will simply be nonsense ([54], p. 3).

First, he postulates that the text's objective is to draw a limit to the expression of thoughts. To do that, he adds, one must go beyond the limit and think what cannot be thought; and what lies outside the limit is called nonsense. If one must cross the limits of thought to find its limits, one must enter the realm of nonsense. Indeed, further in the text he will explain that the very book is *nonsensical*:

> My propositions serve as elucidations in the following way: anyone who understands me eventually recognizes them as nonsensical, when he has used them – as steps – to climb up beyond them. (He must, so to speak, throw away the ladder after he has climbed up it.) He must transcend these propositions, and then he will see the world aright ([54], proposition 6.54).

This view of nonsense once more corroborates the first etymological interpretation I proposed: to say that paradox is "beyond belief simply because it is nonsense" does not actually mean that paradoxes are chimeras, but that they lie outside the limit of thought and, according to Wittgenstein, one must go beyond the limit to be able to trace it. This strengthens the idea that paradoxes go beyond "mere opinion," in which no truth can be found. At last, the two pictures I proposed in section 1.1 seem to collapse into one, which takes paradoxes as the road outside *doxa*.

However, we cannot go as far as to say that, for Wittgenstein himself, paradoxes are nonsense. First, we must examine the difference he draws between two words: *Unsinn*, translated by Pears and McGuinness as "nonsense," and *Sinloss*, which they interpret as "senseless" ([30], topic 2). Tautologies and contradictions lack sense, but are not nonsense ([54], prop. 4.461): tautologies are senseless because they have no truth-conditions, they are unconditionally true, while contradictions are senseless because they are true under *no* condition. Thus, tautologies and contradictions say nothing about the world, they do not *represent* possible situations in any way, for tautologies accept all situations and contradictions accept none ([54], prop. 4.462). For example, the sentence "it is raining or it is not raining" tells me nothing about the weather, and neither does "it is raining and it is not raining," which is true under no condition. In this way, they are senseless, but not nonsensical: they lie within the boundaries of language and thought, not beyond them.

This might lead us to believe that paradoxes, in Wittgenstein's sense, are not nonsensical, but senseless – after all, contradictions seem to be essential to paradoxes, and they are taken to be merely senseless. However, paradoxes are not simply contradictions. Let's take the Liar Paradox as an example: as will be better explained in chapter 2, the paradox happens when we seek to decide whether a certain sentence L, stating "L is false," is true or false. If we assume that L is true, then what it says is true; but, given that it says of itself that it is false, we must conclude that it's false. On the other hand, if we assume that L is false, then what it says must be false; but, given that it says of itself that it is false, we conclude that it is true. We can see, through this intricate argument, that the Liar is not a contradictory sentence such as "$p \wedge \neg p$," which would, according to Wittgenstein, be called senseless. It is, rather, an *argument* where both assumptions – that the sentence is true or that it is false – lead to contradiction. This is a subtle but crucial point: while a sentence such as "it is raining and it is not raining" is simply a contradiction, the Liar Paradox is a *logical impossibility* to the eyes of classical logic.

Judging from these data, one cannot quickly decide what would be the status of (logical) paradoxes for Wittgenstein: if they are nonsensical, senseless or should be placed in yet another category; a deeper analysis would be needed to determine it with more certainty.

We haven't yet examined in detail the reason why the very propositions of the *Tractatus* are nonsense. By taking a closer look at it, we will realize that it is the paradoxical nature of those propositions that render them nonsense. In fact, as we will see, the heart of the matter is that we cannot call them "propositions." In [39], the logician Graham Priest walks the reader through Wittgenstein's argument with a degree of clarity which is hard to achieve, specially when it comes to such a dense and convoluted philosopher. Thus, I will follow Priest's steps and walk the reader through why Wittgenstein's claim that his very book is nonsensical.

First, we must have in mind that the world is the totality of (atomic) facts ([54], prop. 1.1). Facts, in their turn, are collections of objects that are arranged in a specific way: they cannot be connected randomly; instead, their connection must be determined by possibilities intrinsic to the objects (see [39], p. 204). The manner in which they connect is called the *form* of the fact. On the other side, we have language, which is constituted by propositions. "Structured" propositions have their truth values determined by atomic propositions, the entities they are composed of. Additionally, atomic (and non-atomic) propositions are composed

of *names* that correspond to *objects*. Just as facts are not mere bundles of objects, propositions are not mere bundles of names: the latter must be structured in a certain manner to form propositions, and such a manner is called the *form* of the proposition.

Let us recapitulate: on one side, we have the world, which is made of facts – which, in turn, are made of objects. The way objects are assembled in a fact is called the form of the fact. On the other side, we have language, which is made of propositions; those are, in turn, made of names. The manner in which names are assembled in a proposition is called the form of the proposition. It is not hard to conclude that an atomic proposition is true if and only if there is a fact whose objects are objects named by the proposition, and whose form is equal to that of the proposition. As Priest states it, a proposition is true if and only if its names are isomorphic to the objects of a fact ([39], p. 204).

To get to the main point, we still need to address two preliminary ones: first, the distinction between *saying* and *showing*. A proposition *says* that objects in the world are such and such; it expresses something about certain objects. But it also *shows* its own form; the proposition "snow is white," for example, not only says something about objects but shows the way names are assembled in it (its form). The second point has to do with *sense:* for Wittgenstein, sense is determinate. This means that each proposition has a sense, and all senses of all propositions can be determined. Their determination is made through a process of philosophical analysis, which breaks down facts into simpler facts, until one reaches the realm of objects, which are *simple.* By "simple," Wittgenstein means that they have no constituents – they cannot be reduced any longer to other smaller facts.

Simple objects are *necessary* to avoid an infinite regress of senses. Say we want to determine the sense of proposition 1. Its sense is given by *analysandum 1.* If 1 is not already about simples, it can be analyzed. Through this analysis, we reach *analysandum 2*, which determines the sense of *analysandum 1.* If we cannot reduce the analysandums to simple objects, we fall in a regress: the sense of a proposition must be determined by another, and that by yet another. Thus, we would not have access to the senses of propositions.[4] But, for Wittgenstein, we *do* have access to senses – if we didn't, we wouldn't be able to "create a picture of the world, whether true or false" ([54], prop. 2.0212). Hence, because senses are determinate, there must exist simple objects.

[4]See argument in [39] p. 206.

From the distinction between saying and showing, we conclude that "structural *facts*"[5] – facts about propositions, about the structure of facts etc – are *shown*, but cannot be said. They cannot be said because propositions are composed by names, and names correspond to objects in the world. If structural "facts" could be said, then they would have to be constituted by objects in the world; but this cannot be the case, because they are the very structure that binds objects together. An example might make it clear: if a proposition could say something about the form of propositions, the form of propositions would have to be an object of the world. But it cannot be an object of the world, because it is what determines the structure of propositions themselves. Even clearer than my explanation is Priest's commentary on this:

> Thoughts are articulated. To form them we must combine simpler building-blocks. But thoughts are no mere lists of their components. There must therefore be things which hold them together as unities. Let us call these (with apologies to modern physics) gluons. Gluons are not the same kind of thing as the components they glue, and hence not the kind of thing one can express claims about ([39], p. 212).

Thus, we cannot say anything about gluons, just as we cannot use a canoe itself as another wooden board to build a canoe. But, Priest adds, "anything can be an object of thought; in particular, we can think about gluons. Thus they are the same kind of thing as other constituents, and we can express claims about them" ([39], p. 212). In fact, the very *Tractatus* is all about gluons, for the entire book talks about the form of facts, propositions, world and language. Finally, we reach Wittgenstein's conclusion: the propositions in his book *must be nonsense (Unsinn)*, for they talk about things which one cannot talk about. Since propositions have sense and sense is determinate, if we conclude that the propositions in the book are nonsense, it follows that they cannot even be called propositions. Just then, when the reader has realized such conundrum, he can throw away the ladder, transcend the propositions and see the world aright.

Priest, however, sees this conclusion as a sign that something has gone wrong in the philosopher's argument. After all, he is telling the reader that the whole book is nonsense, when the reader has read *and understood* the book and what each of its propositions says. We can only understand what has sense; thus, the

[5] This is a terminology used by Priest in [39]. He uses it to make the argument easier but recognizes that "fact," in this locution, is not the Wittgensteinian kind of fact – it cannot be so because, as we will see, there are no structural facts.

book cannot be nonsense. One might reply, in Wittgenstein's defense, that we understood not what the propositions said, but what they showed. However, that will not do: a proposition shows *something* about its form: that it is a proposition, that it relates such and such components. But if the *Tractatus* claims are nonsense, they are not even propositions, so they cannot even show their propositional form. Alternatively, one could claim that the book's propositions are literally ineffable. That wouldn't necessarily be false, but Priest claims that it would be strange: what would it mean for the Tractarian propositions to be ineffable? They seem perfectly effable in the eyes of the reader of the book.

One way out of the paradox – which Priest claims will be adopted by many thinkers after Wittgenstein – is to reject that sense is determined and, ultimately, accept the infinite regress of senses. Maybe sense is not grounded in simple objects after all, maybe there's no way to get to the bottom of the meaning of a proposition. This position will be adopted by Deleuze, which I will examine in the next section. First, however, let me show a possible deduction of Wittgenstein's argument and Priest's objection to it. It is extremely reductionist and is supposed to be so: I tried to extract only the absolutely necessary components that lead to the conclusions wanted. My hope is that the derivation can clarify the logical structure of the arguments, even if it sacrifices some important components involved in Wittgenstein's reasoning.

Argument for the nonsensical status of "structural propositions"

Premise 1. Propositions *show* their own form, but cannot *say* that their forms (structures) are such and such.

Premise 2. The sense of propositions is determinate.

Premise 3. A proposition that supposedly says something about what cannot be said is nonsense.

Derivation.

If premises 1, 2, and 3 are propositions, then by premise 2 they have a determinate sense. By premise 1, propositions cannot say that their forms (structures) are such and such. But premises 1, 2, and 3 are about the structure of propositions. Thus, by premise 3, premises 1, 2, and 3 are nonsense.

Priest's objection

However, **(Premise 4)** I can only understand what has a determinate sense and **(Premise 4')** I understand premises 1, 2, and 3. Thus, 1, 2, and 3 cannot

be nonsense, which is a contradiction with the previous conclusion.

Therefore, we must reject one of the premises 1, 2, 3, 4 or 4'. We can choose (as many authors did) to reject premise 2 and admit that sense is *underdetermined*. If we do so, we cannot say that premises 1, 2, and 3 are *about* the structure of propositions, for they would not have a determinate meaning. This way, the contradiction can be avoided.

There is a final additional remark that was not mentioned in Priest's interpretation: to reject premise 2, either one must reject premise 4 along with it, or admit that we do not understand propositions at all. If we accept the infinite regress and grant that propositions do not have a determinate sense, then by premise 4 we cannot understand any proposition. Another alternative would be to reject one of the other premises – perhaps premise number 1, for example: maybe we can, after all, talk about the forms of propositions. This alternative is tempting due to the simple fact that we seem to have been doing so for the length of this whole section. Wittgenstein, however, would certainly not let this premise go, as this seems to be a crucial hypothesis of the *Tractatus*. He might have been content with the slightly jocose rejection of premise 4', suggesting that we have actually not understood premises 1, 2, 3 or the *Tractatus* as a whole.

1.2.2 Deleuze

Now, let us drift to another conception of nonsense, which certainly includes paradoxes: that of the french philosopher Gilles Deleuze. In case anyone finds it strange to place such distinct philosophers one after another, I will say that Deleuze and Wittgenstein have a notable similarity of style: they are both obscure, cryptic, and somewhat unsystematic. In addition, as I will show below, they are both interested on the problem of paradoxes and the inexpressible.

We may begin this analysis with the following excerpt, from [13]: "[...] le mécanisme du non-sens est la plus haute finalité du sens, de même que le mécanisme de la bêtise est la plus haute finalité de la pensée" (p. 201).[6] Surprisingly, both authors, whose philosophies are often considered as opposites, give nonsense an extremely central role: according to Wittgenstein, nonsense is the realm outside of thought from which the limits of thought can be traced and, according to Deleuze, it is the *very goal of sense*.

[6]My translation: "the mechanism of nonsense is the highest purpose of sense, just as the mechanism of gibberish is the highest purpose of thought."

Two steps must now be executed: to examine the way in which nonsense is the goal of sense, and to consider how and why paradoxes are said to be nonsense. Only after these steps the role played by paradoxes in Deleuze's philosophy will be unveiled.

Deleuze first exposes the wide spread belief that that the realm of sense is attached to the *expression* of a proposition, but separated from its *designation* – which is the realm of truth and falsity ([13], p. 198) – only to later reject this separation. In this view, sense becomes a question of mere logical formalism or a psychological matter, for it would have no effect on the truth or falsity of propositions, but only on the way they are expressed (their intended meaning; their logical form). Instead of rendering sense such a weak, unimportant concept, we should view it as that in which the relation between proposition and the object it designates is established. Sense goes beyond the the linguistic dimension of proposition, towards the designated object.

In this picture, it is difficult to determine the sense of a proposition: it is not entirely linguistic nor entirely in the designated object. It is, rather, in between those two worlds, so much so that Deleuze even calls it *extra-propositional*. Mind that he is not denying that sense is tied to what is expressed in a proposition – it is intimately tied to it, but its expression cannot be reduced to the designated object, to the state of the utterer or to its linguistic form. Since this reduction is impossible, we can never formulate the sense of a proposition in the very proposition itself. For example, the proposition "snow is white" says something about an object in the world, but does not express its *own sense*. However, we could still express the sense of a proposition by appealing to another: we could make the *sense* (the expressed) of a proposition the *designated* of another. Thus, a second proposition would designate the sense of the first. But now, if we want to establish the sense of the second proposition, we need a third that designates it; to establish the sense of the third, we need a forth; and so on indefinitely. This is exactly the problem faced by Wittgenstein, which he solves by assuming that there are simple objects. If the reader goes back to the derivation in the end of section 1.2.1, she will notice that Deleuze rejects premise 2 of the argument.

Deleuze baptizes this infinite regress of propositions expressing the senses of others as *paradox of proliferation,* the first of the "paradoxes of sense" ([13], p. 202). From it, one can conclude that we can never grasp the sense of *all* propositions; if we try to do so, we will be trapped in an endless regress. It is possible to escape this paradox, but only to fall into another: the *paradox of the*

neutralizing unfolding. We may try to escape the first paradox by suspending the proposition and retaining only its "ideal content," its expression. This content, as I said before, is neither in the proposition itself nor in the object it designates, so it cannot be fully located somewhere or another. In Deleuze's vocabulary, it *almost is*, or it has an *extra-being*. Sense is only a vapor moving within the limits between words and things. The philosopher provides two pictures of the paradoxes found in Lewis Carroll's two masterpieces of fiction ([9] and [10]): the paradox of the neutralizing unfolding is pictured in the smile of the cat without the cat, while the paradox of proliferation is pictured in the knight who names the name of his song, the name of the name and so on. Below, I reproduce the ingenious passage of the knight to illuminate the idea behind Deleuze's paradox of proliferation:

> 'You are sad,' the Knight said in an anxious tone: 'let me sing you a song to comfort you.'
> 'Is it very long?' Alice asked, for she had heard a good deal of poetry that day.
> 'It's long,' said the Knight, 'but very, very beautiful. Everybody that hears me sing it— either it brings the tears into their eyes, or else— '
> 'Or else what?' said Alice, for the Knight had made a sudden pause.
> 'Or else it doesn't, you know. The name of the song is called "Haddock's Eyes."'
> 'Oh, that's the name of the song, is it?' Alice said, trying to feel interested.
> 'No, you don't understand,' the Knight said, looking a little vexed. That's what the name is called. The name really is "The Aged Aged Man."'
> 'Then I ought to have said "That's what the song is called"?' Alice corrected herself.
> 'No, you oughtn't: that's quite another thing! The song is called "Ways and Means": but that's only what it's called, you know!'
> 'Well, what is the song, then?' said Alice, who was by this time completely bewildered.
> 'I was coming to that,' the Knight said. 'The song really is "A-sitting On A Gate": and the tune's my own invention' ([10], p. 71).

This is the way in which sense leans towards nonsense and, furthermore, has nonsense as its very goal: by trying to determine the sense of propositions, one reaches paradoxes that lead to nonsense. Paradoxes are, in a way, roots to nonsense: from the perfectly regular activity of trying to determine the sense of

propositions, paradoxes guide us through a forest of infinite regress, being and non-being. It now becomes clear that nonsense is not the opposite of sense, or the lacking of it: it is engendered in sense itself. Even though this description may sound esoteric to some, it is quite accurately aligned to the mechanism of logical paradoxes, which are so extensively analyzed by the Anglo-Saxon philosophical tradition. The Liar Paradox, for example, as I will show in the next chapter, arises from simple sentences such as "This sentence is false." Not involving any complex concepts and being perfectly understandable, this sentence generates one of the deepest paradoxes of all – again, we witness the moving arrow from sense to nonsense.

Moreover, this path from sense to nonsense has something like a freeing quality to thought, as explained in this passage: "C'est que le paradoxe s'oppose à la *doxa*, aux deux aspects de la doxa, bon sens et sens commun" ([14], p. 93).[7] The meaning of the etymology of the word "paradox," as analyzed in section 1.1, is endorsed here: paradoxes are opposed to *doxa* which, for Deleuze, is made of *bon sens* (which I translated as good judgment) and *sens commun* (common sense). The technical difference between those two aspects of doxa is not so important here – it suffices to say that *bon sens* is tied to the necessity of an order and of choosing one – only one – direction within this order and following it. This is what orients, for example, the established order of time: the need for going from the past towards the future. *Sens commun,* in its turn, is a faculty of identification, which reports a diversity of elements under the "form of the Same" ([14], p. 96); it is the *sens commun* which permits us to regard two different states as belonging to the same subjects, or unite many different aspects under one sole Self.

The freeing aspect I alluded to in the last paragraph is given by the opposition between paradoxes and *doxa*. In chapter three of [13], Deleuze provides eight postulates of the *dogmatic* image of thought, which are obstacles to non-dogmatic thought and the philosophy that might come from it. One of this postulates is precisely that of doxa: these two aspects of doxa, that things have to be ordered and must follow only one route, and that different aspects must be united under one sole Self, are obstructions in the way of thought. Therefore, paradoxes, as the route from sense to nonsense, should not be avoided as obstacles to thought or ignored as silly charades, but exactly the opposite: they should be examined as deep concepts capable of removing barriers from thought. Lastly, thought is

[7] My translation: "It is that the paradox is opposed to *doxa*, to the two aspects of doxa, common sense and good judgment."

not a clear faculty which is darkened by paradox and nonsense – thought is not clear and simple in the first place, the idea that it is so comes from a dogmatic picture of thought. Therefore, paradoxes do not darken what was once clear, but show that the supposed clearness was only a dogmatic mirage.

1.2.3 Paradox and Apory

The idea that paradox is opposed to *doxa* and that nonsense is the very goal of sense allows for an interesting parallel between two notions: paradox and *aporia* (in English, apory). *Aporia* is a Greek word that can be translated as "lacking passage," "puzzlement" or "impasse." Already in antiquity, it was not a simple puzzle, but a way to *purge one's false beliefs*. The one who exercises this method in its deepest form is undoubtedly Plato: many of his dialogues result in apories, with Socrates driving the characters to a state of total puzzlement brought by insidious contradictions from the assumed premises. Thus, apories are far from being sophistic methods to win arguments – they have a profound *teaching effect*. In the dialogue [33], Socrates asks Meno's servant what would be the size of the side of an eight square feet quadrangle. The boy answers that it should be a line four feet long, and Socrates drives him to realize that this is wrong, since this line would be the side of a sixteen square feet quadrangle. When the servant realizes his mistake, Socrates turns to Meno and says:

> There now, Meno, do you observe what progress he has already made in his recollection? At first he did not know what is the line that forms the figure of eight feet, and he does not know even now: but at any rate he thought he knew then, and confidently answered as though he knew, and was aware of no difficulty; whereas now he feels the difficulty he is in, and besides not knowing does not think he knows ([33], 84b).

I will overlook the delicate usage of the word "recollection" instead of "learning," since such discussion falls out of the scope of this text.[8] In the above passage, Socrates characterizes the servant's state of apory as *progress:* at first, the boy did not know the answer to Socrates' mathematical challenge but thought he knew; after the conversation with Socrates he still does not know the answer, but

[8] One of Socrates' aims in the dialogue is to show that the servant already had mathematical ideas in his soul, and by conversing with Socrates the servant is merely recollecting them. Thus, Socrates is not teaching the boy anything, but only bringing back the mathematical knowledge that was already in him. This is why the philosopher employs the word *recollection* and not *learning*.

at least *he knows that he does not know*. We might even say that, before Socrates' intervention, the boy was submerged in a state of *doxa* and, after reaching an apory, this false belief was purged and now he is aware of his own ignorance.

I maliciously hinted at the relation between doxa and apory to induce the reader into drawing a relation between apory and paradox: just as an apory has the teaching effect of eliminating one's false beliefs, paradoxes have the teaching effect of distancing one from *doxa*, which stands in the way of knowledge and a true philosophy. Both apories and paradoxes make us realize the depths of the difficulty we are in; both go against the dogmatic image of thought as a faculty of clarity and simplicity; both show that things can be more complex than they seem. In fact, Rescher, in [45], goes as far as to say that "paradoxes are the very model of apories arising when we have a plurality of theses, each individually plausible in the circumstances but nevertheless in the aggregate constituting an inconsistent group" (p. 84). Thus, paradoxes and apories are intimately connected: since paradoxes usually arise from commonly accepted premises and result in contradictions, we can say that they are always set in an *aporetic stage*.

1.3 Logical Paradoxes

In the previous sections, I investigated paradoxes in a general sense, in the word's etymology and its approximation to nonsense. When logicians define paradoxes, particularly logical paradoxes, they typically employ a distinct approach from the one I have utilized thus far. They often provide a precise and commonly accepted definition for the concept. In this section, I will introduce some versions of this definition and evaluate their accuracy.

The most common definition of paradox among logicians is something along these lines: it is said that a paradox arises when *plausible* premises $\{p_1, ..., p_n\}$ logically imply a given conclusion C such that its negation $\neg C$ is also plausible. So each of the assumptions is plausible (as is $\neg C$), but the set $\{p_1, ..., p_n, \neg C\}$ is inconsistent. This is the definition of paradox provided by Rescher in [44] (p. 6), for example. He explains the definition above by saying that a paradox emerges when a set of individually plausible propositions is collectively inconsistent.

Let us examine this definition in more detail. Its first intriguing feature is that it makes use of quite a slippery notion: that of *plausibility*. What does it mean that a premise is plausible? The concept of plausibility seems to depend more on external factors (social groups; different eras) than we would want it to.

For example, take the Paradox of the Barber (a weaker kin of Russell's Paradox): there is a village in which lives a barber that shaves all and only the men who do not shave themselves. And so I ask you: does the barber shave himself? Well, since he shaves only the men who do not shave themselves, he shaves himself if and only if he does not shave himself. Now, this certainly has the looks of a paradox: a set of premises which, taken collectively, lead to inconsistency. But the definition above presupposes the plausibility of the premises, and deciding whether the premise "in a certain village, there exists a barber who shaves the beards of all and only the men who do not shave themselves" is quite a difficult task. Some could say it does not look very plausible; after all, how would this barber operate? Every man who walked into the barbershop would be asked "do you shave yourself? If so, get out of here!" And what if the costumers lied? Clearly, this situation is unsustainable, one could say. At the same time, the situation does not offer any logical impossibility: it doesn't evoke round squares or any of such objects. So, someone else could argue, it is plausible in a deeper *logical* sense.

The problem is that, independently of the answer to the argument's alleged plausibility, it does not cease to be a paradox. If it is decided that the premise is, after all, implausible, this changes nothing in the core of the argument: it still moves, by sound reasoning, from its premises to a contradictory conclusion. This makes it seem absurd for a definition of paradox to say something about the "believable quality" of its premises. We shouldn't reach this conclusion so fast, however, for it is not entirely unprincipled to require plausibility to premises in a paradox. The reason for such requirement is less tied to the argument's soundness than to its *reach*: if an argument has very implausible premises, it may be an apory about which people wouldn't care much. In particular, logicians (the natural researchers of paradoxes) would not care one bit: they would simply reject the premise and the apory would be gone with it. The reason why Russell's Paradox is so important is because it comes from a premise which, at the time, seemed so plausible that Frege introduced it as one of the axioms of his set theory: the Unrestricted Comprehension Schema. I will present it in detail in section 3.4.4; here, it suffices to say that it expresses the idea that *every property determines a set*. From such useful premise, one can construct the Russell set, which is contradictory. Hence, what makes this paradox so important is that it forces Frege and his followers to review the edifice of set theory as a whole. The "quality" of a paradox can be measured by its insidiousness, and how insidious it

is has to do with how plausible its premises are.

A curious anecdote related to plausibility tells us that one of the first mentions of the word "paradox" in English, recorded in the Oxford English Dictionary from 1616, reads as follows: "Paradox, an opinion maintained contrary to the commonly allowed opinion, as if one affirme that the earth doth move round and the heavens stand still" ([44], p. 6). Hence, according to the Oxford dictionary, we are all immersed in paradox after the acceptance of Galileo's theory. Such are the dangers of having our definition of paradox depend on the mutable notion of plausibility (or "contrariness to common opinion," as the dictionary suggests): we might become subject to mockery by future generations which will have accepted as truths ideas that, for us, seem absurd. Given the difficulty of getting rid of plausibility and related notions, maybe we should make amends with unavoidable future mockery.

Ironically, we seem to have reached an apory ourselves: either we require that the definition of paradox mentions the plausibility of its premises, and so we permit an imprecise concept to sneak into our (otherwise precise) definition, or we allow for silly arguments to be called "paradoxes," which will be solemnly ignored by logicians. Is there a way out of this apory? Maybe. Allow me to present one more definition, by the English logician Graham Priest:

> The paradoxes are all arguments starting with apparently analytic principles concerning truth, membership, etc., and proceeding via apparently valid reasoning to a conclusion of the form α and not-α'. Prima facie, therefore, they show the existence of *dialetheias* [italics added] ([40], p. 9).

What first catches the eye in this definition is its conclusion: that, prima facie, they show the existence of *dialetheias*. Dialetheias are *true* sentences of the form "α and it is not the case that α." What is surprising here – and which is different from classical assessments of the problem – is the classification of paradoxical sentences as *true*. This reveals the central thesis of the theory elaborated by Priest, called *dialetheism*: that there are true contradictions. Of course this does not mean, as the author emphasizes several times, that *all* contradictions are true, but that there are some cases of such phenomenon, one of them being precisely the Liar sentence. If it is at all possible to talk about a theory's attitude, one should say that the attitude of dialetheism is to *accept the paradoxes*. Language does not need paladins of common sense, whose duty is to defend it from the threat of contradictions. On the contrary: it is necessary to accept them as

elements contained in the language. Language is ultimately inconsistent, but not trivial – and this is by no means a failure of reasoning.

Returning to the plausibility issue, in Priest's definition, "plausible premises" is replaced by "analytic principles." Now, the burden is thrown on deciding what is an analytic principle. However, at any rate, this definition is considerably more precise than the previous one, because analyticity is a concept much more technical and philosophically charged than "plausibility." Evidently, it has more than one definition, but only a handful: it is not stained by the contingency of time and space as plausibility, which may vary upon cultures and upon eras.

Analytic statements have been historically characterized as those whose truth depends solely upon the meaning of each word; these are opposed to *synthetic statements*, whose truth depends upon other external factors. An example of the first kind is the statement "all bachelors are unmarried," whereas an example of the second is "all dogs are cute." The first sentence is true simply because "being unmarried" is contained in the definition of the word "bachelor," so one need not know more about bachelors than the definition of the word to be able to evaluate the truth of this sentence. Since *cuteness* is not, prima facie, contained in the definition of the word "dog," evaluating the truth or falsity of the sentence involves other factors other than merely knowing the word's meaning. Although the opposition between analytic and synthetic can be put in simple terms, evaluating whether a specific statement is analytic or not is not at all a simple task, and has actually generated much disagreement among philosophers: a fruitful source of historical controversy has been, for example, mathematical statements, which some, such as Frege, believe to be analytic, and others, most notably Kant, believe to be synthetic.

Even if analyticity is more precise than plausibility, it is not without quarrel that one will reach a reasonable degree of certainty regarding mathematical statements. After all, how are we to decide whether a certain premise is an analytic principle? Is the premise "there is a village in which lives a barber that shaves all and only the men who do not shave themselves" an analytic principle? Hardly. Maybe the premise should be put as "*assume that there is* a barber that such and such...," perhaps in this case it becomes analytic, since we would be stating the mere possibility of there being such a barber. Or maybe, just maybe, the concept of *paradox* is indeed quite a difficult one to grasp, and we must analyze arguments case by case to judge whether they are paradoxes or not. It may be that no definition can put safe and firm boundaries into paradoxes, keeping inside

the borders all arguments which are actually paradoxes and throwing away all which are not. In the next section, I will analyze some arguments case by case to determine whether they are, according to the definitions exposed here (and perhaps others), truly paradoxes or only mirages.

1.4 Investigating Arguments

To investigate arguments and determine their paradoxicality (or lack of it), I will take into account, in addition to the definitions already exposed, Quine's threefold distinction between *falsidical* paradoxes, *veridical* paradoxes and *antinomies* (in [43]). Falsidical paradoxes are arguments that lead to absurd conclusions but, on a closer examination, are discovered to be fallacious. Quine gives as an example the misproof that $2 = 1$ ([43], p. 5): Let x = 1. Then $x^2 = x$. Thus, $x^2 - 1 = x - 1$. Dividing both sides by $x - 1$, we have that $x + 1 = 1$, and so $2 = 1$. As credible as the "proof" may be, it is easily overthrown once we consider that, in dividing both sides of the equation by $x - 1$, we would be dividing them by 0, which cannot be done. Hence, this is a falsidical paradox: it is an argument that reaches a bizarre conclusion, but it is actually fallacious. Quine adverts that those should not be mistaken to be simple fallacies: fallacies can have true or false conclusions, and their conclusions might be surprising or seem trivial. Falsidical paradoxes, on the other hand, always have absurd and false conclusions, which arise from a fallacious step in the argument. Quine places paradoxes such as the barber's on the side of *veridical* paradoxes. Those are sound arguments that have strange, but true, conclusions. On the barber's case, for example, the contradictory conclusion implies that there can be *no such barber*. These are truly paradoxical, but are not so insidious; they do not defy the grounds of a well established belief or theory.

To the most profound type of paradox Quine reserves the name *antinomy*. These are sound arguments that bring on the crises in thought. They challenge deep principles of reasoning and force us to review a whole field of knowledge. The difference between them and veridical paradoxes is that their reach goes much further than the latter's; while veridical paradoxes can be seen as a demonstration of an impossibility (for example, the demonstration of the impossibility of there existing such a barber), antinomies cannot be brushed off that easily. A useful parallel arises: antinomies are deeper paradoxes than veridical ones, such as Priest's definition in [40] (p. 9) requires "stronger" paradoxes than Rescher's, in [44] p. 6. For the latter, it suffices that premises be plausible for an argument

to be a paradox. For the former, they must start with analytic principles. Of course, as I said in the last section, those two notions are, to some degree, vague; but requiring premises to be analytic is undoubtedly a stronger demand. Thus, Rescher's definition is closer to Quine's notion of veridical paradox, whereas Priest's definition is closer to Quine's notion of antinomy.

An example of an antinomy would be Grelling-Nelson's Paradox, or the Heterological Paradox. I will introduce it formally in the next chapter. An informal version of it would be thus: I call *autological* an adjective that describes itself, and *heterological* an adjective that does not. "Short," for example, is autological, since the word "short" is short. "Long," on the other hand, is heterological, for the adjective is not long. The paradox arises when we ask: is the adjective "heterological" an autological or heterological one? If it is autological, then it describes itself. But it defines exactly the adjectives that do not describe themselves, so it is heterological. On the other hand, if it is heterological, then it does not describe itself, but that is in precise agreement with the definition of heterological adjectives; so, it is autological. One cannot simply brush off this paradox as one could do with the barber, by just negating the premise that there can be a barber who shaves all and only men who do not shave themselves. Grelling-Nelson's paradox does not come from any interim premise such as the existence of some entity. The mere definition of an adjective ended up being paradoxical, and it seems unreasonable to conclude that "there is no such thing as the "heterological" adjective," for I just defined it using perfectly sound language rules.[9]

I said that antinomies are paradoxes that defy deep entrenched principles of reasoning. The principle, in Grelling-Nelson's case, is that adjectives are true of things if and only if they apply to things – the adjective "loud" is true of something if and only if this thing is loud; the adjective "black" is true of something if and only if it is black; the adjective 'not true of itself' is true of something if and only if it is not true of itself. This is such a rooted principle that it would be quite strange to abandon or to restrict it: how could language function without the feature that, when I say "red glass," I am denoting something that has the properties of being a glass and being red? This is what antinomies do: they force us to either abandon or restrict deep principles in a rather artificial

[9]Furthermore, Quine remarks that we do not even need the adjectives "autological" and "heterological" to bring about the paradox. We could simply substitute that by the adjectival phrases "true of itself" and "not true of itself" and ask whether "not true of itself" is true of itself or not. Again, the paradox arises. ([43], p. 6).

fashion or, on the other hand, accept the paradoxical extension of some principle by adopting an inconsistent (or incomplete) semantics of some sort.

The Liar Paradox is, by all measures, one of the most insidious paradoxical arguments. It is a paradox according to Priest's definition (and certainly to Rescher's as well) and an antinomy according to Quine's. I have exposed it briefly in the previous section (and will go over it extensively in the next chapter), but to refresh the reader's memory, here it goes again: the paradox comes from a sentence which says that it, itself, is false – for example, in the statement "this sentence is false." It arises with the question: is the sentence true or false? If we assume it is true, we conclude that it is false; if we assume it is false, we conclude that it is true. This attacks the principle that sentences are either true or false, and cannot be both true and false or neither true nor false. Moreover, this disturbs the principle that Tarski named *T-schema*, which will be explained in detail in the next chapter, and states that a sentence "φ" is true if and only if φ is the case. Again, to cope with the antinomy, we must either abandon or restrict this principle or adopt an inconsistent or incomplete semantics.

Russell's Paradox, in its turn, is a more delicate case. Let us now examine it. It arises if we take as a premise the unrestricted Comprehension Axiom, that states that any formula $\Phi(x)$ determines a set $\{x : \Phi(x)\}$, whose members are exactly those objects that satisfy $\Phi(x)$. If we take $\Phi(x)$ to mean $x \in x$ and define the set $R = \{x : \neg\Phi(x)\}$ as the set of all sets that are not member of themselves, then the question "is R a member of itself?" gives us a contradiction. If R is a member of itself, then it satisfies $\Phi(x)$, so it cannot be one of elements of R, for it is not among the objects such that $\{x : \neg\Phi(x)\}$, so it is not a member of itself. On the other hand, if it is not a member of itself, then it does not satisfy $\Phi(x)$, making it one of the objects such that $\{x : \neg\Phi(x)\}$, and so it is a member of itself. Therefore, we've concluded that R both is and is not a member of itself – which makes the set of premises and conclusion inconsistent.

The argument is a genuine paradox according to all definitions: Priest's, Rescher's and Quine's. Now, in Quine's case, should it be called an antinomy, or simply a veridical paradox? He chooses antinomy, albeit admitting that it belongs to a different paradoxical family than Grelling-Nelson's or the Liar. Indeed, the argument is more similar to the one in the paradox of the barber than in the other two: we define the set of all sets that are not member of themselves and, through sound reasoning, derive a contradiction. In view of this, our conclusion

(in a classical appreciation of the problem, at least)[10] should be that there can be no such thing as the set of all sets that are not member of themselves (or simply *the set of all sets*, by the same reasoning). Well, but Quine dropped the barber's paradox in the veridical paradoxes batch because its conclusion was that a certain barber cannot exist, and in Russell Paradox, the conclusion must be that a certain set cannot exist (classicaly). Thus, shouldn't Russell's Paradox be seen simply as a veridical paradox?

The reasons that lead Quine to establish that the paradox is an antinomy are tied to its consequences to mathematics: as I said in the last section, it stroke the Axiom of Comprehension of Frege's set theory and, because of this, changed the whole of mathematical practice. Furthermore, Quine argues that, in mathematics, there has been an overwhelming presumption that there is a set whose members are all and only the sets which are not members of themselves; whereas there hasn't been such a strong presumption in the case of the barber. The following quote can make the point clearer:

> A veridical paradox packs a surprise, but the surprise quickly dissipates itself as we ponder the proof. A falsidical paradox packs a surprise, but it is seen as a false alarm when we solve the underlying fallacy. An antinomy, however, can be accommodated by nothing less than a repudiation of part of our conceptual heritage ([43], p. 11).

The reasons that lead Russell's Paradox to be characterized as an antinomy are completely contingent; so much so that Quine even adds that perhaps in the future, when it becomes blatantly absurd that there can be a set whose members are sets that are not members of themselves, and when there is another strong enough principle to put in its place, the argument may cease to be an antinomy, to become simply a veridical paradox. The logician underlines this contingency with a strange statement: "One man's antinomy can be another man's veridical paradox, and one man's veridical paradox can be another man's platitude" ([43], p. 14). This might be more of a catchphrase than a considerate argument, but it still doesn't seem to suit well to the text's atmosphere: if one man's antinomy is another man's veridical paradox (and so on), how could the very author himself argue that the Liar, Grelling-Nelson's and Russell's Paradox as antinomies, and the barber as a veridical paradox? If qualifying paradoxes is a subjective or social

[10] This is a waive to non-classical solutions to Russell's Paradox: there is a whole new field of studies called Inconsistent or Paraconsistent Set Theory. In it, the Russell set exists and is accepted, and the theory is inconsistent but non-trivial. On that, see [40] chapter 18.

activity, none of his arguments should mean much. The very position embraced by Quine that contingent factors are allowed to enter our criterion of qualification of paradoxes can be criticized. In the case of Russell's Paradox, for example,

The Liar, Russell's, Grelling-Nelson's and the barber paradoxes are all cases where the definition of paradox applies. It is much harder to decide upon problematic cases such as Zeno's paradoxes. Firstly, Zeno himself did not call his arguments *paradoxes*,[11] even though this is the term with which they are referred to today: being part of the Eleatic philosophical school, Zeno sought to confirm Parmenides' doctrine using clever logical constructions. In some sense, therefore, it can be said that Zeno intended to extract *positive conclusions* from his arguments, even if those are rather alien to common sense, such as that "there can be no movement" or "there can be no plurality". I say that his conclusions are *positive* because the arguments are always aimed at proving *something*, rather than stating a logical impossibility, as in the Liar case. Examining his *Arrow* argument might make the matter clearer:

> The third is ... that the flying arrow is at rest, which result follows from the assumption that time is composed of moments .... he says that if everything when it occupies an equal space is at rest, and if that which is in locomotion is always in a now, the flying arrow is therefore motionless. (Aristotle Physics, 239b30) Zeno abolishes motion, saying "What is in motion moves neither in the place it is nor in one in which it is not". ([15], ix. 72).

To make this argument, Zeno appealed to the premise that time is composed *only* by moments. As a moment is something without duration, an arrow is motionless during each one of the moments that compose time – "everything when it occupies an equal space is at rest." Thus, if everything in motion "is always in a now" – that is, if time is composed by moments and everything that moves is always in one –, then the arrow must be at rest, not in motion. With this conclusion, Zeno aims to reinforce the idea that there can be no movement in the world. Two questions emerge: Is this a paradox? Is his argument sound?

Let us examine the first question. Rescher's definition of paradox, for example, states that it arises when plausible propositions are collectively inconsistent. Is that what is happening here? In a way, yes. From the premises which state that time is composed of moments (P_1), everything when it occupies an equal space is at rest (P_2) and that which is in locomotion is always in a now (P_3), he concludes

[11] This is supported by [44] p. 3.

that the arrow is not in motion (C). And so the set $P_1, P_2, P_3, \neg C$ is inconsistent, even though $\neg C$ is plausible. However, according to our definition of paradox, all propositions included in the reasoning must be *plausible*. Is P_1 plausible? That is, is it a common belief that time is composed of moments and nothing more? Again, the ambiguity resides in the meaning of the word *plausible*. How can we evaluate the plausibility of an assertion? Rescher's definition again shows itself to be vague to evaluate such complicated cases. But the problem is not easily solved with Priest's definition either: deciding whether the premises are analytic or not is as complicated as deciding if they are plausible.

One important objection that defies the argument's soundness is: to determine the velocity of the arrow, one must divide the distance traveled in some time by the length of that time; but if the arrow travels 0m in 0s, the fraction $\frac{0}{0}$ m/s is no number whatsoever, so the formula would not make sense.[12] Thus, Zeno's argument would not be sound, because we should be able to calculate that the arrow's velocity equals zero. Maybe, however, Zeno could argue that the velocity cannot be calculated due to the very reason that it is a failed concept – if there is no movement, the concept of velocity indeed does not make sense. But this would be a circular move, for he would be using the fact that movement does not exist to argue that the velocity formula does not work. If the argument is indeed invalid, then it cannot be a paradox. But even if it were valid, the problem of deciding if it was a paradox or not would still be present.

[12] This argument can be found here [22].

Chapter 2

The Liar and Other Paradoxes

2.1 What is necessary to create a Liar

The first version of the Liar Paradox is due to Eubulides of Miletus, a renowned philosopher of the Megarian school, known for his paradoxes. One of his creations is the Liar Riddle (*pseudomenos*), which poses the question: does the man who says 'I am lying' lie? (Also "Does the witness who declares 'I am perjuring myself' perjure himself?") ([44], p. 199). The problem here, as I explained briefly in the previous chapter, is that we cannot evaluate the sentences consistently. From such riddles, we can conclude that the man is both lying and not lying, and that the witness is both committing and not committing perjury. Eubulides' Liar riddle was well known in antiquity ([44], ibid.) and gave rise to the famous statement "all Cretans are liars," uttered by Epimenides of Crete. The point of Epimenedes' proclamation is that he is a Cretan himself, implying that if his statement is true, he must be a liar, thereby rendering the statement "all Cretans are liars" false. Of course, the dilemma only arises if we interpret "liar" as someone who is inherently deceitful and *always* tells lies. Differently from Eubulides' riddles or the modern versions of the Liar, Epimenedes' sentence is a self-falsifying statement, rather than a paradox: concluding that it is false does not lead to further contradiction, for it is perfectly consistent to say that some Cretans are liars and others are not. Curiously, his statement was quoted in nothing less than the New Testament, where it reads:

> For there are many insubordinate men, empty talkers and deceivers, especially the circumcision party; they must be silenced [...].
> One of themselves, a prophet of their own, said, "Cretans are always liars, evil beasts, lazy gluttons." This testimony is true. (Titus 10-13).

Considering the moralizing tone of the passage, it is rather comical that Paulus has completely misunderstood the essence of Epimenides' sentence. Had he grasped that the seemingly innocent statement is a riddle that forbids him from consistently qualifying the testimony as true, he might have taken it as further evidence of Cretans' insubordination.

After its first formulation, many more effective versions of the Liar Paradox have been written (in which the paradox does not depend on the meaning of the word "liar," for example). One of those is the following:

(FL) FL is false.

FL (Falsity Liar) is the *name* of the sentence which states that it itself is false. The paradoxical character of the sentence emerges with the question: is this sentence true or false? If it's true, then *what it says* is true. As it says of itself that it is false, it must be false. On the other hand, if it is false, then *what it says* is false; but if says of itself that it is false, so if that's false, it must be true.

Embedded in this informal demonstration of the paradox are some principles widely adopted by classical logicians, which need to be spelled out so that the paradox is derived formally. Two of them are the Principle of Non-Contradiction (PNC), which states that $\neg(A \wedge \neg A)$, and the Law of Excluded Middle (LEM), which states that $A \vee \neg A$.[1] There are still others to be introduced: the *Ex Contradictione Quodlibet* principle (ECQ), which states that $A \wedge \neg A \vdash B$ for any B; the Principles of Conjunction (PC) and Disjunction (PD), of which the first says that if $A \vdash B$ and $A \vdash C$, then $A \vdash B \wedge C$ and the second guarantees that, if $A \vdash C$ and $B \vdash C$, then $A \vee B \vdash C$. Lastly, the Closure Principle, which states that if A is the case and $A \vdash B$, then B is the case. The demonstration of the paradox starts from the union of these principles to the principles of capture and release, whose definitions are below.[2]

Capture: A logically implies $Tr(\ulcorner A \urcorner)$, symbolized as $A \vdash Tr(\ulcorner A \urcorner)$.
Release: $Tr(\ulcorner A \urcorner)$ logically implies A, symbolized as $Tr(\ulcorner A \urcorner) \vdash A$.

Such rules[3] were elaborated by Tarski (1936) to form the T-schema, which I will define below. The intuitive idea is this: if we claim that a sentence is the

[1] I am using $\wedge$ and $\vee$ as the usual connectives for conjunction (and) and disjunction (or), respectively.

[2] [5], p. 18.

[3] Tr is the truth predicate. So $Tr(\ulcorner A \urcorner)$ means that the truth predicate is being applied to the name of the sentence A.

case, then we are allowed to claim that it is true – that is, capture it with the truth predicate. If we claim that a sentence is true, we can assert it, detach it from the truth predicate. It is not difficult to see why such rules are necessary to derive the Liar Paradox: assuming the sentence is true, the paradox is only consummated if we can affirm what it says (release); on the other hand, the paradox is only consummated if, from its affirmation, we can affirm that it is true (capture). Together, the rules result in the famous T-schema:

(T-Schema) $Tr(\ulcorner A \urcorner) \leftrightarrow A$

The T-schema is nothing more than the unification of the Capture and Release rules in a biconditional: it expresses that if a sentence A is true, then what it states must be the case. Likewise, if what a sentence A states is the case, it follows that it is true, which allows us to write $Tr(\ulcorner A \urcorner)$. This simple scheme captures the crucial distinction between *mention* and *use*, two different employments of a sentence. When I *mention* a sentence, by writing it under corner quotes (or quotation marks, as in non-formal speech), I am referring not to the what the sentence effectively says, but to the sentence itself, to the set of words that compose it, as if they were a single object. When I *use* a sentence, I am affirming what the sentence *says*, its content. On the right side of the T-schema, the sentence A is used; on the left side, it is mentioned.

Now that we have defined the principles and rules, we can finally demonstrate the paradox. There is, however, another even more effective version of it, and I will reproduce its proof instead of the proof for the FL sentence. Here is how this other version goes:

(UL) UL is not true.

The UL (Untrue Liar) paradox is derived very similarly to the FL paradox. However, the question that raises the antinomy is the following: is this sentence true *or not true*? The difference between FL and UL is that UL does not require the concept of falsity to generate the paradox, but only of "non-truth". In the proof, λ is the Liar Sentence. So we set λ to be $\neg Tr(\ulcorner \lambda \urcorner)$.

Proof 1.

1. Show that $Tr(\ulcorner \lambda \urcorner) \vdash Tr(\ulcorner \lambda \urcorner) \wedge \neg Tr(\ulcorner \lambda \urcorner)$:

(a) $Tr(\ulcorner \lambda \urcorner)$ [premise]

(b) λ [1a: release]

(c) $\neg Tr(\ulcorner\lambda\urcorner)$ [1b: definition of λ]
(d) $Tr(\ulcorner\lambda\urcorner) \wedge \neg Tr(\ulcorner\lambda\urcorner)$ [1a, 1d: PC]

2. Show that $\neg Tr(\ulcorner\lambda\urcorner) \vdash Tr(\ulcorner\lambda\urcorner) \wedge \neg Tr(\ulcorner\lambda\urcorner)$:

(a) $\neg Tr(\ulcorner\lambda\urcorner)$ [premise]
(b) λ [2a: definition of $\neg(\ulcorner\lambda\urcorner)$]
(c) $Tr(\ulcorner\lambda\urcorner)$ [2b: capture]
(d)$Tr(\ulcorner\lambda\urcorner) \wedge \neg Tr(\ulcorner\lambda\urcorner)$ [2a, 2c: PC]

3. $Tr(\ulcorner\lambda\urcorner) \vee \neg Tr(\ulcorner\lambda\urcorner) \vdash Tr(\ulcorner\lambda\urcorner) \wedge \neg Tr(\ulcorner\lambda\urcorner)$ [1, 2 PD]

4. $Tr(\ulcorner\lambda\urcorner) \wedge \neg Tr(\ulcorner\lambda\urcorner)$ [3: LEM, Closure]

■

This version is more insidious than FL, as it generates problems for those who intend to respond to the Liar by saying that the sentence is "neither true nor false," i.e., that it is in the space between truth and falsity. Given this version of Liar, that is not enough: one cannot solve it just by creating a third truth value, non-truth, and saying that the paradox is not truth, for UL poses a contradiction between truth and non-truth.[4] Paracomplete solutions to the Liar must find a way to overcome this difficulty.

For a more detailed view on paracomplete solutions to the paradox and many other formal theories of truth, see the appendix "possible solutions to the Liar Paradox" at the end of this thesis.

2.2 Other versions of the Liar

In this section, I will expose different versions of the Liar paradox and the Herelogical Paradox. After this, I will end the chapter with a discussion about Yablo's Paradox.

An interesting version of the paradox was introduced by the logician Philip Jourdain, called the Card Paradox. The paradox is achieved with a card that has written on one side "the sentence on the other side of this card is true" and, on the other, "the sentence on the other side of this card is false." Later, another version of the same paradox was introduced by Kripke.[5] It occurs not in a single

[4] See [5], p. 11.
[5] [25], p. 691.

sentence, but in a cycle, formed by the speeches of two different people. See below:[6]

(A) The following statement is not true.
(B) The preceding statement is true.

At first glance, the claims do not seem to conflict at all. However, I will demonstrate below that A can be regarded as both true and not true. It should be noted that, in the proof, I use two rules not yet stated. One is the Double Negation principle (DN), which states $\vdash A \Leftrightarrow \neg(\neg A)$. While this principle is valid in classical logic, this is not the case for other logics – notoriously, it is not valid in the intuitionistic logic (consequently, it might be more difficult to derive the paradox in intuitionistic logic). The other rule assumed here is the contraposition of the T-scheme: $\neg Tr(\ulcorner A \urcorner) \Leftrightarrow \neg A$. I use it in steps 1d and 2b, to apply the release rule inside the parentheses.

(A) $\neg Tr(\ulcorner B \urcorner)$
(B) $Tr(\ulcorner A \urcorner)$

Proof 2.

1. Prove that $Tr(\ulcorner A \urcorner) \vdash \neg Tr(\ulcorner A \urcorner) \wedge Tr(\ulcorner A \urcorner)$

(a) $Tr(\ulcorner A \urcorner)$ [premise]
(b) A [1a: release]
(c) $\neg Tr(\ulcorner B \urcorner)$ [1b: definition of A]
(d) $\neg B$ [1c: release (T-schema contraposed)]
(e) $\neg Tr(\ulcorner A \urcorner)$ [definition of B]
(f) $Tr(\ulcorner A \urcorner) \wedge \neg Tr(\ulcorner A \urcorner)$ [1a, 1e: PC]

2. Prove that $\neg Tr(\ulcorner A \urcorner) \vdash Tr(\ulcorner A \urcorner) \wedge \neg Tr(\ulcorner A \urcorner)$

[6] The version I present here is a simplified version of the one contained in the paper by Kripke (1975). There, the sentences that generate the paradox are as follows

Jones: Most of Nixon's claims about the *Watergate* are false.
Nixon: All of Jones' statements about the *Watergate* are true.

In this version, the paradox arises only in the case where half of Nixon's statements is true and half is false, except for one problematic case - that of the statement "all of Jones' statements about the *Watergate* are true". We must further assume that Jones' only claim about the *Watergate* is the one described above, "most of Nixon's claims about *Watergate* are false". Only in this configuration is it possible to derive the paradox in a similar way as we derive it here. Kripke selects this complex situation to highlight the contingent components that make it paradoxical. With this, the author emphasizes the importance of meditating on the Liar in everyday language, as I discuss below.

(a) $\neg Tr(\ulcorner A \urcorner)$ [Premise]
(b) $\neg A$ [2a: release (T-schema contraposed)]
(c) $Tr(\ulcorner B \urcorner)$ [2b: definition of A; DN]
(d) B [2c: release]
(e) $Tr(\ulcorner A \urcorner)$ [2d: definition of B]
(f) $\neg Tr(\ulcorner A \urcorner) \wedge Tr(\ulcorner A \urcorner)$ [2a, 2e: PC]

3. $Tr(\ulcorner A \urcorner) \vee \neg Tr(\ulcorner A \urcorner) \vdash Tr(\ulcorner A \urcorner) \wedge \neg Tr(\ulcorner A \urcorner)$ [1, 2: PD]

4. $Tr(\ulcorner A \urcorner) \wedge \neg Tr(\ulcorner A \urcorner)$ [3: LEM, Closure]

■

With this version of the Liar, Kripke aimed to point out the importance of studying the paradox in everyday (non-formal) language, because, as the author says, "many, probably most, of our ordinary assertions about truth and falsity are liable, if the empirical facts are extremely unfavorable, to exhibit paradoxical features"([25], p. 691). Kripke's version shows that the paradox can arise in an absolutely common context, depending on certain empirical factors that he calls "unfavorable". Therefore, it is not enough to exhibit solutions to the paradox that forbid certain statements using the truth predicate – our solutions must allow *risky* statements to be uttered using such predicates, without sifting the "bad" cases from the "good" ones, rejecting the bad ones and preserving the rest. For a more detailed discussion of this topic, see the appendix at the end of this text.

Another alternative to enunciate the paradox is what Beall, Glanzberg and Ripley call Boolean compounds (in [5] p. 11). A possible example is the following:

(DL) DL is not true or rats play beach volleyball.

The Disjunctive Liar (DL) is a little different from the previous ones: instead of resulting in contradictions, it allows us to prove, for example, that rats play beach volleyball, as we will see in the demonstration below. Let A be the name of the sentence "$\neg Tr(\ulcorner A \urcorner) \vee B$" and B be the sentence that states "rats play beach volleyball."

(A) $\neg Tr(\ulcorner A \urcorner) \vee B$

Proof 3.

1. Suppose $Tr(\ulcorner A \urcorner)$. Demonstrate $\vdash B$.
(a) $Tr(\ulcorner A \urcorner)$ [Premise]

(b) A [1a: release]
(c) $\neg Tr(\ulcorner A \urcorner) \vee B$ [1b: definition of A]
(d) $\neg(\neg Tr(\ulcorner A \urcorner))$ [1a: PNC]
(e) B [1c, 1d: disjunction]

2. Suppose $\neg Tr(\ulcorner A \urcorner)$. Prove $\vdash B$.

(a) $\neg Tr(\ulcorner A \urcorner)$ [Premise]
(b) $Tr(\ulcorner A \urcorner) \wedge \neg B$ [1a: negation; definition of A]
(c) $Tr(\ulcorner A \urcorner)$ [2b: conjunction]
(d) A [2c: release]
(e) $\neg Tr(\ulcorner A \urcorner) \vee B$ [2d: definition of A]
(f) $\neg(\neg Tr(\ulcorner A \urcorner))$ [2c: PNC]
(g) B [2e, 2f: disjunction]

■

The problem raised by this version of the paradox is that it allows us to prove any declarative sentence, however absurd it is. Such a version is also interesting as it is equivalent to a Curry sentence. Curry sentences are material implications such that the first conditional states that the sentence itself (the entire implication) is true, and the second expresses any sentence. In this case, DL would look like:

(DL') If DL' is true, then rats play beach volleyball.

Note that DL and DL' are classically equivalent (that is, if the classical implication and disjunction rules are followed): DL is formalized as $\neg A \vee B$, and DL' as $A \rightarrow B$.

2.3 The Heterological Paradox

The Heterological Paradox (or Grelling-Nelson's paradox) is an important semantic antinomy that has its roots in impredicativity. A definition is said to be *impredicative* when it depends on a set of objects, at least one of which is the very object being defined. To produce the paradox, we first define two categories of words:

1. *Autological* words are those that describe themselves, and
2. *Heterological* words are those that do not describe themselves.

For example, the word "polysyllabic" is autological, because it is a polysillabic word; "English" is also autological, for it is a word of the English language. On the other hand, words such as "monosyllabic" or "long" are heterological, since the first is not a monosyllabic word and the second is not long. The paradox arises when we ask: is the word "heterological" autological or heterological? We cannot, consistently, place the word in any of the two boxes. The proof that this question leads to paradox is very different from the variations of the Liar. I will show it below:

Proof 5.

For any predicate p, let $A(\ulcorner p \urcorner)$ be mean that "p" is autological, and $H(\ulcorner p \urcorner)$ mean that "p" is heterological. Additionally, for any predicate p, let $D(\ulcorner p \urcorner)$ mean that "p" describes itself. According to the definitions above, we have the following equivalencies:

Eq. 1: $D(\ulcorner p \urcorner) \leftrightarrow p(\ulcorner p \urcorner)$
Eq. 2: $A(\ulcorner p \urcorner) \leftrightarrow D(\ulcorner p \urcorner)$
Eq. 3: $H(\ulcorner p \urcorner) \leftrightarrow \neg D(\ulcorner p \urcorner)$

1. Prove that $A(\ulcorner H \urcorner) \vdash \bot$

(a) $A(\ulcorner H \urcorner)$ [premise]
(b) $D(\ulcorner H \urcorner)$ [Eq. 2]
(c) $\neg H(\ulcorner H \urcorner)$ [Eq. 3, contraposition]
(d) $\neg D(\ulcorner H \urcorner)$ [Eq. 1, contraposition]
(e) $D(\ulcorner H \urcorner) \wedge \neg D(\ulcorner H \urcorner)$ [b, d]
(f) $\bot$

2. Prove that $\neg A(\ulcorner H \urcorner) \vdash \bot$

(a) $\neg A(\ulcorner H \urcorner)$ [Premise]
(b) $\neg D(\ulcorner H \urcorner)$ [Eq. 2, contraposition]
(c) $H(\ulcorner H \urcorner)$ [Eq. 3]
(d) $D(\ulcorner H \urcorner)$ [Eq. 1]
(e) $\neg D(\ulcorner H \urcorner) \wedge D(\ulcorner H \urcorner)$ [b, d]
(f) $\bot$

3. $A(\ulcorner H \urcorner) \vee \neg A(\ulcorner H \urcorner) \vdash \bot$ [1, 2, PD]

■

By comparing the formal proof of the paradox with its informal exposition, the reader will notice that we defined the predicates $A(\ulcorner p \urcorner)$ ("p" is autological)

and $H(\ulcorner p \urcorner)$ ("p" is heterological) as ranging over *predicates*, not words (as in the informal version). This is necessary because, to generate the contradictions, the predicate "heterological" must range over itself. If we defined the predicate as ranging over words, we would have a *type* problem: the the predicate H and the word "heterological" would belong to two different types, so $H(\ulcorner heterological \urcorner)$ would not be describing exactly itself, but another type of entity. Thus, we could not go from $H(\ulcorner heterological \urcorner)$ to $D(\ulcorner H \urcorner)$, which is a necessary step of the proof.

It is fascinating how each logical paradox, so simple yet so disruptive, seems to *poison* one of more parts of a classical formal theory: the Liar Paradox contaminates formal theories of truth; Russell's Paradox contaminates set theory (as we will see further ahead in chapter 3); the Heterological Paradox, in its turn, infects logics of predicates that range over predicates. Anyone who wishes to build a second-order logic with predicates ranging over predicates must be aware of the paradox, and must either embrace non-classicality or take steps to avoid it. If the first option is preferred, then one may choose from a wide range of heterodox logics that can "solve" the paradox, be it a paraconsistent, paracomplete, or substructural logic (or yet another). If the second option is chosen, one way to deal with the paradox is by introducing a type restriction, according to which predicates can range only over predicates of a lower type; this will stop a predicate from ranging over itself. Another way to deal with the problem while remaining classical is through a Tarskian sort of hierarchy of languages. In this alternative, predicates can range over predicates of languages of lower levels, but not over predicates of its own language.

2.4 Yablo's Paradox

Yablo's Paradox was elaborated by the logician Stephen Yablo in [57], and was originally called the ω-Liar. The author aimed to create a non-circular version of the Liar Paradox to show that this is not a necessary item to formulate the antinomy – and, ultimately, to formulate semantic paradoxes in general. This is, however, a controversial point: many have argued that there is a hidden circularity in Yablo's construction. For this proof, we use two additional principles that have not yet been introduced: $\forall$-introduction, which states that, if $P(x)$ is the case, where x is an arbitrary element from a set, then $\forall x P(x)$ is the case; and $\forall$-elimination, which states that, if $\forall x P(x)$ is the case, then $P(a)$ is the case, with a being a specific element from the set. Now, let's see the paradox' formulation

and proof:

Yablo's Paradox

$Y(0)$: For all $n > 0$, S_n is not true.
$Y(1)$: For all $n > 1$, S_n is not true.
$\vdots$
$Y(n)$: For all $n > m$, S_n is not true.
$\vdots$

Proof 5.

(1) $Tr(\ulcorner Y(n) \urcorner)$ [Premise]
(2) $Y(n)$ [1: release]
(3) $(\forall m)(m > n \rightarrow \neg Tr(\ulcorner Y(m) \urcorner))$ [2: def. of $Y(n)$]
(4) $\neg Tr(\ulcorner Y(n+1) \urcorner)$ [3: arithmetic]
(5) $(\forall m)(m > n+1 \rightarrow \neg Tr(\ulcorner Y(m) \urcorner))$ [3: df. of $Y(n)$, arithmetic]
(6) $Y(n+1)$ [5: df. of $Y(n+1)$]
(7) $Tr(\ulcorner Y(n+1) \urcorner)$ [6: capture]
(8) $\bot$
(9) $\neg Tr(\ulcorner Y(n) \urcorner)$ [1-8: *reductio*]
(10) $(\forall n)(\neg Tr(\ulcorner Y(n) \urcorner)$ [$\forall$-intro.]
(11) $(\forall m)(m > 0 \rightarrow \neg Tr(\ulcorner Y(m) \urcorner))$ [10: arithmetic]
(12) $Y(0)$ [11: df. of $Y(0)$]
(13) $Tr(\ulcorner Y(0) \urcorner)$ [12: capture]
(14) $\neg(Tr(\ulcorner Y(0) \urcorner))$ [10: $\forall$-elim.]
(15) $\bot$

■

The proof shows that the Liar's *explicit* circularity is not present here: none of the sentences of the infinite sequence $(Y(0), Y(1), ..., Y(n), ...)$ says of itself that it is not true, as most common versions of the paradox do. The circularity seems to dissolve itself into the infinite character of the sequence, as each S_n always finds sentences above itself to denote, so there is no failure of denotation nor does any sentence denotes itself. However, such considerations are not enough to

demonstrate that the paradox really escapes circularity. For this, we must first answer the question: what is circularity?

In this section, I will explore the idea of taking circularity to be in the existence of fixed points, which can be said to be the mainstream view of the concept in logic. I will show how, by this train of thought, Yablo's Paradox is actually circular. This is not, however, the view that I will personally argue for: in chapter three, I claim that defining circularity through fixed points leads to triviality, so we should look for other definition of the concept. I then propose a definition of my own, by which Yablo's Paradox ends up being non-circular. But this will appear solely in chapter three. Now, I will merely portray the discussion that has led to the conclusion, prevalent in today's literature on the topic, that the paradox is circular; further ahead, it will be clear how and why I do not subscribe to it.

To talk about the mainstream concept of circularity, we should first introduce the mathematical notion of *fixed point*, for it is believed that this two notions are tied together (a claim that I will challenge in the next chapter). However, Roy T. Cook [11] shows that this concept too suffers from some vagueness: there are at least two possible definitions of fixed point, the strong and the weak one, that can be found below ([11], p. 74).

Definition.[7] A sentence Φ is a *strong sentential fixed-point* of an unary predicate $\Psi(x)$ if and only if:

$$\ulcorner\Phi\urcorner = \ulcorner\Psi(\ulcorner\Phi\urcorner)\urcorner$$

Definition. A sentence Φ is a *weak sentential fixed-point* of a unary predicate $\Psi(x)$ if and only if:

$$\Phi \Leftrightarrow \Psi(\ulcorner\Phi\urcorner)$$

is a theorem.[8]

[7] This definition is a little more precise then the one contained in [11]. There, it is said that $\Phi = \Psi(\Phi)$ for a sentence Φ and predicate Ψ. The problem with this is that Φ cannot be *equal* to $\Psi(\Phi)$, because they are not identical. What makes sense is to say that their *names* are the same; *i.e.* that they are designated by the same Gödel number.

[8] This definition is also somewhat vague in that it does not say *what* it is a theorem of. Is $\Phi \Leftrightarrow \Psi(\ulcorner\Phi\urcorner)$ a theorem of Peano Arithmetic? Is it any theory? Should we have said that a sentence Φ is a *weak sentential fixed-point* of a unary predicate $\Psi(x)$ in the theory $\mathfrak{T}$ if and only if $\Phi \Leftrightarrow \Psi(\ulcorner\Phi\urcorner)$ is a theorem of $\mathfrak{T}$? Let us assume so and believe that the definition works relative to some theory.

The difference between the two definitions above is subtle: while the strong sentential fixed point requires that the (name of the) sentence be *identical* to the (name of the) application of the unary predicate to the sentence, the weak sentential fixed point requires that they be equivalent. Note that, if the classical rules of identity are maintained, a strong fixed point is always also a weak fixed point. The first type captures well the fixed points contained in the most common versions of Liar, such as FL or DL. The second, in turn, captures what Cook calls Arithmetic Liar. This is one way to build the Liar paradox into arithmetic. Assuming we are working on a first-order (or stronger) theory of arithmetic that contains the truth predicate and in which T-Schema holds, we obtain the Liar from the following lemma.

Gödelian diagonalization lemma ([11], p. 22):

For any unary predicate $\Psi(x)$, there exists a sentence Φ such that

$$\Phi \Leftrightarrow \Psi(\ulcorner \Phi \urcorner)$$

is a theorem of arithmetic.

Returning to the definitions quoted above, the lemma states that every unary predicate has a weak sentential fixed point. Applying the lemma to the negation of the truth predicate, we get the Arithmetic Liar, which states that

$$\Lambda \Leftrightarrow \neg Tr(\ulcorner \Lambda \urcorner)$$

is a theorem of arithmetic.

The contradiction is, in this case, obtained by the incompatibility of the result with the instance of the sentence Λ from T-Scheme: $Tr(\ulcorner \Lambda \urcorner) \Leftrightarrow \Lambda$. The Arithmetic Liar is therefore a weak sentential fixed point.

It is misleading to call this version the "Arithmetic Liar," however, for it suggests that a Liar with a strong fixed point cannot be constructed within arithmetic. That is not true, though: whether you have a strong or a weak sentential fixed point depends on the expressive resources available in the theory. In particular, if you have an arithmetic with a diagonal function, then you can construct strong fixed points. Hence, although it is true that you can define fixed points in those two different ways, nothing of much significance hangs on this distinction. We will see below that the Yablo Paradox can contain a strong fixed point, not only a weak one.

In any case, these distinctions do not have much relevance to the specific discussion of the Yablo Paradox, as it does not in fact contain a strong or weak sentential fixed point, since the construction does not include any *sentence* that refers to itself. As will become clear below, what the paradox includes is a predicate fixed point ([11], p. 75).

Definition[9]**.** A unary predicate $\Phi(x)$ is a *strong predicate fixed point* of a binary predicate $\Psi(x, y)$ if and only if:

$$\ulcorner\Phi(x)\urcorner = \ulcorner\Psi(\ulcorner\Phi(w)\urcorner, x)\urcorner$$

Definition. A unary predicate $\Phi(x)$ is a *weak predicate fixed point* of a binary predicate $\Psi(x, y)$ if and only if:

$$\Phi(x) \Leftrightarrow \Psi(\ulcorner\Phi(w)\urcorner, x)$$

is a theorem.[10]

One can use this framework to interpret fixed points as the manifestation of circularity. It is, of course, a loose definition which will lead to conceptual problems, as I will show in chapter three. But for now, let's stick with this view to try to think about the following question: does the Yablo Paradox really escape circularity? Priest, in [36], argues that it does *not*. The paradox does contain an implicit "self-referential circularity," mitigated by the infinite character of the construction. The central point of Priest's objection lies in step (2) of the proof: at first glance, it seems clear that the Tarskian rule allows us to go from (1) to (2), for if we assert $Tr(\ulcorner Y(n)\urcorner)$, we can use it to assert $Y(n)$. However, the T-schema is valid for sentences, and the definition of sentence is that of a closed formula – *i.e.* with *no free variables*, and $Y(n)$ does have free variables. Priest claims that it makes no sense to say, for example, "$Tr(\ulcorner x$ is white $\urcorner) \Leftrightarrow x$ is white" ([36], p. 237). For a correct formulation of the paradox, we need a generalization of the T-schema that holds for formulas with free variables, which Priest calls *predicates*. For the purposes of this section, I will follow this nomenclature and, henceforth, call open formulas predicates. To build the Yablo sequence correctly,

[9] This is also a little more precise than Cook's definition, for the reason explained on footnote 11.

[10] This definition suffers from the same kind of vagueness as that of sentential fixed points: Cook states that $\Phi(x) \Leftrightarrow \Psi(\ulcorner\Phi(w)\urcorner, x)$ is a theorem, but does not say the theory it is a theorem of. I will assume, again, that the definition actually reads: a unary predicate $\Phi(x)$ is a *weak predicate fixed point* of a binary predicate $\Psi(x, y)$ in a theory $\mathfrak{T}$ if and only if $\Phi(x) \Leftrightarrow \Psi(\ulcorner\Phi(w)\urcorner, x)$ is a theorem of $\mathfrak{T}$.

it is necessary to make use of the concept of satisfaction, and rewrite lines (1) and (2) as follows:

(1) $Sat(n, S^*)$
(2) $\forall k > n, \neg Tr(\ulcorner Y(k) \urcorner)$

Where Sat is the binary predicate of satisfaction and S^* is defined by $\forall k > x, \neg Tr(\ulcorner Y(k) \urcorner)$. We must also repeat the procedure on all lines, substituting the truth predicate for the satisfaction predicate. But then S^* itself must be understood as the predicate $\forall k > x, \neg Sat(k, S^*)$. However, if this is so, then $S^* = \forall k > x, \neg Sat(k, S^*)$ is a fixed point and, concludes Priest, of exactly the same type as the usual Liar Paradox. Basically, S^* is the predicate "no number greater than x satisfies this predicate" ([36], 238). The circularity, according to the author, is not exactly in the argument, but in the simple fact that we define the predicate S by specifying each of its values, but the values themselves are defined with reference to S. Note that $S^* = \forall k > x, \neg Sat(k, S^*)$ is a strong fixed point, not a weak one: the name "S^*" names precisely $\forall k > x, \neg Sat(k, S^*)$.

Cook's construction of Yablo's Paradox in arithmetic is a little different than Priest's, and he concludes from it that the paradox contains only a weak fixed point, not a strong one, since the sequence is constructed by the formulas $Y(1), Y(2), Y(3), ...$ such that the following biconditionals are theorems of the theory in question ([11], p. 24):

$$
\begin{aligned}
&Y(1) \leftrightarrow \forall n(n > 1 \rightarrow \neg Sat(\ulcorner Y(x) \urcorner, n)) \\
&Y(2) \leftrightarrow \forall n(n > 2 \rightarrow \neg Sat(\ulcorner Y(x) \urcorner, n)) \\
&\qquad\qquad\vdots \\
&Y(m) \leftrightarrow \forall n(n > m \rightarrow \neg Sat(\ulcorner Y(x) \urcorner, n)) \\
&\qquad\qquad\vdots
\end{aligned}
$$

The contradiction, as in the case of the Arithmetic Liar, is obtained only by the incompatibility of the above predicates with (an instance of) the T-scheme equivalent for the satisfaction predicate, which Cook calls $Y(x)$-*Simple Satisfaction Scheme ($Y(x)$-SSS)* ([11], p. 25). For any τ term, we have:

$$Sat(\ulcorner Y(x) \urcorner, \tau) \leftrightarrow Y(\tau)$$

Cook uses this construction to argue that, while the Yablo Paradox is, indeed, circular, its circularity is "weak," for it is achieved only with a weak fixed point, not a strong one. This is not a very good point, though: we have seen that,

through Priest's formulation of the paradox, the paradox amounts to a strong fixed point, not a weak one. Plus, getting a weak or strong fixed point depends only on the expressive power of the theory in question, not on the paradox itself. Hence, I will now leave this point behind.

Priest gives yet another example of an infinite version of Liar that, according to him, masks its circular character by appealing to infinity (even though he does not explain exactly the way in which the circular character gets masked by the infinite construction). According to him, this version is due to Soresen:[11]

> At the gates that lead to Heaven, there is an infinite line of people. Each of them is thinking only one thought, and it is this: the thought that each of the people behind me is thinking is not true. God, being aware of the thoughts of all the people in the queue, can deduce the contradiction and verify the paradox ([36], p. 240).

For Priest, the circularity of this paradox lies in the fact that everyone is thinking the *same* thought, t, and this thought is, precisely, "t is not true." The kind of circularity present in the version attributed to Soresen is even closer to the common construction of the Liar than to the Yablo Paradox: this version is, Priest argues, a sentence that says of itself that it is not true – it is, then, a strong sentential fixed point, while Yablo's Paradox contains only predicate fixed points.

It is not that simple, however: is it really the *same* thought being thought by everyone? The way Priest formulates the problem certainly makes it seem so, but many would disagree: it can be argued that the *content* of each thought is different, for the range of the expression "each of the people behind me" will be distinct for each person in the line. If the first person thinks "the thought that each of the people behind me is thinking is not true," the expression "each of the people behind me" applies to person 2, person 3, person 4 and so on, indefinitely. If the second person thinks that very thought, it will apply to person 3, person 4, person 5 and so on. So, while everyone seems to be thinking the same thought, the content of each thought is formed by a different set of individuals – which, according to this view, makes it a different thought.

Indeed, the range of a quantifier may drastically change the meaning of a sentence. Picture a child that has just learned the natural numbers $\mathbb{N}$, and only

[11] Unfortunately, I could not find Soresen's text that Priest identifies as the source of this paradox. It should be called "Extroverted epistemic paradox," but it might not have been published.

them. For her, all the numbers that exist are 1, 2, 3, 4, and so on, to infinity. Insightful as she is, she realizes an important property of this set of numbers. She thinks: "I could count all numbers that exist just by adding 1 each time, had I an infinite amount of time to do it! By adding 1 to 1, I have 2; by adding 1 to 2, I have 3; with this process, I can keep counting and, in an infinite amount of time, reach all the numbers there are!" Since the child is referring *only* to natural numbers – because these are the only numbers she knows –, she is thinking rightly, for this is true for this set of numbers. If someone else, say, an adult that has completed school and has learned about the existence of the set $\mathbb{R}$ of real numbers, thinks the same thought, the *content* of his thought will be different: for him, the quantifier "all" in "all numbers" does not range over the natural numbers only, but over the integers, rationals, real numbers and all others. Thus, his thought is different than the child's, and we can conclude that his reasoning is incorrect. In this situation, the range of a quantifier substantively changed the content of a thought whose propositional expression seemed to be the same. But that is exactly the situation described in Soresen's Paradox; hence, it is quite reasonable to defend that each person's thought is different. This avoids the circularity that was attributed to that paradox.

Going back to Yablo's Paradox, if we assume circularity to be in the presence of fixed points, then we can conclude that there is, in fact, circularity in Yablo's paradox, contrary to what the author had intended. This raises the following philosophical question: does the paradoxical nature of *every* semantic paradox reside in circularity (if it is understood as the presence of fixed points)? Is it the one mechanism responsible for generating these kinds of paradoxes? Cook argues that it is not: circularity is too widely distributed to be blamed for the paradoxicality of both the Yablo sequence and semantic paradoxes in general. The reason for Cook's argument is simple: "given a suitably strong background theory (eg, one containing a truth predicate, or one containing enough arithmetic to diagonalize) every sentence is a weak sentential fixed point of some unary predicate, and every unary predicate is a weak predicate fixed point of some binary predicate" ([11], p. 77). It is easy to see this in a theory that contains a truth predicate satisfying the (unrestricted) Tarskian T-scheme – since every sentence must satisfy the T-scheme, every sentence is a weak sentential fixed point of the truth predicate[12]. If fixed points are so common and so well distributed among sentences and predicates, how can we see the nature of paradoxicality in them?

[12]Even in a theory without the truth predicate or the T-scheme, it is possible to obtain a similar result, as shown in [11], p. 78.

Why, after all, do they work innocuously for most sentences, but supposedly bring out paradoxes in others? It would be insufficient to affirm, faced with this question, that fixed points only cause paradoxes when combined with semantic concepts, and not in any other contexts. After all, as I said it, every sentence is a fixed point of the truth predicate (if, of course, the necessary conditions are fulfilled).

Although Cook's claim that every sentence is a fixed point to some predicate (or function) is perfectly true, I argue that he is drawing the wrong conclusions from it: we should not conclude, from this, that "circularity is widespread," but that this must *not* be the right way to understand circularity. If we say that a sentence is circular when it is a fixed point to some function, then this notion is trivial, for every sentence is a fixed point to some function. It is not circularity that is widespread, it is just that we have been understanding circularity wrongly: we should look for a definition that does not depend on fixed points. As I said before, I will approach this question in more detail in the next chapter.

For now, let's look at another argument provided by Priest to defend the existence of circularity in Yablo's paradox. He says:

> This answers a question that should have been obvious as soon as one reads Yablo's description of the situation. He asks us to imagine a certain sequence. How can one be sure that there is such a sequence? (We can imagine all sorts of things that do not exist.) As he presents things, the answer is not at all obvious. In fact, we can be sure that it exists because it can be defined in terms of S: the n-th member of the sequence is exactly the predicate $S(x)$ with 'x' replaced by $\ulcorner n \urcorner$ ([36], p. 238 - notation changed to agree with this text).

The point illustrated is of an epistemological nature: we can only *know* Yablo's paradox using the Gödelian method of diagonalization, which itself involves circularity. Beall, in [4], explains Priest's point more clearly:

> If we have fixed the reference of "Yablo's Paradox" at all, then we have fixed the reference of "Yablo's Paradox" via (attributive) description. But, now, the upshot of Priest's point is plain: Priest has shown that any description we employ to pick out (or otherwise define) a Yabloesque sequence is circular [...] From here it is a small step to the circularity of the sequence itself ([4], p. 105).

Cook, in turn, sees this argument as a misunderstanding on the part of the other two authors of what actually happens in Gödel's diagonalization technique.

The method allows us to derive the following *Uniform Fixed-Point Yablo Principle* (UFPY) ([11], p. 24 – nomenclature taken from [23]) as theorem of arithmetic:

$$\forall z(Y(z) \leftrightarrow \forall n(n > z \rightarrow \neg Sat(\ulcorner Y(x) \urcorner, n)))$$

This principle is a generalization of the previously $Y(1), Y(2), ...$ sequence of sentences described before. If UFPY is a theorem, it means that there is a proof of this sentence in arithmetic. Although it is undecidable whether a sentence is a theorem of arithmetic, theorems are enumerable through valid proofs. Cook also adds that it is possible to decide whether a Gödel number is a number of a valid proof ([11], p. 107). To find, know, or construct Yablo's Paradox, we can simply run through the valid theorems of arithmetic, along with some necessary tools such as the truth predicate, and we will eventually find some sentence like UFPY for any predicate. Diagonalization guarantees that we will eventually find the theorem. Therefore, Cook argues that circularity is not necessary to construct or know the paradox, even though it does exist as a predicate fixed point. Nonetheless, the simple method of going through the enumerated proofs of a theory is not circular.

From the discussion, it's safe to conclude that, if circularity manifests itself in the presence of fixed points, then Yablo's Paradox is undoubtedly circular, for it contains a predicate fixed point: a predicate stating that no number greater than x satisfies itself. If all that holds, then the thesis that circularity or self-reference is a necessary component to semantic paradoxes still holds and Yablo's Paradox is no threat to it. This is important especially for someone like Graham Priest, who has developed the Inclosure Schema, a formal schema made to capture circularity, and which should capture every semantic paradox there is – and it indeed captures Yablo's. What now needs to be drawn into question is, as I've said before, the equivalence of the notions of self-reference and circularity to that of fixed points. As it stands, the discussion remains vague: it's not enough to state that circularity arises with fixed points. Is circularity *a fixed point itself,* or is it a sentence that is a fixed point to a function? Even if that is solved, there is still the question of what self-reference is and how it's different from circularity. In the next chapter, I will provide an answer to those questions and, only after that is settled, decide whether Yablo's Paradox is truly circular or not.

Chapter 3

Self-Reference and Circularity

3.1 Introduction

Certain logical paradoxes, such as the Liar, Russell's and others, have been categorized as *self-referential* or *circular*. In view of this, it is only natural to ask: "*What is self-reference and what is circularity?* And why do all these paradoxes fit one or the other label?" However natural those questions may seem, little attention has been given to these problems in the literature in logic. Much has been said, of course, on *how to achieve* self-reference and circularity by using logical and mathematical tools, but logicians have not devoted much time to defining the concepts directly. The absence of this discussion becomes all the more perplexing if we consider the active debate on whether certain paradoxes, such as Yablo's Paradox, are self-referential, circular or neither; and also whether there can be non-circular or non-self-referential semantic and set-theoretic paradoxes at all.[1] Without a clear idea of what self-reference is, what circularity is and what the differences between the two are, this debate is bound to remain vague and obscure: we must know *what we are talking about* before we can classify logical paradoxes and judge whether those are necessary features of them. This is precisely my aim in this chapter: to construct formal definitions for the concepts of self-reference and circularity, and analyze some paradoxes in light of these definitions.

I will start by providing an idea of the concept of reference in formal settings, which I aim to capture with one of my definitions. Then, I will consider the most

[1] Here, I am addressing the widely held belief described by Roy Cook: "The Yablo paradox threatens to overthrow the rather entrenched idea that paradoxes (or, at least, the semantic and set-theoretic paradoxes) are intimately and ineliminably tied to self-reference or circularity of some sort" ([11], p. 2).

common appreciation of the concept of circularity in logic, that circularity is manifested in the presence of fixed points. Next, I will provide reasons to argue that this notion is not sufficient, and that we should look for other definitions of these concepts. This will lead me to my proposal, in which I argue that reference is interpretation-dependent and hyperintensional, for it escapes logical equivalence. I will explain how, under my definitions, the Liar Paradox is self-referential, the Liar Cycle is circular, Russell's Paradox is non-well-founded and Yablo's Paradox is neither self-referential nor circular. With the last result, I dissent from the common diagnosis (defended by Priest in [36] and, later, by Beall in [4]) that Yablo's Paradox contains a hidden circularity, despite Yablo's attempt to formulate a non-circular paradox (in [58]). The question whether Yablo's Paradox is non-well-founded remains open, as I show in 3.4.5.

3.2 What is Reference?

Before diving into self-reference and circularity, one must first provide an answer to the question "what is reference?," or at least determine the type of reference which is at stake. This is the question I will start the paper with, and I will begin by considering some philosophical motivations for the view I argue for. My aim in this section is not yet to give a precise definition of reference, but a philosophical outline of the concept.

First of all, my approach will consider reference for formulas and sentences (*i.e.* both closed and open formulas).[2] Since, in the context of formal languages, self-reference and circularity are essentially features of formulas, sentences and sequences of sentences, they must be able to *refer*. For example, we say that the sentence which expresses the Liar Paradox is self-referential – so, the sentence must refer to itself. Therefore, I take the referring objects to be open and closed formulas, and the referents of a whole formula to be the referents of each of its terms; this means that a formula can refer to *more than one thing.* Another important caveat is that my primary worry in this paper concerns definitions for formal languages. Even though I believe my definitions *could* be extended to non-formal languages, it would take many more pages to defend this point (I will

[2]Someone may find it strange to allow for *open* formulas to refer, not just sentences. I must allow this to consider reference in Yablo's Paradox (which is one of my main objectives). The Yablo sequence, as pointed out by Priest in [36], is formed by open formulas. Moreover, there is no good reason to prohibit open formulas from referring: as the reader will see later, with one small exception, a formula refers to what its *closed* terms refer to. So the fact that the formula is open is merely "accidental," for only its closed terms refer.

come back to this in some paragraphs).

What may reference be? A possible view is to take reference to be analogous to the notion of *occurrence*, so a sentence φ refers to u when the name of u – let's say $\ulcorner u \urcorner$ – occurs in φ. That seems fairly natural at first glance: reference would then be a purely syntactic instrument to count the incidences of certain terms in a sentence. This is more or less the way Picollo, in [34], views reference.[3] Although this definition seems perfectly plausible, there is more to reference than mere occurrence: even though 'Cicero' and 'Tully' are different terms, – that is, different marks on a sheet of paper – they refer *to the same object.* So reference is something other than occurrence, and their fundamental difference is that reference is a *semantic* relation, rather than a purely syntactic one: the previous example exhibited two syntactically different terms whose meaning is the same, and so they have the same referent. To establish the referents of a sentence, we must think about the meaning of each of its terms. While occurrence denotes simply the incidences of a symbol in a sentence, reference is a relation that holds between words or sentences in a language and the objects they refer to, somewhere *outside* the language. The referent 'snow' of the sentence "snow is white," for example, is a certain object known to us, with certain definite characteristics.[4] To grasp it, it seems we must understand the term's meaning.

This leads to the following position: that the referents of a sentence depend on the objects we assign to each of its terms. In formal languages, which are my primary concern here, the object we assign to each of the terms depends on the *interpretation* we choose. By 'interpretation' I wish to denote a concept from Model Theory that can also be called *structure*, and is used to provide a semantic framework for formal languages. This also shows that reference is a relation that requires two linguistic levels, for it can only be expressed in a *metalanguage*. So the reference of a term will depend on the metalanguage – the model-theoretic interpretation – we choose. This point will be further developed in section 4.1. This motivates the careful step of limiting this research (at least at first) to formal languages: when it comes to formal languages, it is simply a fact that their terms

[3] Picollo does not approximate reference to occurrence explicitly in her paper, but her *Definition 1* is similar to the one I sketched in this paragraph (with the difference that, in her definition, the subsentence t must be tied to the truth predicate, because she aims to define *alethic reference*). For her, reference is connected to the syntactic structure of the sentence (hence its approximation with occurrence).

[4] There is, of course, the problem of characterizing 'snow': is it really *one* object? If so, what does it consist of? Quine suggests that the object 'snow' is the mereological sum (the fusion) of all portions of the spaciotemporal world that consist of snow ([42], p. 47). I will not, however, dwell on such problems more thoroughly, for they fall out of the scope of this text.

can be interpreted in different ways by various distinct structures, so it seems natural to defend the concept of reference as interpretation-dependent. But it would take many more pages to defend this view for non-formal language itself, for I would first have to determine what an interpretation of non-formal language is, and then decide if it can have more than one interpretation, which are quite demanding tasks. So this rests as a future endeavor.

Lastly, I must add that I will not adopt a Fregean view – I will not take a sentence's referent to be its truth value, as the logician argues in [16]. However, if desired, one could adopt this view while maintaining the key characteristic of the one outlined here: the model-theoretic notion of interpretation could still be used to formalize reference, with certain adaptations to include the values 'True' and 'False' in its domain.

3.3 Circularity as Fixed Point

In [27], Hannes Leitgeb puts forward two different notions[5] of self-reference and circularity that, he believes, underlie logicians' discourse about paradoxes, but are not explicitly defined by them. The first one takes circularity to correspond to the presence of fixed points, and is in the core of Graham Priest's argument in [36] (and Beall's [4] endorsement of this argument) for the thesis that Yablo's Paradox, despite appearances, is circular. Leitgeb defines this notion of circularity in the following way[6] (here, C_1 stands for the predicate "is circular"):

$C_1(x) \leftrightarrow_{df} x$ *is a sentence* $\wedge\ \exists y \exists z \exists f$ *(*y *is a term* $\wedge\ x$ *contains* $y \wedge$ *ref* $(y, z) \wedge f$ *is a syntactical mapping* $\wedge f(z) = z$*)*

A syntactical mapping is, roughly, a mapping that is defined by syntactic operations, such as certain concatenation of strings. Fixed points are understood as usual: z is a fixed point of f if and only if $f(z) = z$. Note also that *ref* (y, z) means that y refers to z (Leitgeb takes *ref* as a primitive relation, a point on which I will dwell in section 4). What the definition says is that a sentence is circular whenever the referent of one of its terms (which can be the very term itself) is a fixed point to a syntactical mapping. This view of circularity is not

[5] In this section, I will present one of Leitgeb's definitions and explain why it cannot be a good one, at least as it is currently formalized. His other definition is somewhat incomplete, and I try to complete it in my proposal. Therefore, Leitgeb's other definition will be exposed in section 4 together with my view.

[6] [27], p. 4.

restricted to Priest and Beall, but rather widespread in logic texts, such as in Smullyan [48], Cook [11] and others (even though it is not formalized thus by the authors). In a way, although it is rather far from the folk understanding of the words "self-reference" and "circularity," the fixed point idea might be said to be the standard interpretation of those terms in logic. And it is not randomly so: it does capture, to an extent, some statements that are intuitively circular, such as the (Untrue) Liar Paradox:[7]

(1) Liar Paradox

$$\textbf{(L)}\ \neg Tr(\ulcorner L \urcorner)$$

Here, L is the sentence which states that L itself is not true.[8] L is the fixed point of a mapping f that maps each formula A to the code of the formula concatenated with the negation symbol and the truth predicate symbol. The mapping is such that $f(\neg Tr(\ulcorner b \urcorner)) = \neg Tr(\ulcorner \neg Tr(\ulcorner b \urcorner) \urcorner)$ ([27], p. 5). This definition is, then, in perfect agreement with the pre-theoretic idea that there is something circular in the Liar Paradox. Yablo's Paradox also falls under this definition of circularity – or, rather, the arithmetical reconstruction of Yablo's Paradox, as seen below:[9]

(2) Yablo's Paradox

$$\begin{aligned} s(1) &\leftrightarrow \forall n(n > 1 \rightarrow \neg Sat(\ulcorner s(x) \urcorner, n)) \\ s(2) &\leftrightarrow \forall n(n > 2 \rightarrow \neg Sat(\ulcorner s(x) \urcorner, n)) \\ &\vdots \\ s(m) &\leftrightarrow \forall n(n > m \rightarrow \neg Sat(\ulcorner s(x) \urcorner, n)) \\ &\vdots \end{aligned}$$

Each formula in the Yablo sequence states that all formulas *above it* in the sequence are not true; or, better, that the formula $s(x)$ is not satisfied by n, for

[7] In (1), Tr is taken to be the truth predicate, $\neg$ is the usual symbol of negation and the corner quotes $\ulcorner$, $\urcorner$ signal sentences' names (so $\ulcorner L \urcorner$ denotes the name of L). Hence $\neg Tr(\ulcorner L \urcorner)$ is the sentence which says that L is not true.

[8] To see how this leads to paradox, suppose that L is true. If so, then what it says is true; but the sentence says that it itself is not true, so L is not true. Contradiction. Now suppose L is not true. Well, if this is the case, then what it says must not be true; but it states that it itself is not true, and that is true by supposition – so the sentence is, after all, true. Contradiction. Hence the paradox.

[9] In (2), Sat is the satisfaction predicate, and so the formula $\forall n(n > m \rightarrow \neg Sat(\ulcorner s(x) \urcorner, n))$ says that for all $n > m$, n does not satisfy the code of the predicate $s(x)$.

all n larger than the the subscript number of the formula in question.[10] The fixed point arises when $s(x)$ is taken to be a function, rather than each Yablo sentence. In Priest's own words,

> The function s is defined by specifying each of its values, but each of these is defined with respect to s ... It is now the function s that is a fixed point. s is the function which, applied to any number, gives the claim that all claims obtained by applying s itself to subsequent numbers are not true. Again the circularity is patent. ([36], p. 239).

Therefore, the function s is a fixed point of the binary predicate $Sat(\ulcorner s(x) \urcorner, n)$ because it figures on both sides of the biconditional.[11] If this notion captures, at least to an extent, our intuition of the concept of circularity, why not adopt it? The first reason is that it does not allow us to differentiate between self-reference and circularity. In Priest's paper, for example, the terms 'circularity' and 'self-reference' are used interchangeably (or at least without clear distinction).[12] This is a problem because, although closely related, those two terms seem to point to different relations: while self-reference is obviously tied to the notion of reference, circularity is not necessarily linked to it – even if reference plays a role in it, it could be an indirect one.

The second and most important problem with this picture is that it is ultimately *trivial*: "virtually every sentence is the fixed point of some mapping, in particular if only equivalence in one or another sense is demanded." ([27] p. 8). One example given by Leitgeb is that of a mapping g that maps the name of each formula A to the concatenation of its name + the equality sign + the conjunction sign + the formula itself – that is, $g(\ulcorner P(A) \urcorner) = \text{“}(\ulcorner P(A) \urcorner = \ulcorner P(A) \urcorner) \wedge P(A)\text{”}$.

[10] The paradox arises from the following argument (which I will explain using the truth predicate, not the satisfaction predicate, for clarity's sake): Suppose $s(m)$ is true. $s(m)$ says that all formulas above it are not true; so, in particular, $s(m+1)$ is not true. But $s(m+1)$ states that all formulas above *it* are not true, and if $s(m)$ is true, then all formulas above $s(m+1)$ are indeed not true. Therefore, $s(m+1)$ is true – contradiction. Now suppose $s(m)$ is *not* true. Then, there must be a *true* $s(k)$, for $k > m$. But now the same reasoning undertook to $s(m)$ can be followed to show that $s(k)$ cannot be true, and we have another contradiction. Hence the paradox.

[11] The fixed point in Leitgeb's definition of circularity does not capture the fixed point in Yablo's Paradox so precisely. A more precise notion would be what Roy Cook, in [11], defines as *weak predicate fixed point*: A unary predicate $\phi(x)$ is a weak predicate fixed point of a binary predicate $\psi(x, y)$ iff $\phi(x) \leftrightarrow \psi(\ulcorner \phi(x) \urcorner, x)$.

[12] In the introduction, Priest says that he will demonstrate that "self-referential circularity" is involved in Yablo's Paradox ([36], p. 236). Further on, he argues that the fixed point in Yablo's Paradox is of a self-referential kind and, in the same paragraph, that "the circularity is now manifest" (p. 238).

Thus $P(A)$ "is a fixed point of g, or, rather, it is a fixed point of g up to arithmetical equivalence, i.e., the formula encoded by the g-image of the code of $\ulcorner P(A) \urcorner$ is equivalent to $\ulcorner P(A) \urcorner$ in the standard model of arithmetic, *i.e.*, it is arithmetically true that $(\ulcorner P(A) \urcorner = \ulcorner P(A) \urcorner \wedge P(A)) \leftrightarrow P(A)$" ([27] p. 9, notation changed to match mine). Furthermore, there is another even simpler example to illustrate the point: if we let f be a function that assigns the truth predicate to the name of sentences and A a sentence, so that $f(A) = Tr(\ulcorner A \urcorner)$, then, by the Tarskian T-schema,[13] every sentence is a fixed point of f, for the equivalence $A \leftrightarrow f(A)$ holds (if the unrestricted T-schema is adopted). Now, if we hold C_1 as our definition of circularity, we must conclude that *every sentence is circular*, which simply cannot be true.

That should be enough justification to motivate us to look for other definitions of self-reference and circularity. It is true that there might be a way to circumvent some of these problems: maybe the definition of circularity as fixed point should not be that a sentence is circular if it is a fixed point for *any* mapping; perhaps it should be something less general. I do not see, however, any simple solution in sight that can both escape triviality and distinguish the concept of circularity from self-reference.

3.4 My Proposal

3.4.1 Reference

In this and in the next section, I will explain what it is for a *formula* (or, in particular, sentence) to refer to something and, moreover, to refer to itself. My definition of reference is inspired by Leitgeb's other definition ([27], p. 2), which goes like this:

$ref_1(x, y) \leftrightarrow_{df} x$ *is a sentence* $\wedge\ \exists z(z$ *is a closed term* $\wedge\ x$ *contains* $z \wedge$ *ref* $(z, y))$

Closed terms are terms with no free variables. Note that $ref_1(x, y)$ means that "x refers to y." From this, the author defines self-reference as expected:

[13] As shown in section 1 from chapter 2, the Tarskian T-schema is the equivalence $Tr(\ulcorner A \urcorner) \leftrightarrow A$, formed from the rules of Capture: $A \vdash Tr(\ulcorner A \urcorner)$ ('A' semantically implies $Tr(\ulcorner A \urcorner)$), and Release: $Tr(\ulcorner A \urcorner) \vdash A$, ($Tr(\ulcorner A \urcorner)$ semantically implies 'A'). The idea of the schema is that, if we claim that a sentence is the case, then we can claim that it is true – that is, capture it with the truth predicate. If we claim that a sentence is true, we can assert it, detach it from the truth predicate.

$$selfref_1(x) \leftrightarrow_{df} ref_1(x, x)$$

In Leitgeb's definition, a sentence[14] refers to what its closed terms refer to, and nothing more. I consider this view to be on the right track to a satisfactory definition of reference, and I maintain that a sentence's reference depends on the referents of its closed terms. However, in ref_1, the relation ref remains undefined: Leitgeb takes the definition of reference of terms as a primitive, describing it as "the usual reference relation for terms" ([27], p. 2). The problem is that there is no standard formalization of reference; furthermore, if we really wish to explain this concept (and the concepts of self-reference and circularity), we must understand exactly what the reference relation consists of. This is what I will do now. But first, I shall define the model-theoretic concept of *interpretation* since, as mentioned in section 2, my definition of reference depends on it:

Definition. Let $\mathfrak{L}$ be a first-order language. An *interpretation* I of $\mathfrak{L}$ consists of

(1) A non-empty set D_I, named *domain* of I.

(2) For each constant symbol c of $\mathfrak{L}$, a specific element c_I of D_I.

(3) For each n-ary function symbol $\mathcal{F}$ of $\mathfrak{L}$, an n-ary operation $\mathcal{F}_I$ on D_I.

(4) For each n-ary relation symbol $\mathcal{R}$ of $\mathfrak{L}$, an n-ary relation $\mathcal{R}_I$ in D_I.

We write $I(a)$, for any closed term a of the language $\mathfrak{L}$, to mean "the interpretation of a in I," which is an element of D_I. This is precisely how reference will be understood: the referent of a closed term will be given by the interpretation attributed to the term in D_I. I take a formula to refer to what each of its terms refers to, and so one formula may have various different referents. We then have the following definition:[15]

Definition 1. Let $\mathfrak{L}$ be a first-order language, I an interpretation of $\mathfrak{L}$ and $\mathsf{Fm}(\mathfrak{L})$ the set of formulas of $\mathfrak{L}$. The reference relation for formulas relative to I is defined as a subset $\mathscr{R}_I$ of $\mathsf{Fm}(\mathfrak{L}) \times D_I$:

$$\mathscr{R}_I \subseteq \mathsf{Fm}(\mathfrak{L}) \times D_I$$

[14] Leitgeb defines reference and self-reference for *sentences*, not simply formulas. This is a problem since he wishes to evaluate Yablo's Paradox, but the paradox, if formalized correctly, consists on a sequence of open formulas, as Priest showed in [36]. So his definitions would not apply to the paradox just because the sequence does not contain sentences.

[15] In this chapter, all definitions created by me are enumerated. The definition of interpretation, as it is a standard definition in logic, is listed but not labeled.

For all pairs $(\varphi, a) \in \mathsf{Fm}(\mathfrak{L}) \times D_I$, $(\varphi, a) \in \mathscr{R}_I$ if and only if there is a closed term t occurring in φ and $I(t) = a$. Henceforth, I will write the relation $\mathscr{R}_I(\varphi, a)$ as $ref_I(\varphi, a)$.

3.4.2 Self-Reference

Having a precise definition of reference at hand, we can now define the notion of self-reference, which is the phenomenon of a formula (either a sentence or an open formula) which refers to itself. I will provide two possible definitions of the same concept: an interpretation-dependent self-reference and a *generalized* self-reference. The former is a particular case of the latter. Let me present the interpretation-dependent definition now:

Definition 2.1. Let φ be a formula of a language $\mathfrak{L}$ and I an interpretation of $\mathfrak{L}$. Moreover, let t be a term occurring in φ. Then, φ is *self-referential in* I if it satisfies the two following conditions:

(i) $\varphi \in D_I$

(ii) $ref_I(\varphi, \varphi)$ – *i.e.*, there is a term t occurring in φ such that $I(t) = \varphi$.

If conditions (i) and (ii) are satisfied, I write $selfref_I(\varphi)$.

The definition states that a formula φ is self-referential when it is in D_I and contains a term t such that the interpretation of t in I is φ itself.[16] Albeit its direct and straightforward character, it has the flaw of being restricted to a specific kind of interpretation, a restriction coming from condition (i): φ must figure in D_I to be evaluated as self-referential, because the interpretation of terms of the language $\mathfrak{L}$ are objects in D_I. If φ is not in D_I, then it is *vacuously* not self-referential. Thus, a very special kind of interpretation is needed to evaluate self-reference: one which contains the formulas of the language in its domain. However, to make self-reference a functional concept for formal languages, another definition is desirable: one that is not limited to such a specific kind of interpretation and, furthermore, that can evaluate self-reference across all interpretations of the language $\mathfrak{L}$ (*i.e.* one that is not interpretation-dependent). This will be achieved via the notion of generalized self-reference.

[16] Urbaniak, in [53], provides a similar definition using model-theoretic interpretations, but he identifies this phenomenon as *aboutness*, not self-reference ([53], p. 244). Accordingly, his definitions depend on different concepts, such as a constant occurring "informatively" in a formula. I chose the different path of defining reference and working with self-reference, not aboutness.

But first, before I present the generalized definition, a quick word about a language that can express self-reference: to contain self-referential formulas, the language $\mathfrak{L}$ must be able to *name its own formulas*. This can be done with some labeling technique; I will assume the constants method, which consists of the addition of a constant symbol to each and every formula of $\mathfrak{L}$; these constants function as labels of the formulas. [17] The necessity of this condition becomes clear when we look at the Liar Paradox (1): L is a sentence which applies the negation of the truth predicate *to its own name* – this application is necessary to generate the paradox.[18] Moreover, if φ is a formula of $\mathfrak{L}$, its name $\ulcorner\varphi\urcorner$ is a closed term of $\mathfrak{L}$. Since I is an interpretation of $\mathfrak{L}$, each term t of $\mathfrak{L}$ receives an interpretation $I(c) \in D_I$. In particular, for each formula φ of $\mathfrak{L}$, I contains an interpretation of the name of φ:

$$I(\ulcorner\varphi\urcorner) \in D_I$$

Bearing this in mind, I now present the generalized definition of self-reference:

Definition 2.2. Let $\mathfrak{L}$ be a first-order language. A formula φ of $\mathfrak{L}$ is *self-referential* if and only if $ref_I(\varphi, I(\ulcorner\varphi\urcorner))$ for all interpretations I of $\mathfrak{L}$. From now on, if $ref_I(\varphi, I(\ulcorner\varphi\urcorner))$ for all interpretations I, I write $selfref(\varphi)$.

This definition acts as expected: a self-referential formula φ is one that contains a term t such that $I(t)$ in all interpretations I of $\mathfrak{L}$ is the interpretation of the name of the formula φ, $I(\ulcorner\varphi\urcorner)$. Note that, even though this definition depends on the concept of interpretation, it does not depend on *one* interpretation in particular, and so it can be said that φ is self-referential *tout-court*, not self-referential in an interpretation I; for this reason, I omit the subscript I of $selfref$.

[17] The language $\mathfrak{L}$ plus the set of constants added to each formula φ of $\mathfrak{L}$ is called the *diagram language* of $\mathfrak{L}$. To learn more about this method, refer to [49]. I should add that, whereas the specific method chosen to name sentences is not *that* important to the overall idea, some methods might not work as well as others. Standard Gödel numbering is a method that presents a problem to the evaluation of self-referential sentences. There can be, for example, a sentence such as "1011 is an odd number" whose Gödel number is 1011; i.e. $\ulcorner$1011 is an odd number$\urcorner$ = 1011. Here, 1011 would be doing double duty as both the name of a formula and as a number itself. Hence, according to my definition, the sentence would be qualified as self-referential, although it really isn't, because the mention of 1011 refers to the very number 1011, not as the sentence itself. Therefore, even though the choice of naming device does not affect the definition in *most* cases, some might raise problems, such as the one outlined above.

[18] If, in addition to names for its own formulas, a language contains a truth or a satisfaction predicate that evaluates the sentences of the language itself, this language is called *semantically closed*. This terminology comes from Tarski in [51] and was used by the logician to designate languages which can formulate paradoxes. This will be discussed in detail in chapter 4.

Moreover, to show that a formula φ is *not* self-referential, it suffices to provide an interpretation under which no term t of φ is such that $I(t)$ is the same as $I(\ulcorner\varphi\urcorner)$.

From this last definition, one can derive the following result:

Proposition 1. If the name of the formula φ occurs in φ, then $selfref(\varphi)$, *i.e.* φ is self-referential.

Proof. Let $\ulcorner\varphi\urcorner$ be the name of a formula φ of $\mathfrak{L}$ and suppose $\ulcorner\varphi\urcorner$ occurs in φ. Then, $selfref(\varphi)$ if and only if there is a closed term t occurring in φ such that $I(t) = I(\ulcorner\varphi\urcorner)$ for all I of $\mathfrak{L}$. The result follows trivially from the fact that $\ulcorner\varphi\urcorner$ is a constant of the language $\mathfrak{L}$ – hence, a closed term occurring in φ – and so it is obvious that $I(\ulcorner\varphi\urcorner) = I(\ulcorner\varphi\urcorner)$ for all interpretations I of $\mathfrak{L}$.

■

With the generalized definition of self-reference and the above proposition, the difference between reference and occurrence is evident: while the right side of the implication follows, the left side *does not*, for a formula can be self-referential without having its name as one of its terms. The formula may contain another term t which is not exactly its own name, but is such that $I(t) = I(\ulcorner\varphi\urcorner)$ for all interpretations I of $\mathfrak{L}$.

Finally, with the above proposition, it follows trivially that the Liar Paradox as formulated in (1) is self-referential, for the name $\ulcorner L\urcorner$ of L occurs in the sentence $\neg Tr(\ulcorner L\urcorner)$. So $selfref(L)$. Moreover, a "standard" interpretation I could easily be constructed (that is, one that interprets the terms of $\mathfrak{L}$ as we understand them) such that $selfref_I(L)$.

3.4.3 Circularity

Now for circularity. How are we to capture the phenomenon when a sequence of formulas forms a referential circle? An immediate idea is to use the definition of reference elaborated in the previous sections to account for this notion. As with self-reference, we have an interpretation-dependent definition and a generalized one. The restricted definition of circularity goes like this:

Definition 3.1. Let $a_1, a_2, ..., a_n$ be a sequence of formulas of a language $\mathfrak{L}$ and I an interpretation of $\mathfrak{L}$. The sequence is said to be *circular in* I if it satisfies the following conditions:

(1) $a_1, a_2, ..., a_n \in D_I$

(2) $ref_I(a_1, a_2), ref_I(a_2, a_3), ref_I(a_m, a_{m+1}), ..., ref_I(a_n, a_1)$.

If conditions (i) and (ii) are satisfied, I write $Cir_I(a_1, a_2, ..., a_n)$. Any sequence of formulas with a circular-in-I segment is said to be circular-in-I too.

The last condition stated in the definition is necessary to account for infinite sequences: without it, an infinite sequence would be non-circular simply because circularity applies for finite ones. With the last condition, I guarantee that any infinite sequence containing a circular segment is also circular. In this way, circularity *spreads* from the finite to the infinite case. If a sequence possesses a circular-in-I segment, it cannot get rid of its circularity just by appealing to infinity.

This definition possesses the same virtues and flaws as $selfref_I$: on the one hand, it captures the *direct* character of circularity, for the sequence is circular whenever $a_1, a_2, ..., a_n \in D_I$ and there is a term t in a_1 such that $I(t) = a_2$, there is a term t in a_2 such that $I(t) = a_3$ and so on; finally, there is a term t in a_n such that $I(t) = a_1$. So the interpretations of the terms are the *very formulas themselves*. On the other hand, the concept is once more dependent on an interpretation whose domain includes the formulas of the language $\mathfrak{L}$. To solve this issue, one may define a generalized notion of circularity, as I did with self-reference. I will now spell out the definition of generalized circularity (even though the reader might have guessed it already):

Definition 3.2. Let $a_1, a_2, ..., a_n$ be a sequence of formulas of the language $\mathfrak{L}$. The sequence is said to be *circular* if and only if

$$ref_I(a_1, I(\ulcorner a_2 \urcorner)), ref_I(a_2, I(\ulcorner a_3 \urcorner)), ..., ref_I(a_m, I(\ulcorner a_{m+1} \urcorner)), ..., ref_I(a_n, I(\ulcorner a_1 \urcorner))$$

for all interpretations I of $\mathfrak{L}$. If the sequence is circular, I write $Cir(a_1, a_2, ..., a_n)$. Any sequence with a circular segment is said to be circular too.

Note that we say "circular in I" when referring to the interpretation-dependent definition circularity, and "circular" *simpliciter* when referring to generalized circularity. Thus, from now on, when I state that a sequence is circular, I mean that it is circular according to the generalized definition of circularity.

For the same reason as before, Cir_I is a special case of Cir (when the Correctness of Names assumption is preserved). Furthermore, both definitions of circularity are expansions of the definitions of self-reference: a self-referential formula is a circular formula (but the opposite is not always true). The loop-like phenomenon present in self-reference is *augmented* with circularity; from another

perspective, circularity can be seen as indirect self-reference. Hence, one may derive an equivalent of Proposition 1 to generalized circularity:

Proposition 2. If the name of the formula a_2 occurs in a_1, and the name of a_3 occurs in a_2 etc and, finally, the name of a_n occurs in a_1, then the sequence $a_1, a_2, ..., a_n$ is circular.

Proof. For any formula a_n, let its name be designated by the symbol $\ulcorner a_n \urcorner$. Now, consider the sequence of formulas $a_1, a_2, ..., a_n$ of the language $\mathfrak{L}$ and suppose $\ulcorner a_2 \urcorner$ occurs in a_1, $\ulcorner a_3 \urcorner$ occurs in a_2 etc and, finally, $\ulcorner a_1 \urcorner$ occurs in a_n. The reasoning is exactly the same as in Proposition 1: $\ulcorner a_2 \urcorner$, $\ulcorner a_3 \urcorner$, ..., $\ulcorner a_1 \urcorner$ are closed terms occurring in a_1, a_2, ..., a_n respectively, and so it is obvious that $I(\ulcorner a_2 \urcorner) = I(\ulcorner a_2 \urcorner), I(\ulcorner a_3 \urcorner) = I(\ulcorner a_3 \urcorner)$, ..., $I(\ulcorner a_1 \urcorner) = I(\ulcorner a_1 \urcorner)$, for all interpretations I of $\mathfrak{L}$. Hence, the sequence $a_1, a_2, ..., a_n$ is circular; in symbols, $Cir(a_1, a_2, ..., a_n)$. ■

From Proposition 2, we conclude that the following paradox is circular:

(3) Liar Cycle

$A = \neg Tr(\ulcorner B \urcorner)$
$B = Tr(\ulcorner A \urcorner)$

The Liar Cycle is a sequence of two sentences in which the first states that the second is not true, and the second states that the first is true.[19] Since $\ulcorner B \urcorner$ occurs in A and $\ulcorner A \urcorner$ occurs in B, by Proposition 2, the sequence is circular, and so $Cir(A, B)$.

The generalized definition of circularity enables us to conclude that *Yablo's Paradox is not circular*. To do that, I must simply present an interpretation in which the sequence comes out as non-circular. This is what I will do now.

Proposition 3. The Yablo Paradox is not circular.

Proof sketch.

Let $\mathfrak{L}_Y = \langle S, >, Sat, \ulcorner, \urcorner, 0 \rangle$ be a first-order language such that S is an unary function symbol, $>$ is a binary relation symbol, Sat is a binary relation symbol, $\ulcorner$ and $\urcorner$ are logical symbols and 0 is a constant symbol. A language with this structure can formulate Yablo's Paradox, for it possesses the successor function

[19] To understand the paradoxicality of the sequence, ask yourself if A is true or false. If A is true, then B is certainly false. However, B says that A is true, and this is true by assumption – contradiction. Now suppose A is false. Then, B must be true, but B says that A is true and, by assumption, that cannot be true – contradiction. Hence the paradox.

(from which it is possible to construct all natural numbers, starting with the constant symbol 0), the $>$ relation (so it is possible to write $m > n$), the satisfaction predicate Sat and $\ulcorner, \urcorner$, which name formulas of $\mathfrak{L}_Y$ (formulas with brackets on each of its corners are closed terms). Now, $I_Y = \langle S^{I_Y}, >, Sat^{I_Y}, 0^{I_Y} \rangle$ is an interpretation of $\mathfrak{L}_Y$, with domain D_{I_Y} as the set of natural numbers $\mathbb{N}$. Now let 0^{I_Y} denote the number 0, $S^{I_Y}(0^{I_Y})$ the number 1 and so on, as usual. Now consider the first formula in the Yablo sequence:

$$s(1) \leftrightarrow \forall n(n > 1 \rightarrow \neg Sat(\ulcorner s(x) \urcorner, n))$$

This formula has two closed terms: '1' and $\ulcorner s(x) \urcorner$. '1' refers to $S^{I_Y}(0^{I_Y})$, which is interpreted as the number 1. So the first term in $s(1)$ refers to *the number 1* (not the formula $s(1)$ itself). The other closed term in $s(1)$ is $\ulcorner s(x) \urcorner$. Since each constant symbol of the language is assigned to a specific element in D_{I_Y}, $\ulcorner s(x) \urcorner$ must be assigned to a number (because the elements of D_{I_Y} are numbers). I assign $\ulcorner s(x) \urcorner$ to 0^{I_Y}. So the second term of $s(1)$ refers to the number 0. Now, for each name $\ulcorner s(1) \urcorner$, $\ulcorner s(2) \urcorner$, $\ulcorner s(3) \urcorner$, ..., $\ulcorner s(n) \urcorner$, ..., I assign, respectively, $S^{I_Y}(S^{I_Y}(0^{I_Y}))$, $S^{I_Y}(S^{I_Y}(S^{I_Y}(0^{I_Y})))$ and so on; hence, $\ulcorner s(1) \urcorner$ refers to the number 2, $\ulcorner s(2) \urcorner$ refers to the number 3, $\ulcorner s(3) \urcorner$ refers to the number 4 and, in general, each $\ulcorner s(n) \urcorner$ refers to the number $n + 1$.

It is easy to see that Yablo's sequence is neither self-referential nor circular in this interpretation: no $s(n)$ is such that $ref_I(s(n), I(\ulcorner s(n) \urcorner))$, and no initial segment $s(1), ..., s(n)$ is such that $ref_I(s(1), I(\ulcorner s(2) \urcorner)), ref_I(s(2), \ulcorner s(3) \urcorner)..., ref_I(s(n), I(\ulcorner s(1) \urcorner))$. Since a sequence is circular when it is circular *for all interpretations* I, this model demonstrates that the Yablo sequence is *not* circular after all. Moreover, it would be easy to construct a "standard" interpretation I containing the formulas of $\mathfrak{L}$ in D_I, such that the sequence is not circular in I (*i.e.* it is not the case that $Cir_I(s(1), s(2), ..., s(n), ...)$.

3.4.4 Russell's Paradox and Non-Well-Foundedness

I started this chapter by saying that one of the objectives of my definitions is to contribute to the study of paradoxes, both by showing why some of them fit the 'self-referential' or 'circular' label and by deciding whether others do. Now that I have laid out the definitions, there is still one important paradox which escapes them: Russell's Paradox.

(4) Russell's Paradox

Comprehension Schema: $\exists y \forall x (x \in y \leftrightarrow \phi(x))$

The axiom allows us to define the set $\mathfrak{R} = \{x : x \notin x\}$. Is $\mathfrak{R}$ a member of itself? Instantiating the axiom, we have

$$\mathfrak{R} \in \mathfrak{R} \leftrightarrow \mathfrak{R} \notin \mathfrak{R}$$

Suppose $\mathfrak{R} \in \mathfrak{R}$. By the implication $\mathfrak{R} \in \mathfrak{R} \rightarrow \mathfrak{R} \notin \mathfrak{R}$, we have $\mathfrak{R} \notin \mathfrak{R}$, which is a contradiction. Now suppose $\mathfrak{R} \notin \mathfrak{R}$. By the implication $\mathfrak{R} \notin \mathfrak{R} \rightarrow \mathfrak{R} \in \mathfrak{R}$, we have $\mathfrak{R} \in \mathfrak{R}$, which is a contradiction. Hence, this is a paradox.

■

The characterization of a set that is a member of itself if and only if it is not a member of itself is undoubtedly a strange phenomenon, and seems, at least on the surface, to have something in common with circularity and self-reference. However, it still does not fit into any of the given definitions, which is an unsettling result. Why is that so? As it turns out, the reason is very simple: Russell's Paradox does not regard a formula referring to itself, nor a referential circle between two or more formulas. The paradox arises through a wicked *description* of a set: a description that presupposes, in itself, the very set described. The non-conformity of the paradox does not unveil a deep problem with the previous definitions, but a new type of phenomenon that must be analyzed. I take it to be *non-well-foundedness*, rather than circularity or self-reference. I will now explain why.

Let me start with the definition of a non-well-founded relation as given by Martin Pleitz:

> A relation R is well-founded (on a collection) if and only if there are no infinitely descending R-chains (within that collection). Otherwise it is non-well-founded ([41], p. 194).

Pleitz gives three different relations examples that are worth mentioning here: the relation expressed by the predicate "x rests on y" is well-founded with respect to the bricks of a house because each chain of bricks resting on one another will end in the basement. On the other hand, the successor relation with respect to the integers (both positive and non-positive) is non-well-founded, since it does not have a minimal element. Lastly, the relation expressed by "x cites y" is non-well-founded in some academics, for there may be citation cycles among their works; but there may also be academics with respect to which the citation relation

is well-founded. In the same way, an *object* is said to be well-founded with respect to a relation R (and a collection of objects) if and only if it is *not* the first element of an infinitely descending R-chain; otherwise, it is said to be non-well-founded. Thus, bricks are well-founded with respect to the resting relation, integers are non-well-founded with respect to the successor relation, and some academics are well-founded but others are not, with respect to the citation relation.

In the case of Russell's Paradox, the relation in question is membership. To see why the Russell "set"[20] is non-well-founded with respect to membership, we must first consider the Foundation Axiom, a set-theoretic axiom that establishes that there can be no ill-founded set. As the reader will see, it claims that there can be no chains of the sort described by Pleitz in [41] regarding membership. There are many ways to formulate this axiom; I will present only two of them, which will be useful for proving that the Russell set is non-well-founded (or, better, that there can be no Russell set in classical set-theory). See below:[21]

The Foundation Axiom (FA)

(1) There are no infinite sequences of sets

$$x_0 \ni x_1 \ni x_2 \ni ... \ni x_n \ni x_{n+1} \ni ...$$

each of whose terms is an element of the previous term.

(2) For every non-empty set X, there is some $y \in X$ such that $y \cap X = \varnothing$

The relationship between the first formulation of the axiom and Pleitz' definition of a well-founded relation is immediate – indeed, the former is an instantiation of the latter to the membership relation. It is a little harder to see why those two ways to write the axiom are equivalent. To make this clear, I will now prove their equivalence, which is a known result in logic.

Proposition 4. Formulations (1) and (2) of the Foundation Axiom are equivalent.

[20] The quotation marks are used to signal that, in classical set theory, Russell's "set" is not a set at all, but a class. In the more recently developed non-classical set theory, however, the set of all sets is indeed a set. I will sometimes use the expression "Russell's set" without quotation marks simply to preserve the aesthetic appeal of this text – the excess of quotation marks is always an ugly sight, just as the abuse of notation in some logic texts. If this happens, the reader should know that it is simply an aesthetic resource, rather than the postulation of an entity.

[21] These formulations of the axiom can be found in [31], as well as other interesting ways to write the same axiom.

Proof.

(i) (1) ⇒ (2).

Suppose (2) does not hold. Then, there is a set S such that, for every $y \in S$, $y \cap S \neq \varnothing$.

Now, select any member y_n of S. By assumption, $y_n \cap S \neq \varnothing$. Thus, there is an element, call it $y_{n+1} \in S$ such that $y_{n+1} \in y_n$. Since $y_{n+1} \in S$, the same argument may be repeated for y_{n+1}, and so there is a y_{n+2} such that $y_{n+2} \in S$ and $y_{n+2} \in y_{n+1}$. With the Axiom of Choice and the Recursion Theorem, we can define a function $f : \mathbb{N} \to S$ such that $f(k+1) \in f(k)$ for all $k \in \mathbb{N}$. The sequence defined by f is precisely

$$y_0 \ni y_1 \ni y_2 \ni ... \ni y_n \ni y_{n+1} \ni ...,$$

Therefore, (1) does not hold (S is ill-founded according to (1).)

(ii) (2) ⇒ (1).

Suppose (1) does not hold. Then, there is a set S defined by the function $f : \mathbb{N} \to S$ such that $f(k+1) \in f(k)$ for all $k \in \mathbb{N}$. Again, f creates the following sequence of members of S:

$$y_0 \ni y_1 \ni y_2 \ni ... \ni y_n \ni y_{n+1} \ni ...,$$

Now, select an arbitrary $y_n \in S$. By the definition of S, there is a y_{n+1} such that $y_{n+1} \in S$ and $y_{n+1} \in y_n$. Thus, $y_n \cap S = y_{n+1}$, which implies that $y_n \cap S \neq \varnothing$. Since y_n is arbitrary, it means that for all y_k, $y_k \cap S \neq \varnothing$. Thus, S is not in agreement with (2).

■

With this equivalence, we can prove that the Russell set is non-well-founded according to the second formulation of the Foundation Axiom and be sure that it is non-well-founded according to the first – and, hence, to the more general definition given by Pleitz. In fact, we will prove a stronger result: if the Foundation Axiom holds, there can be no set of all sets which are not members of themselves. So, if there were such a set, it would be non-well-founded. Before that, however, it is necessary to prove a simple lemma:

Lemma. If $\{S\}$ is non-well-founded, then so is S.

Proof.

Suppose $\{S\}$ is ill-founded. Then, by the second version of FA, $\{S\} \cap S \neq \varnothing$. Since S is the only element of $\{S\}$, $\{S\} \cap S = S$. Therefore, $S \in S$. But this gives rise to the following infinitely descending sequence of sets

$$S \ni S \ni S \ni S \ni ...,$$

which, by the first version of FA, means that S is ill-founded (i.e. S is not a set). ■

Now, we can prove that the Russell "set" $\mathfrak{R}$ is non-well-founded. Follow the proof below.

Proposition 5. If FA holds, then Russell's "set" $\mathfrak{R}$ is not a set.

Proof.

Assume FA and suppose that $\mathfrak{R}$ (in **(4)**) is a set.

Select $\{\mathfrak{R}\}$. If $\mathfrak{R} \in \mathfrak{R}$, then $\{\mathfrak{R}\} \cap \mathfrak{R} = \mathfrak{R}$, and so by FA $\{\mathfrak{R}\}$ is not a set. By the previous lemma, $\mathfrak{R}$ is not a set ($\mathfrak{R}$ is ill-founded). If, on the other hand, $\mathfrak{R} \notin \mathfrak{R}$, then, by the implication $\mathfrak{R} \notin \mathfrak{R} \rightarrow \mathfrak{R} \in \mathfrak{R}$ in **(4)**, it follows that $\mathfrak{R} \in \mathfrak{R}$. Again, by the previous argument $\{\mathfrak{R}\}$ is not a set. Thus, by the lemma, $\mathfrak{R}$ is not a set either ($\mathfrak{R}$ is ill-founded).

■

With these proofs, I intended to show that Russell's Paradox indeed involves non-well-foundedness, regardless of the definition of the concept we prefer to adopt (be it Pleitz' definition, the first or the second version of FA). So, when it comes to diagnosing Russell's Paradox, this is the phenomenon we are looking for, rather than circularity.

Identifying Russell's Paradox as arising from ill-foundedness must not be mistaken as saying that ill-foundedness is *sufficient* to create the paradox. As I mentioned earlier, there are many, perfectly classical, non-well-founded relations, such as the successor relation with respect to the integers. In fact, Peter Aczel created, in his influential book [2], a non-well-founded set theory: by adopting what he called the Anti-Foundation Axiom, the author populated his universe of sets with these strange beings with no recursive foundation and, which is most impressive, proved that his theory, ZFA (Zermelo-Fraenkel *without* the Foundation Axiom and *with* the Anti-Foundation Axiom), is consistent if and only if ZFC (Zermelo-Fraenkel + the Axiom of Choice) is consistent. Hence, non-well-foundedness is a *necessary* but *not sufficient* condition for the paradox to arise.

That non-well-foundedness is necessary but not sufficient to create paradoxes should come as no surprise: when I proved that the Liar Paradox is self-referential – which is a common place in the logic literature, although usually unproven –

or that the Liar Cycle is circular, I did not mean that those are necessary and sufficient conditions for the emergence of such paradoxes. Again, there are many self-referential and circular non-paradoxical sentences, such as "this sentence has five words" or the circle composed of A and B, in which A states "'B' is short" and B states "'A' has three words."

I began this section by exposing Russell's Paradox and mentioning the imprecise *feeling* it produces – a sense that something strange happens in its construction, almost as if we felt the presence of circularity and self-reference lurking behind us. Now, at the end of the section, that elusive presence has been revealed: it was not the phantom of circularity or the spirit of self-reference, but a new concept that plays in the gardens on infinity: non-well-foundedness. But now, as this research itself seems to play in those same gardens, another question arises: what is the relationship between this new entity, ill-foundedness, and the good old self-reference and circularity? Moreover, can it tell us something about Yablo's Paradox? I will tackle these questions in the next section and try to lead us to the end (if this is even possible) of this chapter.

3.4.5 Yablo's Paradox and Non-Well-Foundedness

At first glance, the structure of Yablo's Paradox certainly seems to be, in some way, non-well-founded. But the world is never that simple: there are some subtleties in the construction that prevent us from concluding, with certainty, that it is ill-founded.

One of the problems is that, in spite of the clear similarity of the sequence with a phenomenon of non-well-foundedness, we cannot simply prove that it disrespects the Foundation Axiom, for the axiom concerns *sets* and the *membership relation*, and here we are talking about *sentences* and the reference relation. In view of this scenario, we need a different strategy to consider whether the paradox is ill-founded or not. Fortunately, there is a very simple and straightforward way to capture Pleitz' description of ill-foundedness for the reference relation. It requires no new elements besides the ones we already had from the previous definitions (of reference, self-reference and circularity); using them, we can define non-well-foundedness to cases that are relevant for this text. This definition will allow me to prove that a simplified version of the Yablian sequence is non-well-founded. Another problem is that I can only develop this proof for the simplified version, not for the paradox itself.

A different way to tackle the issue is to draw a parallel between Yablo's

structure and a similar set-theoretic one and, finally, show that this similar structure violates the Foundation Axiom – this second strategy, although perhaps less precise than the first, is important because it relates the paradox to the Foundation Axiom, which is usually the standard "unity of measure" of ill-founded relations.

I will begin with the first task: presenting a definition of non-well-foundedness for the reference relation on sentences, and proving, with it, that a simplified Yablo sequence is non-well-founded. The definition is presented below. Mind that the relations it presupposes have already been defined in the earlier sections.

Definition 4. Let $a_1, a_2, ..., a_n, a_{n+1}, ...$ be an infinite sequence of formulas of the language $\mathfrak{L}$. The sequence is said to be *non-well-founded with respect to the reference relation* if and only if

$$ref_I(a_1, I(\ulcorner a_2 \urcorner)), ref_I(a_2, I(\ulcorner a_3 \urcorner)), ..., ref_I(a_m, I(\ulcorner a_{m+1} \urcorner)), ...$$

for all interpretations I of $\mathfrak{L}$. If the sequence is non-well-founded with respect to the reference relation, I write $\exists_{ref}(a_1, a_2, ..., a_n, a_{n+1}, ...)$.

The definition is exactly the same as the general definition of circularity, with the exception that an ill-founded sequence is *not* necessarily a circle: there is no sentence a_n such that $ref_I(a_n, I(\ulcorner a_1 \urcorner))$, which would form a complete, closed circle. This definition is the *right one* because it captures exactly what Pleitz conveys in his description and what the Foundation Axiom aims at blocking: an *indeterminate* sequence, one in which the first term depends on the second, the second on the third and so on, to infinity. Thus, we can never *truly* determine the reference of sentences $a_1, a_2, ..., a_n, a_{n+1}, ...$, not even of a_1, because a_1 refers to a sentence that, in its turn, refers to another, and that one to yet another; they are all indeterminate, their foundation is wrecked.

To prove that the Yablo sequence of sentences is non-well-founded, we must first prove an intermediate result similar to the other two we have proved before for circularity and self-reference (propositions 1 and 2). See the proof below.

Proposition 6. If the name of the formula a_2 occurs in a_1, and the name of a_3 occurs in a_2 etc, so that the name of every formula a_{n+1} occurs in a_n, then the sequence $a_1, a_2, ..., a_n, ...$ is non-well-founded with respect to the reference relation; in symbols, $\exists_{ref}(a_1, a_2, ..., a_n, ...)$.

Proof.

For any formula a_n, let its name be designated by the symbol $\ulcorner a_n \urcorner$. Now, consider the sequence of formulas $a_1, a_2, ..., a_n$ of the language $\mathfrak{L}$ and suppose $\ulcorner a_2 \urcorner$ occurs in a_1, $\ulcorner a_3 \urcorner$ occurs in a_2 so that, for all a_{n+1}, $\ulcorner a_{n+1} \urcorner$ occurs in a_n. The reasoning is exactly the same as in propositions 1 and 2: $\ulcorner a_2 \urcorner$, $\ulcorner a_3 \urcorner$, ..., $\ulcorner a_{n+1} \urcorner$, ... are closed terms occurring in a_1, a_2, ..., a_n, ... respectively, and so it is obvious that $I(\ulcorner a_2 \urcorner) = I(\ulcorner a_2 \urcorner), I(\ulcorner a_3 \urcorner) = I(\ulcorner a_3 \urcorner)$, ..., $I(\ulcorner a_{n+1} \urcorner) = I(\ulcorner a_{n+1} \urcorner)$, ..., for all interpretations I of $\mathfrak{L}$. Hence, the sequence $a_1, a_2, ..., a_n, ...$ is non-well-founded with respect to reference; in symbols, $\exists_{ref}(a_1, a_2, ..., a_n ...)$.

■

With this proof at hand, let us take a second look at Yablo's sequence of sentences. I will reproduce paradox (2) from 3.3 here to facilitate the reading. The proof does not *immediately* apply to the sequence but, with some modifications, we can use it to conclude that a simpler version of the sequence of sentences $s(1), s(2), s(3), ...$ below is non-well-founded.

(2) Yablo's Paradox

$$s(1) \leftrightarrow \forall n(n > 1 \rightarrow \neg Sat(\ulcorner s(x) \urcorner, n))$$
$$s(2) \leftrightarrow \forall n(n > 2 \rightarrow \neg Sat(\ulcorner s(x) \urcorner, n))$$
$$\vdots$$
$$s(m) \leftrightarrow \forall n(n > m \rightarrow \neg Sat(\ulcorner s(x) \urcorner, n))$$
$$\vdots$$

There is a subtlety in the sentences: they are quantified, and according to my definition there cannot be reference by quantification. This means that $s(1)$ cannot refer to *all* $s(n)$ such that $n > 1$. Reference requires a token of each object to which we refer, so the sentence "All dogs have a heart" does not refer to each and every dog that has ever existed. I will discuss this matter more thoroughly in 3.5.3. Hence, none of the sentences in the sequence can refer to all sentences above it; yet, they all seem to *say something* about all the sentences above them. This tension can be solved by a *simplification* of the sequence which preserves its most important characteristics:

Simplified Yablo Sequence:

$$s'_1 \leftrightarrow \neg Tr(\ulcorner s'_2 \urcorner)$$
$$s'_2 \leftrightarrow \neg Tr(\ulcorner s'_3 \urcorner)$$
$$s'_3 \leftrightarrow \neg Tr(\ulcorner s'_4 \urcorner)$$
$$\vdots$$

$$s'_n \leftrightarrow \neg Tr(\ulcorner s'_{n+1} \urcorner)$$
$$\vdots$$

If any sentence $s(n)$ in the Yablo sequence says that all subsequent sentences are not true, then, in particular, $s(n)$ says that $s(n+1)$ is not true. Hence, in this simplification we are simply particularizing a universal claim – a procedure called universal instantiation, which is allowed in first-order logic by the $\forall$-elimination rule introduced in section 4 from chapter 2. When it comes to the simplified Yablo sequence, it is easy to see that Proposition 6 implies that it is non-well-founded: the name of each sentence s'_{n+1} occurs in s'_n, so, by Proposition 3, it follows that $\exists_{ref}(s'_1, s'_2, s'_3, ..., s'_n, ...)$. The problem with this argument is that the Simplified Yablo Sequence is, as it turns out, *not* paradoxical: there is a consistent truth-evaluation of the sentences in the sequence. Their truth values are always alternated, but not inconsistent: if s'_1 is true, then s'_2 is false, and so s'_3 is true and so on. If s'_1 is false, then s'_2 is true, and so s'_3 is false and so on. No problem there, this is all perfectly classical.

I'm afraid this investigation led us to a cross road: on the one hand, a simplified version of the Yablo sequence is non-well-founded with respect to the reference relation. Under a hasty glance, one could conclude that, if the simplified version of the sequence is ill-founded, so is the full Yablo sequence – specially since the latter is just a universal generalization of the former, and we would assume that properties such as ill-foundedness would be inherited in the universal generalization process. On the other hand, the sequence proven to be non-well-founded is *not* paradoxical, contrary to the standard version of the Yablo sequence. So what can we make of this? Is the simplified sequence ill-founded, while the paradoxical version isn't? If that is so, does ill-foundedness bear no relationship to paradoxicality? Or are they both non-well-founded? I must admit I do not have the answer to all those questions at the moment. All I can do is present another argument hinting to the hypothesis that the sequence might be non-well-founded. This is what I will do now.

The way mathematicians and logicians use the word "foundation" is not dissimilar from the way it is used by architects and engineers in construction projects: if they affirm that a building has *good foundations*, they probably mean that it is laid upon strong bricks and cement. When mathematicians say that a set is well-founded, they mean that the set is "built" upon smaller sets that can be constructed recursively (and, thus, that the set is in agreement with the Foundation Axiom). The idea is, on both sides, the same; what changes is merely

the object in question. Now, when such object is neither a building nor a set, but a sequence of sentences, what would the meaning of foundation be? It is natural to say that a well-founded sentence one formed by smaller, "sentential" parts, that are themselves formed by smaller parts, until we reach terms that cannot be unfolded into other smaller ones. Following this parallel, we can relate sets to sentences by picturing a membership relation on sentences: in a way, every sentence determines a set, one containing as members the words that compose it. Hence, "snow is white" would form the set $W = \{\{snow\}, \{is\}, \{white\}\}$.

In the same way, a sentence referring to another contains it as an element. Thus, s'_1, which says "$\ulcorner s'_2 \urcorner$ is not true," is the set $s'_1 = \{\{s'_2\}, \{is\}, \{not\}, \{true\}\}$. Since s'_2 is the name of a certain sentence, we can also write the same set as $s'_1 = \{\{s'_3\ is\ not\ true\}, \{is\}, \{not\}, \{true\}\}$. Now, it is clear that $s'_2 \in s'_1$. By the same reasoning, $s'_3 \in s'_2$, $s'_4 \in s'_3$ and so on, creating the infinitely descending sequence

$$s'_1 \ni s'_2 \ni s'_3 \ni ... \ni s'_n \ni ...$$

Which is exactly the type of sequence prohibited by version (1) of the Foundation Axiom. Therefore, it follows that the set-theoretic simplification of the Yablo sequence is non-well-founded.

Again, whether this heuristic argument is sufficient to show that non-well-foundedness figures, in some way, in Yablo's Paradox, I do not know. Before ending the section, I would like to add an important detail: given the definition of non-well-founded sequences of formulas with respect to reference, it seems that the Liar Paradox itself is non-well-founded (besides being self-referential, as I proved earlier). The definition presupposes an infinite sequence of formulas, but it does *not* prohibits each formula in the sequence from being the same. In the case of the Liar, $L = \neg Tr(\ulcorner L \urcorner)$, the infinite sequence of formulas would be formed by L itself. It is thus easy to see that $ref_I(a_1, I(\ulcorner a_2 \urcorner)), ref_I(a_2, I(\ulcorner a_3 \urcorner)), ..., ref_I(a_m, I(\ulcorner a_{m+1} \urcorner)), ...$, in which all of the formulas $a_1, a_2, a_3, ..., a_m, a_{m+1}$ are L. This sequence is non-well-founded. This indicates that non-well-foundedness is perhaps a more general property than self-reference and circularity. There are self-referential sentences which are *also* non-well-founded, but there are ill-founded sequences which are neither self-referential nor circular, such as the Simplified Yablo Sequence.

In this section, little was concluded and much was hypothesized. This shows that the work to be done is still abundant: I must investigate the concept of ill-foundedness for sentences (instead of sets); decide whether the Yablo Paradox

is indeed non-well-founded and, lastly, explore the relationship between non-well-foundedness, self-reference and circularity. These remain as open problems to be tackled in the future.

3.5 Considering Objections

3.5.1 The EC Objection and Hyperintensionality

In [27], Leitgeb discards both the definition of circularity as a fixed point and $selfref_1$, which inspired my definition of reference. He rejects the former because it leads to triviality, and the latter because it does not satisfy the Equivalence Condition ([27], p. 7) shown below.

Equivalence Condition (EC): if A is self-referential/circular and B is logically equivalent to A, then B is also self-referential/circular.

Numerous examples suggest the failure of this condition. Take $A =$ "$(Tr(\ulcorner b \urcorner) \vee \neg Tr(\ulcorner b \urcorner)) \vee \neg A$". "$A$" is clearly self-referential, and $selfref_1(A)$ indeed holds. However, the logically equivalent $C =$ "$Tr(\ulcorner b \urcorner) \vee \neg Tr(\ulcorner b \urcorner)$" is not $selfref_1$.[22] Therefore, sentences with logically equivalent extensions can receive different diagnoses as to whether or not they are self-referential or circular. This is as true for Leitgeb's definition as it is for my own definitions, so this is also an objection to my definitions $selfref_I$ and the general $selfref$. While one could think of solving the problem by adding to the definition of $selfref_1$ a clause that states that logically equivalent sentences to self-referential ones are self-referential too, this would not solve it, for this caveat would end up being too liberal – it may cause every sentence to be self-referential, just as in C_1.[23]

However, before trying to solve the issue, we should ask ourselves: is it *really* a problem that our definition of circularity does not satisfy EC? I will argue that it is *not*, and that, rather than taking the failure of EC to mean the failure of our definitions of self-reference and circularity, we should take it to mean that those concepts require a more fine-grained investigation than the

[22] "A" and "C" are logically equivalent because any disjunction that has a tautology as one of its disjuncts is equivalent to the tautology. Two sentences are said to be logically equivalent if they receive the same truth values in all possible models. My definitions also fail the EC condition when it is formulated with the more broad *necessary* equivalence, which states that two sentences are necessarily equivalent when they receive the same truth value in all possible worlds.

[23] For further explanation, see [27], p. 3.

extensional one: they are *hyperintensional* concepts. Here, I am using the concept of hyperintensionality found in [8], which is that a hyperintensional concept is one that draws a distinction between logically and/or necessarily[24] equivalent contents.

Indeed, for what other reason would anyone believe that C = "$Tr(\ulcorner b \urcorner) \vee \neg Tr(\ulcorner b \urcorner)$" is self-referential, if not for the prior consideration that it is equivalent to A = "$(Tr(\ulcorner b \urcorner) \vee \neg Tr(\ulcorner b \urcorner)) \vee \neg A$"? Independently of the conception of self-reference one might have in mind, C = "$Tr(\ulcorner b \urcorner) \vee \neg Tr(\ulcorner b \urcorner)$" never seems self-referential *in its own right*, and that is quite a strong reason to believe that those concepts escape logical equivalence. And that conclusion should be traced not only for self-reference and circularity, but for *reference itself*. The same example showing the failure of EC for self-reference shows it fails for reference simpliciter: "A" refers to the term "A," but the logically equivalent "C" does not refer to "A". Once more, from this, it should not follow that our definition of reference is ill-formed, but that reference draws important distinctions between logically equivalent sentences – hence its hyperintensionality.

The view that reference is hyperintensional has also been sustained by Lavinia Picollo in [34]. In it, Picollo expands Leitgeb's notion of reference to other types of quantified sentences, not only conditional ones, to provide a more thorough treatment of the notion of reference. In this paper, I did something similar, but in an even more general setting: as her paper's title suggests, Picollo defines only *alethic* reference – *i.e.* sentences that predicate truth of their referents. Here, I am interested in reference in general, so it's not possible to simply adopt her definition. Rather, I articulated a hyperintensional account of reference in general, one which allows logically equivalent sentences to refer to different things.

3.5.2 Is Reference Arbitrary?

Another possible objection to my approach is the *arbitrariness concern*: as is known, a language can have multiple interpretations, each in complete disagreement with another about the meaning of the language's terms. If the referent of a term is an element attributed to it in I, how can one select the interpretation that provides its *"correct"* referent? Doesn't reference become an arbitrary matter on this picture?

[24] I will not address the quarrel on the difference between logical and necessary equivalence and whether one can be reduced to the other. The important point is basically that my definitions escape both kinds, and for this reason are said to be hyperintensional.

My response to this concern is that reference is *relative*, but not necessarily *arbitrary*. It is relative because it may change upon model-theoretic interpretation, and so there is no *absolute* reference. Depending on the choice, the meaning of the term changes, and so its referent changes, but that is perfectly normal. An example might make it clearer: to assign the value *true* to the sentence "2 + 2 = 4", written in the language of Peano Arithmetic, one has to choose a certain interpretation; namely, that which assigns the expected meaning to the terms "2" and "4" – *i.e.*, what is called the *standard interpretation* $\mathbb{N}$. Of course there are other possible interpretations: there can be an I in which the term '2' denotes the number 4 and the term '4' denotes the number 2. Hence, the truth value of the sentence changes. While this entails that reference is relative, it does not entail that reference is arbitrary, for we may be able to find a method of *interpretation-ranking* – that is, there might be some interpretations which are better, or more correct, than others. If there is anything like the "right" interpretation to some language, or at least any method to evaluate the level of correctness of interpretations, then reference is certainly *not arbitrary*.

I will not argue that there is indeed a method for choosing the right interpretation for a language. My point in the former paragraph was that, *if there is one*, reference is not arbitrary. But suppose we find out that there is no such method; then, reference is indeed arbitrary. Even in this case, however, arbitrariness would not be a problem: as the arithmetical example suggested, it is to be expected that the reference of a sentence depends on the interpretation attributed to each of its terms. In fact, this arbitrariness makes my approach more general than others, such as Picollo's in [34]. She defines reference by mention as follows:

(Picollo's m-reference) Let φ and ψ be sentences. φ *m-refers* to ψ iff φ contains a subsentence of the form $\mathrm{T}t$ and $\mathbb{N} \models t = \ulcorner\varphi\urcorner$.

By *m-refers*, Picollo means "refers by mention." A subsentence is a sentence contained inside another sentence; a subsentence of the form $\mathrm{T}t$ is a sentence that contains the truth predicate followed by a term t, which is a name of a sentence. $\mathbb{N} \models t = \varphi$ means that the term t is φ in the standard model of arithmetic. What I want to highlight is that Picollo fixes the standard $\mathbb{N}$ as the only interpretation of arithmetic in which reference can be considered, so m-reference will be evaluated in this specific environment. This definition is, therefore, limited in that it does not allow us to evaluate reference using non-standard interpretations, which we know exist, at least in the case of Peano Arithmetic. ref_I, on the other hand,

enables us to evaluate what a formula refers to in any interpretation we choose to work with – be it the standard one or others, maybe non-classical ones – for it does not fix any specific interpretation as setting. This is particularly important to the study of self-referential paradoxes, that must be formulated in non-classical languages and, thus, be interpreted by non-standard structures.

Therefore, if there is a method to decide for a correct interpretation for a setting, then reference is relative, but not arbitrary. If there is not, then reference is arbitrary, but this is to be expected. Moreover, in this case, my definition is general and applies to many different models.

3.5.3 Reference and a General Russell-type Structure

While some non-well-founded relations, such as the successor relation, are unproblematic, others cause paradoxes, such as the membership relation with respect to the Russell set. The reason for such difference lies on some aspects of the theory, rather than in the relations themselves. Actually, there could certainly be a Russell type of paradox involving not membership, but the succession relation. Just imagine that there is a number n that is the successor of all numbers which are not successors of themselves. Is n successor of itself? If not, then n is one of the numbers which are not successors of themselves. Thus, by definition, n is a successor of n. We have a contradiction. On the other hand, if we suppose that n is a successor of itself, then n must figure among those numbers which n is a successor of – namely, numbers that are not successors of themselves. So, n is not a successor of itself. Contradiction.

Unlike Russell's, the successor argument is merely a pseudo-paradox – or, in Quine's nomenclature, a falsidical paradox[25] – not a true one. This is because we defined n as the successor of all numbers which are not successors of themselves; but, as no number is a successor of itself, n would be the successor of all numbers, and, of course, there can be no number which is the successor of all numbers. While the unrestricted Comprehension Schema allows us to define Russell's set, we are defining *no* set when we supposedly defined n: our definition is completely empty.

What prevents the successor relation from generating paradox is not a syntactical feature; it is not the syntax of the theory that makes it impossible to formulate a Russell-type of structure, but a property intrinsic to the theory in question. Actually, literally any binary relation is capable of generating a

[25]Refer to section 1.4 for an explanation of this term.

Russell-like structure with the following schema:

A General Russell-type Structure for any Binary Relation

Let R be any binary relation, and define $n : \forall m(nRm \leftrightarrow \neg(mRm))$

Proposition. The relation R is inconsistent with respect to n.

Proof. Is it the case that nRn? If not, then, by the instance $\neg(nRn) \rightarrow nRn$ of the schema, it can be concluded that nRn, which is a contradiction. If yes, then by the instance $nRn \rightarrow \neg(nRn)$ of the schema, it can be concluded that $\neg(nRn)$, which is a contradiction.

■

This structure shows that any binary relation is subject to paradox if there is no "external feature" to prevent it, the external feature here being a theoretical element that forbids us from defining the necessary n that will lead to paradox. Since this whole chapter is devoted to the definition of the binary relation of reference and other concepts that come from it (self-reference and circularity), it is now imperative to put this very relation to test by subjecting it to the Russell-type structure and checking whether it is paradoxical or not. I would, of course, be in serious trouble if it were indeed paradoxical: I want my definitions to be able to diagnose paradoxical sentences, but not to be paradoxical themselves. Creating a sentence that would both refer and not refer to something would render my definition of reference fruitless. Happily, as I will show now, there are theoretical features that prevent the relation of reference from falling into a Russell type of paradox.

Substituting ref_I for R in the general Russell-type structure, we have the following formula:

Define $n : \forall m(ref_I(n, m) \leftrightarrow \neg ref_I(m, m))$

The first reason why this does not amount to paradox is because ref_I relates two objects of different natures: while n is a sentence of the language $\mathfrak{L}$, m is an object of the domain of some interpretation I. This is already enough to prevent a Russell-type paradox. However, one could argue that we can use the interpretation-dependent definition of reference to create a general, cross-interpretation one; maybe this general definition would be subject to paradox, even if ref_I is not. The general definition could be written along the following lines:

Let $\mathfrak{L}$ be a first-order language. A formula φ *refers* to a if and only if $ref_I(\varphi, I(\ulcorner a \urcorner))$ for all interpretations I of $\mathfrak{L}$; i.e. there is a closed term t in φ such that $I(t) = I(\ulcorner a \urcorner)$ for all interpretations I of $\mathfrak{L}$. I write $ref(\varphi, a)$ when φ refers to a.

The formulation of a Russell type of paradox makes more sense using this definition, for it would ultimately relate the interpretation of two items, $I(t)$ and $I(\ulcorner a \urcorner)$, which are objects "of the same nature." Moreover, the paradox would not depend on one interpretation in particular. Now, it is possible to talk about a sentence n such that $\forall m(ref_I(n, m) \leftrightarrow \neg ref_I(m, m))$. Does this create a paradox regarding the general definition of reference? The first point to take into account is that the universal quantifier in the General Structure sentence is a higher order quantifier, for it quantifies over *formulas*, not individuals. Thus, as my definition of reference applies to first-order languages (and interpretations of first-order languages), the sentence would have to be formulated in a second-order metalanguage, not in the language itself that is being evaluated. This does not, however, prevent the paradox from happening yet.

What really prevents the paradox from happening is not any feature related to the order of the language, but the fact that "the formula of all formulas that do not refer to themselves" is *not* a formula at all. We assume that there might be, in the language $\mathfrak{L}$, some formulas that refer to themselves and others that do not refer to themselves. With very few components (usually available to any first-order language), the language is capable of generating infinite non-self-referential formulas. Actually, a language equipped only with the negation symbol is already capable of generating infinite formulas of the type $p, \neg p, \neg\neg p, \neg\neg\neg p$, and so on. The problem is that, to refer to an infinite number of sentences, n would have to be an infinite formula and, since formulas are formed by recursive rules, no well-formed formula is infinite. Therefore, just as "the successor of all numbers that are not successors of themselves" defines *no* number, the description of n defines no formula whatsoever.

The failure of definition just presented reveals an interesting characteristic of the concept of reference, which I allow myself the poetic license to call its *craftsmanship* character. Reference, as I defined it, happens by mentioning the tokens of each object that is being referred to, one at a time. To refer to the number one, I must mention the numeral 1 directly in the formula I aim to construct. In this sense, referring to things is a craftsman's job: references are knit one by one, and together they form a structured woven. It does not happen

by mentioning many things at once with the help of tools such as quantifiers. By saying, for example, "all men are mortals," I am not referring to each and every man that exists or has ever existed. To refer to all men, I would have to mention each one individually. In the next section, I will argue *why* reference by quantification should not be allowed in an accurate definition of reference.

Since the craft character of my definition of reference is what prevented it from falling into paradoxes of a Russell type, views that allow for reference to happen by quantification must be careful not to fall into paradox. In [27], Leitgeb suggests that a sentence such as "$\forall x(A[x] \rightarrow B[x])$" could be defined to refer to all and only A's. Picollo, in [34], dives deeper in this suggestion and designs a sophisticated definition of reference by quantification (q-reference) in which quantified sentences can refer to the individuals that are under the range of the quantifiers. Under those views, it might be possible to write a formula that refers to all and only the formulas which do not refer to themselves and, if this happens, their notion of reference will be paradoxical.

3.5.4 Reference by Quantification

From the definitions created in this chapter and from the commentaries in the last section, the reader might be wondering why there is no reference by quantification. "It is perfectly reasonable to think that the sentence "all men are mortals" refers to all men," she might object, and judge that my definition of reference is insufficient to portray one of its most important forms.

In my defense, I could try arguing that, as exhibited in 3.5.3, reference by quantification might permit the invasion of a Russell type of paradox, creating a sentence that both refers and does not refer to all non-self-referential sentences and thus rendering the very definition of reference paradoxical. But this would be an ad-hoc move: prohibiting reference by quantification solely because it might lead to paradox is not a strong enough reason to avoid it, specially when we could work with non-classical logics to model this concept. But there is another deeper, stronger, conceptual reason to prohibit reference by quantification: it fails at understanding what reference really is.

To refer to something, one must be able to *singularize* an object – be it empirical or abstract. The role reference plays in discourse is more similar to that of *pointing* than to that of quantifying: it is enabled through tokens of what we aim to refer. How could I refer to all prime numbers, when saying "all prime numbers are divisible only by themselves or by 1," if there is an infinite quantity

of them, and so I can never know them all? When I utter the sentence above, I am establishing a property that applies to a certain set of number, rather than referring to each and every member of such set. A similar problem would be present when uttering the sentence "all men are mortals." By saying this, would I be referring to all *living* men, all men that lived in any historical period, or to each men that has ever lived and will ever live in the future? There doesn't seem to be any reason to prefer one of the options above over the others. This indicates that, by uttering the sentence, I am not referring to any man whatsoever, but merely stating that a property applies to individuals with the determinate quality of being men – those are two quite different things.

This is the reason my account does not allow for reference by quantification: I believe this alternative misses the point of what reference really is. Another important downside of this view is that it might lead to paradoxes, such as the one I constructed in the previous section.

3.6 Consequences of the Results

The first consequence of the presented results is that, as was demonstrated in section 2.4.3, Yablo's Paradox is neither circular nor self-referential, contrary to what is argued by Graham Priest in [36], JC Beall in [4] and others (for example, Cook in [11]). A philosophical consequence that can be drawn from this fact is that *self-reference and circularity cannot be the causes of semantic paradoxes.* There is a widely held belief that the paradoxical nature – the paradoxicality – of semantic paradoxes lies in self-reference and/or circularity, as described by Roy Cook: "The Yablo paradox threatens to overthrow the rather entrenched idea that paradoxes (or, at least, the semantic and set-theoretic paradoxes) are intimately and ineliminably tied to self-reference or circularity of some sort" ([11], p. 2). If my definitions are accepted, Yablo's Paradox does indeed overthrow this entrenched idea, for it shows that there is at least one semantic[26] paradox which does not derive from self-reference or circularity.

Two questions that arise immediately from this conclusion are the following: if not self-reference or circularity, what is the actual root of Yablo's Paradox? Moreover, if semantic paradoxes are not intimately and ineliminably tied to

[26]One might ask why Yablo's Paradox is "semantic," or, better, what are semantic paradoxes after all. Semantic paradoxes can be loosely defined as those that are built upon semantic concepts, such as the concepts of truth or satisfaction. Coming from the Liar family, Yablo's Paradox is clearly semantic, for it too deals with the concept of truth, just as the Liar.

self-reference and circularity, is there some other concept common to all of them, that renders them paradoxical? A hint to address the first question was developed in 3.4.5, where I considered whether Yablo's Paradox can be regarded as being non-well-founded. But this, for now, is only a hypothesis that must be further investigated. This and the second questions will be left for further endeavors, finishing this chapter with a somewhat aporetic tone.

Another consequence of the results given in the chapter is that Russell's Paradox is non-well-founded. This, in itself, is an accepted result: what is different here is that I draw a distinction between non-well-foundedness and circularity; a distinction which arises immediately from the definitions of each of these concepts.

A final possible consequence of this chapter is that the definitions that figure here may create room for a more complete study of the phenomena of reference, self-reference and circularity in formal languages, under a different conception than the standard fixed-point account. We can envision the building, for example, of a formal language containing a predicate of self-reference such as the one we defined here: a unary predicate that applies to sentences that are self-referential. This would make it possible for this language to evaluate its own self-referential character, and eventually create sentences such as "this sentence is self-referential."

Since my definition of self-reference hangs on my definition of reference, and my definition of reference depends on model-theoretic interpretations, such a language would have to be capable of producing sentences that talk about *its own interpretations.* That is because, to evaluate some sentence φ as self-referential, the very language itself would have to recognize that, for all its interpretations, $ref_I(\varphi, I(\ulcorner\varphi\urcorner))$. This means that, in some way, this language would have *access* to all its possible metalanguages. I hope this does not sound too esoteric – it is indeed difficult to imagine a language *so* broad that contains, or can at least express things about, all of its structures. However difficult this is to imagine, I do believe it is worth a try, for it would enable us to achieve a whole new level of semantic closure: a language capable of expressing not only facts about their own sentences, but facts about the whole of their semantics.

3.7 Conclusion

My central aim in this chapter was to formally define the notions of self-reference and circularity. To do that, I had to first provide a formal definition of reference: I

chose to define it through the model-theoretic notion of interpretation, anchoring this choice on the philosophical position that reference is a relation between sentences and objects in the domain of an interpretation I. Moreover, I provided two possible definitions of self-reference and circularity: an interpretation-dependent one and a general one. I also defended that those three notions – reference, self-reference and circularity – are hyperintensional, since they escape logical equivalence and should not be expected to preserve it. Next, I proved that the Yablo Paradox is neither circular nor self-referential and, finally, I explained why Russell's Paradox does not fall under my definitions: it is because it is non-well-founded, rather than circular or self-referential. Hence, under my definitions, the Liar Paradox is self-referential, the Liar Cycle is circular, Russell's Paradox is non-well-founded and Yablo's Paradox is neither self-referential nor circular.

Chapter 4

Semantic Closure and Everyday Language

4.1 What is Semantic Closure and Why it is Interesting

The concept of semantically closed languages makes its first appearance in Alfred Tarski's theory of truth; notably, in the first section of the paper [51]. In general, semantically closed languages are those capable of expressing semantic facts about themselves – *i.e.*, those that contain semantic predicates that apply to the names of their own sentences. In the canonical case, languages contain their own truth predicates, and are thus enabled to decide whether their sentences are true or false. It is exactly this gain of expressive power that causes their evasion from classical logic: they end up being capable of expressing the Liar paradox and other semantic paradoxes that make them either inconsistent[1] or gappy. Shortly, I will define the concept more precisely, but to get an intuitive grasp of it, it suffices to say that semantically closed languages (or theories) are those that can **(i)** name their own sentences and **(ii)** apply to them semantic predicates, like the truth or the satisfaction predicate.

Since semantically closed languages do not behave according to the rules of classical logic, Tarski used this as a criterion to exclude those languages from his theory's scope: if a language is semantically closed, we cannot work with it, and

[1] Of course, as the reader might be wondering, this is a non-standard use of the word *inconsistency*: as is well known, it usually applies to theories, to designate a theory in which A and $\neg A$ can be derived as theorems. Here, I understand an inconsistent language as one in which both A and $\neg A$ are *true*. This denomination follows Tarski [51]. For a more thorough explanation of the term, see [50], p. 52.

we can barely say anything about it. This attitude is unsurprising given that, at the time when Tarski wrote [51], a system discovered to be inconsistent was taken to be a complete failure. Fortunately, times have changed since then, and inconsistency (and incompleteness) are not considered as logical impossibilities if they are not accompanied by triviality. Therefore, we know that those languages (and their background logics) do not obey all rules of classical logic, but this shouldn't be a reason to exclude them from our field of studies, like Tarski did. If non-classical logics figure in our theoretical framework, there is no reason to avoid the concept of semantically closed languages, nor to state categorically that there can be no formal definition of truth for such languages. Indeed, Tarski's Indefinability Theorem holds only for classical valuations, but it says nothing about defining truth in a non-classical setting.

The reasons that make the concept of semantically closed languages interesting are many. For one thing, the possibility of augmenting the expressive power of a language is valuable *per se*: logicians have for long had the ambition of building a universal formal language, one in which everything expressible in any setting can be expressed. Of course, building such a language is a gigantic ambition which falls far out of the scope of this text (and which I'm not even totally convinced can be achieved). However, constructing semantically closed languages – and more than that: constructing a tool to build them (see section 4.6) – is a step towards realizing this ambition.

In addition, this topic allows for a relation between semantically closed languages and everyday language.[2] Until the late 90's, there was an active debate over whether everyday languages are semantically closed or not – see, for example, [19], [28] or [35]. At first glance, they certainly seems to be so: English contains quotation marks (to quote its own sentences) and a truth predicate that supposedly ranges over its sentences. However, if everyday languages are semantically closed, they can produce self-referential sentences which are both

[2] I use the terms "everyday languages," "colloquial" or "non-formal languages" interchangeably: they all denote the languages we, in our everyday life, speak. English, Portuguese, Spanish are all everyday-colloquial-non-formal languages. They are all in opposition to formal languages such as the language of Peano Arithmetic. Although these terms may sound somewhat pedestrian, I consider them to be more accurate than the broadly used "natural language." Actually, I used to employ this term, until it was brought to my attention that the word "natural" indicates a feature with which one is born, and no one is born as an English speaker. Advocates of the term might say that, although we aren't born as English speakers, we are born with the capacity to learn the language. However, when I say "natural language" I aim to denote the language itself, not my natural capacity to learn it, and the language itself is acquired with one's immersion in culture, not naturally. Thus, in lack of better terms, I use the denominations "everyday," "colloquial" or "non-formal" languages.

true and false – thus, the languages would require a non-classical semantics. Because of this, some classical logicians such as Herzberger in [19] have rejected this claim and found ways to defend that everyday languages are, in fact, *not* semantically closed.[3] Even though this debate has lost traction after the 2000's, it is still an important and unsettled matter.

Tarski was a proponent of the thesis that colloquial languages are semantically closed. They must be so because, according to the logician, they are *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" ([51], p. 164). If universality truly is the defining feature of colloquial languages, then they *must* be semantically closed, and thus contain paradoxes. He continues:

> If we are to maintain this universality of everyday language in connection with semantical investigations, we must, to be consistent, admit into the language, in addition to its sentences and other expressions, also the names of these sentences and expressions, and sentences containing these names, as well as such semantical expressions as 'true sentence', 'name', 'denote', etc (Idem, ibidem).

If everyday languages were found to be semantically closed, then formalizing semantically closed languages would help in their formalization, since some issues that may appear within the construction could also appear in an attempt to formalize colloquial language. Formalizing everyday language has for long been an objective of many authors in the domain of philosophy of language, so researching semantically closed languages would be quite useful for that. In the last part of this chapter, I will argue for the theses that languages such as Portuguese or English are, indeed, semantically closed. This is not, however, a settled matter, and there are many authors who disagree with this thesis.

In what follows, I'll first show the formal definition of semantically closed languages and theories, and then the construction of a first-order bisorted semantically closed language. My definition is not *exactly* the one found in [51]: I discovered a technical ambiguity in Tarski's text that is solved here. I will address

[3] His defense feels a little question-begging, specially if analyzed through our modern gaze: he proves that if everyday languages are semantically closed, they are capable of producing paradoxes. Since this is a logical impossibility, we must conclude that they are not semantically closed. Again, this argument comes from a time when non-classical logics were not yet popularized, and seeing inconsistency as logical impossibility was a common attitude.

these questions more thoroughly in the next section. Afterwards, I will show the construction of a semantic closure operator, a monotonic operator that *closes* theories semantically. A distinction has to be made that can prevent us from slipping into conceptual ambiguities: many times, the concept of semantically closed languages is addressed as "semantic closure," as can be seen, for example, from the title of Graham Priest's paper.[4]. The ambiguity generated is that "closure," in logic, is usually seen as a monotonic operator, not as a property of something – the deductive closure, for example, is an operator whose function is to select the set of all deducible sentences of a theory. So it is quite strange to employ the term "semantic closure" as a property of semantically closed languages, since, while the first is an operator that performs an action, the second is simply a characteristic possessed by some languages. To avoid this problem, this text treats them as two different (although related) concepts: semantically closed languages are a type of language with certain characteristics, and semantic closure is a *logic operator*, that acts over theories to close them semantically.

4.2 A Third Condition for Inconsistency?

When I described semantically closed languages (theories) as those that can **(i)** name their own sentences and **(ii)** apply to them semantic predicates, I left out a third condition required by Tarski in his 1936 paper, which he labels "empirical condition." Tarski requires the condition to ensure that semantically closed languages (theories) are indeed capable of producing paradoxes. The first reason I left this condition out of my definition is because I do not want to capture inconsistency necessarily, but the ability of a language (or theory) to produce sentences making claims about themselves or other sentences of the language (theory). However, as I will show below, this condition is also unnecessary for inconsistency to arise. Let me present it now:

(iii) Semantically closed languages must contain a formula φ such that $\ulcorner\varphi\urcorner$ is the name of $\neg Tr(\ulcorner\varphi\urcorner)$.[5]

Naturally, Tr refers to the truth predicate, and indeed Tarski requires (iii) only in defining inconsistent semantically closed languages *with respect to the truth*

[4] In [35] Here, Priest uses the term "semantic closure" to address the property of some languages of being semantically closed. So these two terms refer to the same object.

[5] [35], p. 119 – notation adapted to match mine. This is Priest's formulation, not exactly Tarski's. This is not a problem, since the idea expressed by Priest is the same as the one expressed by Tarski – he simply puts it in a clearer, more formal, manner.

predicate, and not with respect to the satisfaction or other semantic predicates. As I mentioned earlier, Tarski postulates this condition to determine exactly what a language should have in order for it to express the Liar Paradox, and, in this paper, he seems to consider (iii) necessary to do so. The logician assumes it to be necessary because the contradiction generated by the Liar can only arise when the language in which the paradox is constructed is able to address the code name of the sentence φ – *i.e.*, $\ulcorner\varphi\urcorner$ – to the sentence itself, which states $\neg Tr(\ulcorner\varphi\urcorner)$. The worry was that, if the language was not able to "recognize" $\ulcorner\varphi\urcorner$ as the name of $\neg Tr(\ulcorner\varphi\urcorner)$, the latter wouldn't be expressing that it itself is not true, and no contradiction would be derived. This "empirical" factor is better seen with a different formulation of the Liar:

(BL) The sentence in bold in section 4.2 of this text is not true.

Now, it's easy to see how this version of the Liar depends on an "empirical" factor: if BL (Bolded Liar) was not written in bold characters, or if there was another sentence in section 4.2 written in bold, it would not lead to paradox. What makes it paradoxical is that it is written in bold and there are no other sentences of this kind in the section, and so we can conclude that it is referring to itself and claiming it is not true. For the sake of rigor, the word *empirical* should be kept under quotation marks, since what the condition expresses is not exactly empirical, as the word is commonly understood. After all, BL is paradoxical not because the observer can *recognize* that it is the only sentence in bold in section 4.2, but because *it really is* the only sentence in bold on section 4.2. The paradoxicality cannot be debited to any epistemic factors such as the recognition of an observer. But it's true that, to derive the contradictions that render the Liar paradoxical, the language it's written in must possess the ability to check if the name matches the sentence named. Only in this sense the condition can be said to be "empirical."

However, it just so happens, as Tarski will notice in [52],[6] that this condition can in fact be waived, for it's possible to construct versions of the Liar which do not depend on name matching – *i.e.* with no terms such as "this sentence" or any naming devices. In [29], David Miller exposes one of the possible versions.

A: Every sentence has the property F.

Let A^* be the *transformed* of A, defined as

[6] See [52] p. 348.

A^*: "Every sentence has the property F" has the property F.

Now let B be the following sentence:

B: Every sentence has a false transformed.

If B^* is the transformed of B, then B^* reads:

B^*: "Every sentence has a false transformed" has a false transformed.

In this version, B^* states that it itself is false, because it is precisely the transformed sentence of B. Using the "transformed sentence" trick, we can create a Liar that doesn't mention its own name, and so (iii) is not needed for inconsistency to arise – the two first conditions of the semantic closure definition are enough to render a language or theory inconsistent.

4.3 Semantically Closed Languages and Theories

Before I lay down the definitions of semantically closed languages and theories, a caveat has to be made: the definitions are formulated in respect only to the truth predicate, which means that semantically closed languages or theories are, here, the ones that contain their own truth predicates (and names for their sentences, as required by criterion (i) of our loose definition). But this is not the only possibility: we can apply it to other semantic predicates, like the satisfaction predicate. My choice of formulating it to the truth predicate is due to certain factors, among which are the possibility of deriving the Liar within most semantically closed languages, not any satisfatibility paradox. Moreover, the truth predicate, rather than the satisfaction predicate, allows us to relate semantically closed languages to non-formal languages.[7]

It's necessary to make still another scope restriction: the definition of semantically closed languages applies only to *interpreted languages*, *i.e.* those languages whose sentences can be endowed with meaning. Only in those languages it is interesting to talk about "true sentences" and "semantic closure." Indeed, the issue becomes meaningless if we treat languages under a purely syntactic facet. *Structures* are the tools that will do the job of interpreting the language. To craft the concept of semantically closed languages, it will be necessary to rely on a truth evaluation residing in a certain specific structure that will "supervise" the

[7]However, it's not hard to adapt the definition given here to the satisfaction predicate, and it ends up being quite similar to the one for truth. About this, see [35], p. 118.

application of the truth predicate contained in the language itself. I will call it *standard structure.* For the sake of generality, I will leave the concept of structure undefined for the time being; a definition of it can be found in any model theory text book and in section 4.4 of this text. The reason why we need to count on a structure will become clear once I spell out the definition of semantically closed language, so let's jump right into it.

4.3.1 Definition: Semantically Closed Languages

Definition. A *semantically closed language* is a language $\mathfrak{L}$ that fulfills the following conditions:

(i) *Naming condition.* For every sentence λ of $\mathfrak{L}$, there must be in $\mathfrak{L}$ a name for λ, denoted by $\ulcorner\lambda\urcorner$. From a formal point of view, the condition is satisfied by introducing into the alphabet of $\mathfrak{L}$ a unary function $\ulcorner\ \urcorner$: $E_{\mathfrak{L}} \rightarrow E_{\mathfrak{L}}$, which acts on the expressions of $\mathfrak{L}$. The $\ulcorner\ \urcorner$ function selects each λ sentence of $\mathfrak{L}$ and returns the term $\ulcorner\lambda\urcorner$, name of λ.

(ii) *Truth predicate.* There must exist in $\mathfrak{L}$ a predicate Tr, such that $Tr(x)$ is a sentence of $\mathfrak{L}$ if and only if there exists a sentence λ of $\mathfrak{L}$ such that x is the name of λ, that is, x is $\ulcorner\lambda\urcorner$. Let $\mathfrak{A}$ be the standard structure for $\mathfrak{L}$. The truth predicate must behave in such a way that all instances of the schema "$Tr(\ulcorner\lambda\urcorner) \leftrightarrow \lambda$" are true in $\mathfrak{A}$, i.e.:

$$\mathfrak{A}(Tr(\ulcorner\lambda\urcorner) \leftrightarrow \lambda) = T$$

Note that condition (ii) refers to the structure $\mathfrak{A}$ of $\mathfrak{L}$. In light of these conditions, the necessity of a standard structure becomes clear: firstly, we do not want to say simply that "$\mathfrak{L}(\mathfrak{A})$ contains its own truth predicate." We also want this predicate to *behave well*, which means that it must respect *Tarski's biconditionals.* These biconditionals are all instances of the T-schema, the foundation of Tarski's conception of truth (See section 1.2 of this text). In short, it states that if a sentence can be regarded as true, then it is the case and, if a sentence is the case, it can be regarded as true.[8] "Respecting Tarski's biconditionals" means that all

[8]Tarski inherited his theory's spirit from the Aristotelian notion of truth, expressed in the Γ book of [3]. Aristotle says that "false is to say that being is not or that non-being is; true is to say that being is and non-being is not" ([3], p. 179). According to Tarski, this description expresses the key point of the concept: uttering truths is nothing more than discoursing in a way that my discourse is verified in the world; to say, of the things that are, that they are, and of the things that are not, that are not, and not to say the opposite. This is expressed in the idea of correspondence: the thing said must correspond to the state of things in the world to be true.

instances of the T-schema must be *true*, and here's the core of the issue: where is this second truth evaluation located? It cannot be the same truth predicate that evaluates sentences of the language to evaluate its *own* applications. Thus, there must be a higher order valuation that guarantees the good behavior of the truth predicate of $\mathfrak{L}$, and this higher order is exactly the standard structure $\mathfrak{A}$ of $\mathfrak{L}$. That is expressed by saying that all instances of the T-schema must be true in $\mathfrak{A}$.

This may sound a bit unnecessary for some: why should we go through the trouble of creating a structure just to evaluate the instances of the T-schema with the truth predicate Tr of $\mathfrak{L}$? Couldn't we simply work directly with theories, where we wouldn't have all this trouble? The answer is yes, we could, and the reader will witness how it is much easier to define semantically closed theories than languages. The reasons I decided to also work with languages are two: firstly, as far as I could tell, this difficulty had not yet been regarded by the literature on this topic. Graham Priest's definition of semantically closed languages, for example, states that all sentences of the form $\ulcorner Sat(t\ a_\phi) \Leftrightarrow \phi(t) \urcorner$ must be *true*, but doesn't explain where is this truth evaluation located (in the language itself? In its structure?).[9] Tarski too does not address this particular problem when he provides the definition of semantically closed languages.[10] Secondly, this raises an important philosophical (and logical) issue regarding the semantic closure of colloquial languages: it has been said that semantically closed languages are, first of all, *interpreted* languages. It would make no sense to talk about a semantically closed language where no meaning is attributed to its sentences. So, if we regard colloquial languages as being semantically closed, they too must be interpreted languages. But how could a structure for them be built, if they are, as Tarski stated, the sole universal languages? Could they have their own interpretation embedded within them? This problem will be further dealt with in this essay's last section.

Let's now move to the definition of semantically closed theories, for it will lead the way to the construction of the semantic closure operator (section 4.6). Before the operator, however, I will provide a detailed construction of a first-order bisorted semantically closed language.

[9] [35], p. 120. Note that Priest's definition regards the satisfaction predicate instead of the truth predicate. However, the same problem emerges.

[10] See [51], p. 32. Of course, the point can be made that Tarski's definition is informal and does not aim to solve all formal issues that could arise, but rather only provide an idea of inconsistent self-referential languages. Still, if we use his as a model definition, the problem persists if we're to dive deeper into the issue.

4.3.2 Definition: Semantically Closed Theories

Let's begin with the formal definition of a theory.

Definition. A *theory* $\mathfrak{T}$ is a set of sentences of a given language $\mathfrak{L}$ such that every logical consequence of $\mathfrak{T}$ belongs to $\mathfrak{T}$.

Definition. Let $\mathfrak{T}$ be a theory and $\mathfrak{L}$ its language. $\mathfrak{T}$ is called *semantically closed* if it meets the following conditions:

(i) *Naming condition.* For every sentence λ of $\mathfrak{T}$, there must be in $\mathfrak{L}$ a name for λ, denoted by $\ulcorner\lambda\urcorner$. From a formal point of view, the condition is satisfied by introducing into the alphabet of $\mathfrak{T}$ a unary function $\ulcorner\ \urcorner$: $E_{\mathfrak{T}} \rightarrow E_{\mathfrak{T}}$, which acts on the expressions of $\mathfrak{T}$. The $\ulcorner\ \urcorner$ function selects each λ sentence of $\mathfrak{L}$ and returns the term $\ulcorner\lambda\urcorner$, name of λ.

(ii) *Truth predicate.* There must exist in $\mathfrak{T}$ a predicate Tr, such that $T(x)$ is a sentence of $\mathfrak{T}$ if and only if there exists a sentence λ of $\mathfrak{T}$ such that x is the name of λ, that is, x is $\ulcorner\lambda\urcorner$. The Tr predicate must behave in such a way that all instances of the schema

$$Tr(\ulcorner\lambda\urcorner) \leftrightarrow \lambda$$

are theorems of $\mathfrak{T}$.

The reason why this definition is simpler than that of semantically closed languages becomes clear with the second condition: when working with theories, we have at our disposal the notion of *theorem*, and thus it's possible to state that all instances of the T-schema must be theorems of $\mathfrak{T}$. In languages, the concept of truth is somewhat analogous to theoremhood, but we couldn't use the same truth predicate of the language $\mathfrak{L}$ to state that the instances of the T-schema are true in $\mathfrak{L}$. So we had to create another instance to evaluate that of the language we're working on. This is, as the reader might have noticed, the only difference between the definition of semantically closed languages and theories – the other clauses are exactly the same.

4.4 A First-Order Bisorted Semantically Closed Language

This section is dedicated to the construction of a first-order bisorted semantically closed language. I hope the construction can disclose interesting characteristics

of this type of language. The first is that the language must contain a type distinction, because there are regular terms and terms that are names of formulas. So, if a function is to range over those two kinds of terms, there must be a type distinction and formation rules of both types. Hence, the language I will present is bisorted, with two kinds of terms: regular terms and names of formulas. Another interesting feature of the construction is that it requires a simultaneous definition of terms and formula; this also stems from the fact that there are terms which are names of formulas. An important development of this work would be to derive the Liar Paradox in the language to show its inconsistency; I aim to do that in the future.

I will now define the first-order bisorted semantically closed language $\mathfrak{L}$.[11]

The bisorted language $\mathfrak{L}$ will have two sorts, $\langle$S$\rangle$ and $\langle$F$\rangle$. In practice, S is the sort of simple terms and F is the sort of names for formulas. $\mathfrak{L}$ possesses the following symbols:

(a) A set of constants $\mathcal{C}_s$ of sort $\langle S\rangle$ and a set of constants $\mathcal{C}_f$ of sort $\langle F\rangle$.

(b) Variables for the sort $\langle S\rangle$:

$$x_1, x_2, x_3, ...;$$

(c) Variables for the sort $\langle F\rangle$:

$$y_1, y_2, y_3, ...;$$

(d) for each n, the n-ary function symbols and the n-ary predicate symbols;

(e) the symbols $\neg$, $\vee$, $\forall$ and $\exists$.

(f) The equality predicate symbol: $=$.

(g) The parentheses symbol (,) and the naming symbol $\ulcorner, \urcorner$.

(h) The unary predicate symbol Tr.

The equality symbol $=$ and the truth predicate symbol Tr are called *logical symbols* and other function and predicate symbols that are not $=$ or Tr are called *nonlogical symbols*.[12] In addition, a formula is said to be *atomic* if it is of the form $Px_1, ..., x_n$ or $Py_1, ..., y_n$.

Now, let's provide an inductive definition of terms and formulas simultaneously.

[11] I partially follow Shoenfield [49] to elaborate this definition. There, however, he dismisses the usage of parenthesis by modifying the notation from $(A \vee B)$ to $\vee AB$. Then, he introduces the usual notation $A \vee B$ by definition. Here, the parenthesis will be adopted to facilitate reading, and will be removed in unambiguous contexts in a more intuitive way.

[12] In [49], p. 14.

Definition 1.

(i) Variables $x_1, x_2, x_3, ...$ are terms of sort $\langle S \rangle$.

(ii) Variables $y_1, y_2, y_3, ...$ are terms of sort $\langle F \rangle$.

(iii) If f is an n-ary function symbol and $t_1, ..., t_n$ are terms of sort $\langle S \rangle$, then $ft_1, ..., t_n$ is a S-sorted term.

(iv) If f is an n-ary function symbol and $t_1, ..., t_n$ are terms of sort F, then $ft_1, ..., t_n$ is a F-sorted term.

(v) If t_1 and t_2 are terms of the same sort, then $t_1 = t_2$ is a formula.

(vi) If α is a formula, then $\ulcorner\alpha\urcorner$ is a F-term.

(vii) If P is a n-ary predicate and $t_1, ..., t_n$ are terms of either sort, then $Pt_1, ..., t_n$ is a formula.

(viii) If α is a formula, then $\neg\alpha$ is a formula.

(ix) If α and β are formulas, then $(\alpha \vee \beta)$, $(\alpha \wedge \beta)$, $(\alpha \rightarrow \beta)$ and $(\alpha \leftrightarrow \beta)$ are formulas.

(x) If α is a formula and x is a variable of sort $\langle S \rangle$, then $\exists x\alpha$ and $\forall x\alpha$ are formulas.

(xi) If α is a formula and y is a variable of sort $\langle F \rangle$, then $\exists y\alpha$ and $\forall y\alpha$ are formulas.

(xii) If α is a formula, then $Tr(\ulcorner\alpha\urcorner)$ is a formula.

These are the basic features of the language $\mathfrak{L}$. We now introduce a $\mathfrak{L}$-structure $\mathfrak{A}$ to construct a truth evaluation for $\mathfrak{L}$ in $\mathfrak{A}$. Let $For_{\mathfrak{L}}$ be the set of formulas of $\mathfrak{L}$.

A *structure* $\mathfrak{A}$ for the first-order bisorted semantically closed language $\mathfrak{L}$ consists of the following things:

Definition 2.

(i) A non-empty set $|\mathfrak{A}|$ called the *domain* of $\mathfrak{A}$. The elements of $|\mathfrak{A}|$ are called *individuals* of $\mathfrak{A}$.

(ii) A non-empty set $\mathcal{U}_{\mathfrak{A}} := |\mathfrak{A}| \cup For_{\mathfrak{L}}$ called the *universe* of $\mathfrak{A}$. The elements of $\mathcal{U}_{\mathfrak{A}}$ are the individuals of $\mathfrak{A}$ together with the set of formulas of $\mathfrak{L}$.

(iii) For each S-sorted constant symbol, an element $c_{\mathfrak{A}}$ of $|\mathfrak{A}|$.

(iv) For each F-sorted constant symbol, an element $c_{\mathfrak{A}}$ of $For_{\mathfrak{L}}$. (I.e. a formula of $\mathfrak{L}$ corresponds to each F-sorted constant symbol).

(v) For each S-sorted $n-$ary predicate symbol p of $\mathfrak{L}$ rather than $=$ and Tr, an $n-$ary predicate $p_{\mathfrak{A}} \subseteq |\mathfrak{A}|^n$.

(vi) For each F-sorted $n-$ary predicate symbol p of $\mathfrak{L}$ rather than $=$ and Tr, an $n-$ary predicate $p_{\mathfrak{A}} \subseteq For_{\mathfrak{L}}$.

(vii) For each S-sorted $n-$ary function symbol f of $\mathfrak{L}$, an $n-$ary function $f_{\mathfrak{A}}$ from $|\mathfrak{A}|^n$ to $|\mathfrak{A}|$.

(viii) For each F-sorted $n-$ary function symbol f of $\mathfrak{L}$, an $n-$ary function $f_{\mathfrak{A}}$ from $For_{\mathfrak{L}}^n$ to $For_{\mathfrak{L}}$.

As previously pointed out, the language must be interpreted for us to be able to state that it is semantically closed. This will be done using the concept of *diagram languages*, following Shoenfield's [49]; more specifically, sections 2.3 and 2.5. I of course dissent from his approach in that I work with bisorted semantically closed languages, whereas he works with regular first-order languages. I also omit some subtleties of the text, such as the bolding of syntactical variables – in general, u, v will designate variables that vary through expressions and A, B, C and D, variables that vary through formulas. The ranging of the other variables will be determinable by context. Before actually constructing a structure for the language $\mathfrak{L}$, I shall define truth functions that will prove to be useful to the construction. Let a truth function H be a function such that $H : \{T, F\} \rightarrow \{T, F\}$. We then have:

$H_{\neg}(F) = T$
$H_{\neg}(T) = F$
$H_{\wedge}(T, T) = T$
$H_{\wedge}(T, F) = H_{\wedge}(F, T) = H_{\wedge}(F, F) = F$
$H_{\vee}(T, T) = H_{\vee}(T, F) = H_{\vee}(F, T) = T$
$H_{\vee}(F, F) = F$
$H_{\rightarrow}(T, T) = H_{\rightarrow}(F, T) = H_{\rightarrow}(F, F) = T$
$H_{\rightarrow}(T, F) = F$
$H_{\leftrightarrow}(T, T) = H_{\leftrightarrow}(F, F) = T$
$H_{\leftrightarrow}(T, F) = H_{\leftrightarrow}(F, T) = F$

We now define the diagram language of $\mathfrak{L}$, denominated $\mathfrak{L}(\mathfrak{A})$, by adding S-sorted constants to each individual $b \in |\mathfrak{A}|$. If b is an element of $|\mathfrak{A}|$, its correspondent constant is called its *label*.

Definition. A *variable-free* expression is one which contains no variables.

We then define an element $\mathfrak{A}(t)$ of $\mathcal{U}_{\mathfrak{A}}$ for each variable-free term of t of $L(\mathfrak{A})$. This is done by induction on the length of t:

Definition 3.

(i) If t is the label of an individual $a \in |\mathfrak{A}|$, then $\mathfrak{A}(t)$ is a.

(ii) If t is $\ulcorner\varphi\urcorner$ for a formula φ, then $\mathfrak{A}(t)$ is φ.

(iii) If t is $ft_1...t_n$, then $\mathfrak{A}(t)$ is $f_{\mathfrak{A}}(\mathfrak{A}(t_1), ..., \mathfrak{A}(t_n))$.

Definition. A formula or term A is *closed* if and only if no variable is free in A.

Finally, a truth evaluation $\mathfrak{A}(A)$ is defined for each formula A of $\mathfrak{L}(\mathfrak{A})$. The definition is by induction on the length of A:

Definition 4.

(1) If A is $a = b$, then we let $\mathfrak{A}(A) = \mathbf{T} \Leftrightarrow \mathfrak{A}(a) = \mathfrak{A}(b)$.

(2) If A is $pa_1, ..., a_n$, where p is not $=$, then we let $\mathfrak{A}(A) = \mathbf{T} \Leftrightarrow p_{\mathfrak{A}}(\mathfrak{A}(a_1), ..., \mathfrak{A}(a_n))$.

(3) If A is $\neg B$, then we let $\mathfrak{A}(A) = \mathbf{T} \Leftrightarrow H_{\neg}(\mathfrak{A}(B))$.

(4) If A is $B \vee C$, then we let $\mathfrak{A}(A) = \mathbf{T} \Leftrightarrow H_{\vee}(\mathfrak{A}(B), \mathfrak{A}(C))$.

(5) If A is $B \wedge C$, then we let $\mathfrak{A}(A) = \mathbf{T} \Leftrightarrow H_{\wedge}(\mathfrak{A}(B), \mathfrak{A}(C))$.

(6) If A is $B \rightarrow C$, then $\mathfrak{A}(A) = \mathbf{T} \Leftrightarrow H_{\rightarrow}(\mathfrak{A}(B), \mathfrak{A}(C))$.

(7) If A is $Tr(\ulcorner B\urcorner)$, then $\mathfrak{A}(A) = \mathbf{T} \Leftrightarrow \mathfrak{A}(B) = \mathbf{T}$.

(8) If A is $\exists xB$, then we let $\mathfrak{A}(A) = \mathbf{T} \Leftrightarrow \mathfrak{A}(B_x[i]) = \mathbf{T}$ for some label i in $\mathfrak{L}(\mathfrak{A})$.

(9) If A is $\forall xB$, then we let $\mathfrak{A}(A) = \mathbf{T} \Leftrightarrow \mathfrak{A}(B_x[i]) = \mathbf{T}$ for all labels i in $\mathfrak{L}(\mathfrak{A})$.

(10) If A is $\exists yB$, then we let $\mathfrak{A}(A) = \mathbf{T} \Leftrightarrow \mathfrak{A}(B_y[\ulcorner\varphi\urcorner]) = \mathbf{T}$ for some $\varphi \in For_{\mathfrak{L}}$.

(11) If A is $\forall yB$, then we let $\mathfrak{A}(A) = \mathbf{T} \Leftrightarrow \mathfrak{A}(B_y[\ulcorner\varphi\urcorner]) = \mathbf{T}$ for all $\varphi \in For_{\mathfrak{L}}$.

We have finally completed the construction of the semantically closed language $\mathfrak{L}$ and given an interpretation for it. By looking closely to clauses 9 to 12 of the truth evaluation $\mathfrak{A}(A)$, the reader will notice that, while clauses 9 and 10 determine the use of quantifiers for simple terms, clauses 11 and 12 allow us to quantify over terms of type F, which are names of formulas. Hence, our language can produce sentences such as "there is a false sentence in this language" – or, more precisely, "there is a term t which is a name of a false sentence in this language." It is in agreement with the spirit of semantically closed languages that they are capable of producing sentences that talk about all or some (names of) sentences of themselves.

Now, it must be proven that $\mathfrak{L}(\mathfrak{A})$ is semantically closed. To do that, we simply verify if clauses i and ii of the definition given in 3.3.2 hold – i.e. if the language contains names for its sentences and if it contains a truth predicate that preserves the T-schema.

Proposition. $\mathfrak{L}(\mathfrak{A})$ is a semantically closed language according to the definition of semantically closed languages in 3.3.1.

Proof.

Clause i follows immediately from item iv of definition 1 and item 2 of definition 3. There, we established that if α is a formula, then $\ulcorner\alpha\urcorner$ is a term, and that if t is $\ulcorner\alpha\urcorner$ for a formula α, then $\mathfrak{A}(t)$ is α.

Clause ii follows from two steps:

(1) That $\mathfrak{L}$ contains a truth predicate that applies to names of sentences follows from item (xii) of definition 1: if Tr applies to names of formulas, then it certainly applies to names of sentences.

(2) The structure $\mathfrak{A}$ verifies that the T-schema holds by item 7 of definition 4, by the following reasoning:

$\mathfrak{A}(Tr(\ulcorner\alpha\urcorner) \leftrightarrow \alpha) = \mathbf{T} \Leftrightarrow \mathfrak{A}(Tr(\ulcorner\alpha\urcorner)) = \mathfrak{A}(\alpha)$

Since it is always the case that $\mathfrak{A}(Tr(\ulcorner\alpha\urcorner)) = \mathfrak{A}(\alpha)$, it follows that $\mathfrak{A}(Tr(\ulcorner\alpha\urcorner) \leftrightarrow \alpha) = \mathbf{T}$ for all α. Therefore, the T-schema holds in the language $\mathfrak{L}$.

From (1) and (2), we conclude that $\mathfrak{L}(\mathfrak{A})$ is a semantically closed language. ■

4.5 Comments on the Construction

A difficult conundrum that arises from this construction is the question of the language's classicality or non-classicality. According to Tarski, $\mathfrak{L}$ should indeed contain a true contradiction – a sentence that is evaluated as both true and false. However, we have not yet derived paradoxes in the language, so it is hard to tell whether this is indeed possible. Priest, in [40], had a similar issue: he developed semantically closed theories with respect to the satisfaction predicate, but did not prove their inconsistency, even though he had, before, proven that if a theory satisfies some closure conditions (similar to the ones we devised in section 4, with some exceptions), it is inconsistent. However, because of some technical difficulties concerning the way he defined the satisfaction predicate in the theories, it may be that paradoxes are not derivable in them. He comments:

> To be honest, whether or not they [the semantically closed theories] are inconsistent I do not know. This may seem surprising, since I showed in section 1.2 that any theory which contains its own satisfaction predicate is inconsistent. However, the notion of satisfaction employed there was that of a formula with one free variable by an object. The notion of satisfaction used in the above constructions is the slightly more general one of an (arbitrary) formula by a sequence. We can define the more restricted notion in a fairly obvious way, but it would appear that the appropriate satisfaction scheme for it is not forthcoming, at least not as the theories stand, and we know of no other way in which the theories can be shown to be inconsistent. None the less, if the theories are consistent, they are so for the purely accidental reason that to show the appropriate form of the satisfaction scheme requires some principles about the existence of sequences over and above those so far required; and these principles are entirely unproblematic ([40], p. 130).

This shows that deciding on the classicality or non-classicality of the interpretation for a semantically closed languages is not as simple an issue as it may seem at first. For a language such as the constructed $\mathfrak{L}$ to be capable of producing paradoxes, it must be able to build a sentence α which states "$\neg Tr(\ulcorner\alpha\urcorner)$." Whether $\mathcal{L}$ is capable of doing so is a question which remains as a future goal. Of course, if we can use the Gödel numbering technique to name sentences, $\mathfrak{L}$ will certainly be capable of producing a self-referential sentence such as α. However, to produce a Gödel numbering device, the language must contain some level of arithmetic; since we defined the semantically closed language $\mathfrak{L}$ as generally as possible, it may not contain the degree of arithmetic necessary to name sentences with Gödel numbering. So, we cannot assume that such a technique is available in this case. Without the warranty that we can use the Gödel numbering technique, it is not clear for us if the language $\mathfrak{L}$ is capable of producing formulas referring to their own names. This type of formula is impredicative, and its construction might be blocked by the language's recursive definition of formulas.

Someone could argue that, if the language produces paradoxes, then it requires a non-classical sort of structure, such as the ones Priest develops in [37] and [38], and not the classical one we developed in this paper. Thus, we would be obliged to rebuild the interpretation of $\mathfrak{L}$ differently than we just did. While it is true that the truth evaluation of sentences of $\mathfrak{L}$ would indeed be distinct from Definition 4, the *structure itself*, as defined in Definition 2, would remain the exactly same. More importantly, the truth evaluation shift would *not* produce any alteration in

the proof of the theorem: there, we used only item 7 of Definition 4, and this would remain the same if a non-classical structure were adopted.

It should also be highlighted that we did *not* specify a logic for $\mathfrak{L}$; yet, we were still able to tell that it is semantically closed – which indicates that semantic closure is a structural property of the language in question, one that does not depend on its background logics. Hence, we could determine $\mathfrak{L}$'s background logic after checking whether it does or does not contain sentences such as $\alpha = \neg Tr(\ulcorner\alpha\urcorner)$. If $\mathfrak{L}$ contain such a sentence, we could consider which would be the best non-classical interpretation for $\mathfrak{L}$. If $\mathfrak{L}$ were to have a paraconsistent semantics, for example, we would adjust its truth functions and items 1-11 of Definition 4 accordingly, but the structure and the languages themselves would remain the same.

4.6 The Semantic Closure Operator

Here, I'll sketch the construction of a semantic closure operator. As the verb "sketch" rightly suggests, the construction is still not complete, and it's more of an idea for a semantic operator than the operator *per se.* There are still some formal problems that need to be solved that are now being worked on, so this is a work in progress.

The objective here is not only to take an arbitrary (open) theory and make the necessary changes to render it semantically closed, but to produce a function that, given *any* theory that fulfills some basic requirements, gives as result the theory with (i) names for each of its sentences and (ii) a truth predicate that varies over its sentences. So we're not taking the "constructive" route, that would be to take the theory we're interested in and manually insert names for its sentences and then insert a truth predicate in it to make it semantically closed. Instead, my ultimate goal is to build a tool that can mechanically do this process.

One of the characteristics our operator has to match is monotonicity. It must preserve the order between theories when we semantically close them. The notion of fixed point will also be necessary, as will become clear later on. So, firstly, I introduce the definition of a monotonic operator and fixed points. Note that $\mathcal{P}(X)$ denotes the power set of X.

Definition.[13] Let X be a set and $\Gamma : \mathcal{P}(X) \to \mathcal{P}(X)$ any function. We say that Γ is a *monotone operator* if, given $A \subseteq B \subseteq X$, $\Gamma(A) \subseteq \Gamma(B)$.

[13] [1] p. 744.

Definition. Let $\Gamma : \mathcal{P}(X) \to \mathcal{P}(X)$ be a monotonic operator. We say that $Y \subseteq X$ is Γ-*closed* if $\Gamma(Y) \subseteq Y$. If, in addition, $\Gamma(Y) = Y$, we say that Y is a *fixed point* of Γ.

The semantic closure operator must be a Γ function such that, when applied to a theory $\mathfrak{T}$, it gives as outcome the theory $\mathfrak{T}'$, with $\mathfrak{T}'$ being the theory $\mathfrak{T}$ plus a truth predicate and names for its sentences. Also, $\mathfrak{T}'$ must contain all of Tarski's biconditionals for its sentences. This is what is shown below. Note that Tr is the truth predicate and $\ulcorner\varphi\urcorner$ is a constant symbol, which is the name of φ.

$\Gamma(\mathfrak{T}) = \mathfrak{T}'$, such that $\mathfrak{T}' = \mathfrak{T} \cup \{Tr(\ulcorner\varphi\urcorner) \leftrightarrow \varphi\}$ for every sentence φ of $\mathfrak{L}(\mathfrak{T})$. The language of $\mathfrak{L}(\mathfrak{T}')$ must be such that $\mathfrak{L}(\mathfrak{T}') = \mathfrak{L}(\mathfrak{T}) \cup \{\ulcorner\varphi\urcorner : \varphi$ is a sentence of $\mathfrak{L}(\mathfrak{T})\} \cup \{Tr\}$.

As the description reveals, the operator Γ is a function that adds to the language $\mathfrak{L}(\mathfrak{T})$ the truth predicate and a constant *for each sentence of* $\mathfrak{L}(\mathfrak{T})$ (that is, the constants name the sentences of $\mathfrak{L}(\mathfrak{T})$), and a Tarskian biconditional on $\mathfrak{T}$ for each of its sentences. However, in order for such a function to actually be constructed, it is necessary to find a certain X such that $\Gamma : \mathcal{P}(X) \to \mathcal{P}(X)$, as the definition of a monotonic operator requires, but which set could figure as X, the set that would be the domain for the function Γ? An interesting candidate would be some kind of super language Über $\mathfrak{L}$ ($U\mathfrak{L}$), such that $U\mathfrak{L} = \mathfrak{L}(\mathfrak{T}) \cup \{Tr\} \cup \{\ulcorner\alpha\urcorner : \alpha < \beta\}$, for some ordinal β large enough to ensure that all necessary constants are included in $\mathfrak{L}(\mathfrak{T})$ - that is, $U\mathfrak{L}$ guarantees that there will be no unnamed sentences in the domain of the function Γ. Thus, each application of the operator would yield a subset of $U\mathfrak{L}$ and, recalling the fact that theories are sets of sentences, we have that $\mathcal{P}(U\mathfrak{L})$ gives us all possible theories of $U\mathfrak{L}$.

So the first issue to be solved is that of finding the right β ordinal (sufficiently large to contain all constants needed). However, this is not the only one: if our operator Γ works well, when given a theory $\mathfrak{T}$ as input, it gives as output $\mathfrak{T}$ plus the necessary constants, the truth predicate and Tarski's biconditionals for all sentences. However, nothing guarantees that Γ would also include, in $\mathfrak{T}'$, constants and biconditionals *for the sentences of* $\mathfrak{T}'$. Thus, we would again have to apply the operator Γ to get $\mathfrak{T}''$, which would be $\mathfrak{T}'$ plus biconditionals and constants for the sentences of $\mathfrak{T}'$. Again, the problem is repeated: it is not possible to guarantee that $\mathfrak{T}''$ will contain biconditionals and constants for its own sentences. This is where the fixed point comes to scene: if we find an application of Γ such that $\Gamma(\mathfrak{T}') = \mathfrak{T}'$ for some $\mathfrak{T}'$, we'll have a guarantee that this process

ends at some point in the chain. If we guarantee the existence of a fixed point, the constants problem will be solved: there will be a level for which the application of the operator will result on the very theory we applied it to, and so no more constants will need to be added. Happily, there is indeed a smallest fixed point available, as is shown by the theorem below. [14]

Theorem. If Γ is a monotonic operator, then Γ has a smallest fixed point.

Proof.

(1) Let $\Gamma : \mathcal{P}(X) \to \mathcal{P}(X)$ be a monotonic operator and S a set such that $S = \{Y \subseteq X : \Gamma(Y) \subseteq Y\}$. Show that $\bigcap S$ is a fixed point of Γ.

(a) $\Gamma(\bigcap S) \subseteq \bigcap S$:

Suppose that $x \in \Gamma(\bigcap S)$.

For all $Y \in S$, it is the case that $\bigcap S \subseteq Y$. Since Γ is a monotonic operator, it follows that $\Gamma(\bigcap S) \subseteq \Gamma(Y)$ for all $Y \in S$. Therefore, $x \in \Gamma(Y)$.

By the definition of S, $\Gamma(Y) \subseteq Y$ for all $Y \in S$. Thus, $x \in Y\ \forall Y \in S$.

Then $x \in \bigcap S$ and it follows that $\Gamma(\bigcap S) \subseteq \bigcap S$.

(b) $\bigcap S \subseteq \Gamma(\bigcap S)$:

Suppose that $x \in \bigcap S$. Then $x \in Y$ for all $Y \in S$.

Since $\Gamma : \mathcal{P}(X) \to \mathcal{P}(X)$, it follows that for any Z, $\Gamma(Z) \subseteq X$. In addition, $\Gamma(\bigcap S) \subseteq \bigcap S$. Hence, by the definition of S, it follows that $\Gamma(\bigcap S) \in S$.

So $x \in \Gamma(\bigcap S)$ and it follows that $\bigcap S \subseteq \Gamma(\bigcap S)$.

By **(a)** and **(b)**, we have that $\bigcap S = \Gamma(\bigcap S)$, i.e. $\bigcap S$ is a fixed point of Γ.

(2) Show that $\bigcap S$ is the smallest fixed point of Γ, *i.e.* if Y^* is any fixed point of Γ, $\bigcap S \subseteq Y^*$.

Suppose that $x \in \bigcap S$. Then $x \in Y$ for all $Y \in S$.

Since Y^* is a fixed point of Γ, $Y^* \subseteq X$ and $\Gamma(Y^*) \subseteq Y^*$. But that is precisely the definition of the elements of S. Thus, $Y^* \in S$.

Therefore, $x \in Y^*$ and it follows that $\bigcap S \subseteq Y^*$.

[14] Note that, in the proof, we define a set $S = \{Y \subseteq X : \Gamma(Y) \subseteq Y\}$. This S must be non-empty, otherwise $\bigcap S$ would incur in paradox (because the intersection of the empty set is the set of all sets). How do we guarantee that there is a Y such that $Y \subseteq X$ and $\Gamma(Y) \subseteq Y$? Remember that the operator Γ acts under a domain which we called X. Thus, of course $\Gamma(X) \subseteq X$, because X is the limit we established. Hence, $\bigcap S$ is not a paradoxical entity and so the theorem runs smoothly.

■

From this section, we can see that the construction of the semantic closure operator will go on smoothly if we find the right ordinal β to make up the domain in which Γ will function. This gear is still to be found for the completion of the construction.

4.7 Everyday Languages

In this section, I will address the following question: are everyday languages, such as Portuguese or English, semantically closed? Such a matter may seem simple, but the terrain of colloquial languages is slippery and hides problems that are difficult to resolve. In that spirit, I will start by making an important caveat to the discussion: it is not that clear, at least at first glance, *what it means* to ask if colloquial languages are semantically closed or open. That is because the definition exposed in 4.3.1 is directed to formal languages, and everyday languages could hardly be said to fit into this group. This is not a small issue: transposing formal concepts to informal settings can lead (and has many times led) to disaster and poor analysis. But I believe – and will defend – that this is not the case here. This is, rather, an example of a fruitful transposition of concepts to different settings: although it is necessary to make some concessions, this inquiry can bring about interesting insights to the study of everyday languages.

The logician Alfred Tarski considers this problem two times, and seems to change his views from a first consideration to a second. In his seminal 1936 paper, he states:

> But it is presumably the universality of everyday language which is the primary source of all semantical antinomies, like the antinomies of the liar or of heterological words. These antinomies seem to provide proof that everyday language which is universal in the above sense, and for which the normal laws of logic hold, must be inconsistent ([51], p. 164).

Although the author is referring to the phenomenon of inconsistency of a language,[15] and not directly to semantic closure, as inconsistency is a *product* of semantic closure, the point made is just as valid here. And the point is

[15]There is, of course, a problem with the concept of inconsistency as related to *languages*, and not theories: in a sense, only theories can be inconsistent, and that happens when they contain contradictions as theorems. We will go back to this issue soon in the next section.

precisely that everyday non-formal language is semantically closed (and, therefore, inconsistent) in virtue of its *universality*, *i.e.,* everything that can be stated at all can be stated using everyday language. If this is true, then of course colloquial languages contain the necessary means to create antinomies, which lead to their inconsistency. However, some years later, in 1944, the logician assumes a much more careful position:

> Our everyday language is certainly not one with an exactly specified structure. We do not know precisely which expressions are sentences, and we know even to a smaller degree which sentences are to be taken as assertible. Thus the problem of consistency has no exact meaning with respect to this language. We may at best only risk the guess that a language whose structure has been exactly specified and which resembles our everyday language as closely as possible would be inconsistent ([52], p. 349).

Is Tarski's worry fair? In a sense, yes, and that should certainly be taken into account. It is true that – at least until now – no structure has been found that formalizes the whole of everyday language, and so the discussion revolves around a structure whose rules are unclear.[16] That does not, however, necessarily mean that the problem of semantic closure (or inconsistency) has no exact meaning in colloquial language; even if it's not yet clear which sentences are assertible in this environment, if we conclude that colloquial languages fulfill the necessary conditions to be semantically closed – that is, if they contain their own truth predicate and names for their sentences – then why shouldn't we conclude that they are semantically closed? I believe we certainly could. We often make assertions about structures about which we are, at least partly, ignorant; in fact, that is the case with every attempt to formalize colloquial languages. In what follows, I will provide the arguments usually given to justify both sides: that colloquial languages are semantically closed and that they're open. By the end of these sections, I hope to show how the arguments for the view that everyday languages are semantically closed are more persuasive than the opposite one.

[16] A possible counter-argument to this could be the following: it's not necessary to know the entire structure of a language to derive properties about it. If someone blindfolds me, hands me a copy of the book *Don Quixote* and asks "what is this?," I will probably be able to claim that it is a book and enunciate some properties about it: it is thick, it has a hard cover, the paper is thin, etc, even without knowing its name or what is written on it. Something similar could be happening in this case. If it was decided that everyday languages fulfill the conditions necessary to be semantically closed, that would be enough to say that they are semantically closed, regardless of what is inside them.

4.7.1 Are semantically open

To defend that everyday language is semantically *open*, one must deny clause (i) or (ii) of the definition of semantically closed language written in 4.3.1. Roughly, one must deny one (or both) of the following claims: (1) that non-formal languages can name their own sentences; or (2) that they contain their own truth predicate and that this predicate satisfies the T-schema. I dare say that anyone's reflex response would be to accept both claims: it seems no less than obvious that everyday languages can name their own sentences, for quotation marks fulfill exactly this task; likewise, they should of course contain their own truth predicate – after all, I can say, in English, that the sentence "snow is white" is a true sentence of English. Such scenario compels me to admit that arguing for the semantic openness of everyday language is a much harder task than arguing for its closure, which can be seen as the "naive" position. I must be humble and grant them that their task might be more difficult than my own. Sometimes we find that between common sense and the world there is an almost impassable gulf; at other times, however, our most naive intuitions prove true. I believe this to be the second case, where truth can be found where our immediate intuition lies.

Martin, in [28], remarks with surprise that the orthodox view is on this matter is the counter-intuitive one, arguing that everyday language is semantically open (or, in his jargon, natural language cannot give its own semantics). Of course, we must be mindful that Martin's paper was written in 1976, so the philosophical landscape has surely changed since then. In any event, he names important philosophers who leaned towards the idea that everyday language must not be universal – if anything expressible can be expressed in everyday language, then it can also express its own semantics and, thus, formulate paradoxes. Davidson, in [12], is of this opinion, and flirts with the idea that colloquial language may be organized in hierarchies of a Tarskian kind.[17]

Adherents of colloquial language hierarchies argue that English (or any colloquial language) does not contain a single truth predicate (this should hold

[17]Even though reluctantly, Davidson admits it is unlikely that language is universal, and puts forward the idea of natural language hierarchies: "But it is not really clear how unfair to Urdu or to Hindi it would be to view the range of their quantifiers as insufficient to yield an explicit definition of 'true-in-Urdu' or 'true-in-Hindi'. Or, to put the matter in another, if not more serious way, there may in the nature of the case always be something we grasp in understanding the language of another (the concept of truth) that we cannot communicate to him. In any case, most of the problems of general philosophical interest arise within a fragment of the relevant natural language that may be conceived as containing very little set theory. Of course these comments do not meet the claim that natural languages are universal. But it seems to me this claim, now that we know such universality leads to paradox, is suspect" ([12], p. 314).

for semantic predicates in general) that varies among its sentences, but infinite predicates T_0, T_1, T_2, ..., which vary among *an infinite number of languages*.[18] Such "infinite number of languages" is nothing more than colloquial languages themselves: what we call English, for example, is not a single language, but an *infinite chain of languages*, with infinite semantic predicates. Therefore, English would be structured as is explained in [50], p. 152:

E_0 = English minus semantic predicates.

$E_1 = E_0 + T_0$, where T_0 applies to all true sentences of E_0 and does not apply to anything else.

$E_2 = E_1 + T_1$, where T_1 applies to all true sentences of E_1 and does not apply to anything else.

And so on, so that we can conclude:

$\forall E_n, \forall T_n \, (T_n \in E_{n+1} \wedge T_n \notin E_n)$

We call the language *to which* the truth predicate is defined *object language*, and we call the language *in which* the truth predicate is defined *metalanguage*. In the construction, therefore, the evaluation of the true sentences of a certain language takes place only in its metalanguage. In order to actually be able to build the hierarchy of languages, one has to specify the exact requirements that a language must meet to function as the metalanguage for another. Those are:

(1) The metalanguage must contain the object language.

(2) The metalanguage must contain a citation device.

(3) The metalanguage must contain the truth predicate for the object language, that is, it must contain the predicate "true-in-the-object-language."[19]

The reason why we stipulate **(1)** is clear: in order for me to evaluate object language sentences in a metalanguage, I need to be able to express them in this metalanguage. After all, I cannot write sentences of type "the sentence '*object language sentence*' is true in the object language" in the metalanguage, if the latter is not able to express the symbols written in the sentence. The criterion **(2)** is given by the need to *mention* the sentence: just like a fishhook, the metalanguage grabs sentences from the depths of the object language, brings them to firm land and evaluates them. The citation mechanism is the tool needed for the

[18] For an exposition of the Tarskian hierarchy for formal languages, refer to the appendix at the end of this text.

[19] Adapted from [6], p. 19.

metalanguage to select the sentences. The criterion **(3)** is the very reason for the arrangement of languages in hierarchies, so it does not need further explanation.

Both the first and second conditions do not need to be taken too strictly, and can be relaxed as follows: in the case of (1), it is not necessary for the object language to *literally* contain the metalanguage, it's enough that it is possible to *translate* the first into the second. Of course, it is vague to allow for translations of object languages in metalanguages and not specify what an effective translation would look like. I shall, however, not answer this question and remain vague on purpose, so that translations can be analyzed case by case. Regarding (2), it is not necessary for the metalanguage to contain *quotation marks*, the template of citation devices; if it can name sentences (with no matter what citation device), condition (2) is already guaranteed. Condition (3) is the only one that cannot be loosened, of course, because it is the reason why language hierarchies were developed in the first place.

This characterization is sufficient to contain the following version of the Liar Paradox:

(UL) UL is not true.

Let's say this sentence was formulated in E_1. So the truth predicate must refer to E_0, so what the sentence actually says is:

UL is not $true_0$.

This last sentence is free from paradoxical content, since it says of itself that it is not true *in the object language*, and not in the language in which it was written.

An immediate objection to this kind of approach is the question: why should we believe that English works this way? At first glance, it seems quite outlandish to believe that colloquial languages are constructed in Tarskian hierarchies only to escape paradoxes. This is one of Priest's objections to this view ([35]). The move is *ad hoc*, he claims, for the "surface structure" of English is certainly not of this form. Moreover, this is certainly not the perception speakers have of their mother tongue – no one who is not deprived of reason would try to attribute everything he says to the correct language level and use only the semantic predicates with correct subscripts. To this, the defenders of the hierarchy approach can respond that the indexing of semantic predicates is implicit: the predicates and levels of languages are employed in such a way that no one needs to spell them out literally,

as they are logical substructures of the language. However, even so, to explain the fact that the language usage works – that is, that individuals understand each other – advocates of the approach will have to admit that all speakers understand the hierarchy and use it correctly, which is a strange claim to make, given that most speakers (apart of some logicians) do not even know about the theory of natural language hierarchies.

Another way to explain hierarchies is through *contexts*: a defender of the approach could claim that English is a bundle of different languages, each of which represents a context of communication, and each context occupies a level in the hierarchy. It could be something like "the context of logic conversations at the University of São Paulo"; this would be a determinate language, occupying a level in the hierarchy, which would include words and predicates specific to that context. Then, the boundary between E_n and E_{n+1} would be determined by differences between contexts. Although the picture makes some sense, it's unclear why a context would *contain* another, and why the truth predicate of a context could be defined by another. A much more complex justification would be needed to explain the plausibility of the hierarchy organization.

But now, for the sake of argumentation, let's set the plausibility issue aside and suppose that English is an infinite hierarchy of languages. Another problem then arises: the restrictions imposed by the hierarchy are perhaps too severe, and exclude from the language things that we would like to be able to affirm, and that we can in fact affirm in the ordinary use of language. Priest's example is as follows:[20] I would like to be able to claim that all the sentences written in this text (including this one) are true. But I cannot do such deed, for the sentences said to be true occur under the same truth predicate as the sentence which asserts that they are true. Even more serious than this is the fact that "the language we use to describe the hierarchy is itself a member of the hierarchy." ([50], p. 156) – thus, there are many claims about the hierarchy that simply cannot be made, such as "English does not contain its own truth predicate" ([50], p. 156). If English is a hierarchy, this means that there is no *one* truth predicate that applies to all and only English sentences (which is true according to the hierarchical picture). However, this was stated in English (that is, in some language within the hierarchy) and so I cannot refer, at once, to *all* truth predicates – in particular, I cannot refer the truth predicates located in languages above the one where I uttered the sentence. Since the hierarchy is infinite, there will always be a higher

[20] See [35], p. 122.

level language which cannot be denoted.

Priest points also to the difficulty of the infinite regress, which is related to the previous objection:

> Any semantically open language is forbidden to talk of its own semantics. Yet we can and do wish to talk of its semantics. This requires a semantically open metalanguage. But of course we wish to talk about its semantics; so we need a meta-meta-language. And since this is open, to talk of its semantics we need a meta-meta-meta ([35], p. 122).

Advocates of this approach should settle with the fact that we will never be able to describe English's semantics. The same happens in Tarski's hierarchy of artificial languages but, in that case, this restriction is not problematic. If I wish to describe the semantics of L_0, then I select another language L_1 that contains the first and can work as its metalanguage. If I wish to examine the semantics of L_1, then I select yet another language, L_2, as its metalanguage. As they're artificial languages, there's no point in describing the semantics of all languages in the hierarchy, and one can simply select the ones that seem interesting to him. However, if the hierarchy is itself a colloquial language, the issue is completely different: it suddenly becomes quite an important aim to describe all languages in the hierarchy, because we do want to describe the semantics of English (after all, philosophers and linguists have been working on this enterprise for quite some time). In the case where English is a hierarchy, restricting the scope of semantic analysis to only *some* languages in the hierarchy is damaging and arbitrary.

A last objection to the supposed hierarchical structure of English is that it fails to fulfill its purpose in the first place. We were hoping that it could block the Liar but, as usual, revenge paradoxes always lurk behind every solution to them. This might not, alone, make a strong case against the hierarchy of colloquial languages but, together with the other objections, helps us see the ineffectiveness of the theory. Take the hierarchy of the English language I drew earlier in the text. Imagine that, in some language E_n, we find the following sentence:[21]

> *(S')* Sentence S' is not true in any language in the English hierarchy.

Since the sentence's name is S', it is saying of itself that it is not true (just as the Liar) in any language. To be more precise, we could have stated it as saying "For all $m > 0$, the sentence S' is not true in E_m." The sentence cannot be

[21] This revenge paradox was adapted from [21] p. 58.

evaluated in E_n so, in E_{n+1}, we ask: is S' true-in-E_n or false-in-E_n? At this point in the thesis, the reader must be tired of paradox derivation, so I will leave the proof to excited readers who wish to check it on their own; I will merely say that the structure is quite similar to the Liar Cycle and Yablo's Paradox (chapters 2 and 3). The reason I said this might not be the strongest objection of all is because it supposes that a sentence in some language can quantify over the truth predicates of all languages in the hierarchy, even though it can be evaluated as true or false only in languages that are above it in the hierarchy. To stop this revenge paradox, we should also prevent any language from creating sentences that mention the truth predicate of languages above it. So, as I said, it is not the strongest objection that can be thrown against the colloquial languages hierarchy but, together with all the flaws I presented in this section, this should be enough to withdraw it from the batch of candidates to explain the structure of colloquial languages.

4.7.2 Are semantically closed

As I said in the last section, the belief that everyday languages are semantically closed can be called the *naive* view, and I argue (with the help of Priest in [35]) that it is also the correct one. My argument stems from the implausibility of the view that they are semantically open and from the many logical problems it encounters, as I exposed in the last section (infinite regress, revenge paradoxes and so on). I will also argue for the semantic closure of colloquial languages by defending that (i) they contain multiple citation mechanisms and can, thus, name their own sentences and (ii) that they contain their own truth predicate and the T-schema is valid in those languages. I will now go over these two points.

The first point is the easiest one to argue for; it is only the second item which calls for a more detailed argumentation. To argue for point (i), I can simply exhibit a device that does the job of naming everyday language's sentences: quotation marks, of course. But not only this: demonstratives like "this," "that," or "these" may also be used as a sort of naming device. Naming, here, is referred to in the broad sense of the word. Neither quotation marks not demonstratives explicitly name sentences, but they do exactly the job we need them to do, which is to allow us to *mention* sentences, rather than *to use* them. When I say "this sentence is false," the demonstrative "this" is used to mention the very sentence written under quotes; similarly, when I state that the sentence "I am lying" is paradoxical, I am using quotation marks to mention the sentence I am referring

to. If one prefers, everyday languages also have explicit naming resources. I can name as "John" the sentence "John is true" and produce the same effect as before.[22]

Someone could argue that there may be a non-formal language, say Trulu, spoken, for example, at a tribe in South Malaysia, that does not have any sort of mentioning device, and so speakers of this tribe cannot refer to sentences of their own language. Although I find that hard to believe – this seems to be an essential mechanism of non-formal languages – I must say "fair enough, Trulu is semantically open after all." In this case, I will gladly restrict the argument for all non-formal languages except Trulu. However, even in this scenario, the exception wouldn't be disastrous for my argument, since nearly all non-formal languages would remain semantically closed. In any event, until someone presents me a Trulu-like language, I will argue that colloquial languages are, in general, semantically closed.

Point (ii) of my claim is more contentious and requires extensive examination. Although colloquial languages certainly seem to contain their own truth predicates, one could argue that the rules governing the usage of the word "true" are nothing like what the T-schema depicts. One could say, for instance, that the adjective "true" can be used to emphasize a fact or a characteristic. For example, compare the two uses of the word:

(1) It is *true* that Fernando Pessoa was a poet.

(2) Fernando Pessoa was a *true* poet.

(1) is in complete agreement with the idea expressed in the T-schema: the sentence "it is true that Fernando Pessoa was a poet" is equivalent to the sentence without the truth predicate, "Fernando Pessoa was a poet." As the schema shows, one can simply erase the truth predicate from the sentence and its content will remain the same. In (2), the usage of the word "true" points to a very different direction: it is employed as an antonym of "impostor." To say that Fernando Pessoa was a true poet is to say that he was not a fake poet. In this case, "true poets" are in opposition to fake ones, and the statement is also implicitly asserting that there are fake poets – phonies, impostors – and Pessoa is not one of them. He is the "real deal." If we erase the word "true" from sentence number (2), its whole sense will be lost. This is only one example of a myriad of different uses of the word in everyday language which are not in agreement with the T-schema.

[22] Just as Kripke does in [25].

Does that mean that the truth predicate in everyday language does not respect the T-schema, and so everyday languages are semantically open? Not quite. From a logical point of view, presenting examples of uses of the word "true" which are in disagreement with the T-schema is far from a sufficient argument against semantic closure. This is because it's enough for the word to have *one use* in agreement to the T-schema for the conclusion of semantic closure to follow. Sentence (1) is precisely a use of "true" which is in agreement with the T-schema. Thus, it follows that languages capable of formulating sentences like this will fulfill condition (ii) of the semantic closure definition. The requirement (ii) of the definition of semantic closure is not meant to model all possible applications of the word "true" – or, put another way, it is not meant to capture all its homonyms – but only to establish the minimal conditions for a language to be able to give its own semantics.

With the previous objection resolved, we stumble on another stone on the way to the semantic closure of everyday languages. As said earlier, a semantically closed language must be an *interpreted* language, for it makes no sense to talk about semantic closure in a language whose sentences have no meaning at all. But what would a suitable interpretation of everyday language be like? If everyday language is, as said Tarski, the sole *universal* language, what linguistic realm could possibly serve as an interpretation for it? It would have to be a realm that could interpret all symbols of English (or any other colloquial language) in an unambiguous manner, but it's highly questionable that this is even possible, due to the gigantic plurality of senses and meanings that can exist in a non-formal language.

There are some ways to go about this problem. One is to doubt the universality of everyday languages, at least as Tarski claims it. Maybe not everything that can be said can be put in everyday language's terms. Maybe there are some realms of discourse which require formal languages and cannot be translated to non-formal ones. In this picture, there could be an interpretation of everyday languages which is not contained in them, and through this interpretation we would find that the T-schema holds in those languages. From a certain perspective, this view is quite plausible: it is hard to imagine that mathematics, programming, logic and such fields could be developed entirely in everyday languages. Of course we can try to redo all mathematics or all logic using only everyday language terms, but it seems that new formal languages must always be developed for these areas to really flourish. Formal languages are not simple extras that help us make sense

of what is done in mathematics, logic or programming; they are the very vessels in which essential concepts of those domains come to life.

Nevertheless, the problem remains: the boundaries between formal and non-formal languages are not perfectly clear. It is true that we need formal languages to do mathematics, but we also need English, Portuguese or whatever metalanguage one is doing mathematics in. The point where colloquial language ends and formal language begins is extremely dim; when you open a mathematics or logic textbook, you will probably find a continuous mixture of the two, not one or the other type of language. If we found out that those formal languages are somehow contained in colloquial language, then the latter would be universal. I shall not decide upon this matter for the moment.

Another way to tackle the issue of interpretation for a colloquial language is to regard it as embedding its own interpretation: as there doesn't seem to exist any linguistic domain prior to colloquial language, maybe its interpretation could be located within itself. Logically, it would be an enormous issue to figure out how this could be so, but there are some interesting philosophical reasons that motivate this suggestion: when we learn our mother tongue, it can be argued that we do so with no prior background linguistic knowledge – that is, with no metalanguage of any sort. So it can be said that we learn our mother tongue from *within* language itself, and not from a separate outside realm. In fact, there is no separate realm whatsoever, for once we're in language, there's no leaving it. Therefore, we could ask: how can one understand his mother tongue with no interpretation associated with its symbols? One reasonable conclusion would be to argue that its interpretation is located within it, and this is what enables us to learn its symbols and understand the language.

Perhaps the best idea to tackle the issue is neither one of the above, but to assert that non-formal languages are closer to formal theories than to formal languages, so that there would be something resembling *theoremhood* in everyday language. If that was the case, no metalanguage would be needed to state the truth of all the instances of the T-schema – we could simply say that all instances of it are "theorems of English," so to speak. The analogy would be between theorems and the "establishing" of certain statements. For example, the sentence "the Moon's gravity causes tides in the ocean" could be seen as a theorem, something that has been derived from certain principles and rules, and that has finally been established in English as a derivable sentence. The same would be true for the T-schema: all of its instances could be said to be derivable

from certain basic principles and rules regarding truth, and are now elements of the set of "theorems" of English.

"Assertability" is the notion we are looking for. This would be the analogue of theoremhood for languages such as English, Portuguese and so on. To assert something is to claim that something is the case; so assertability is a measure of "how assertible" a sentence is. It comes from the idea that discourse has implicit rules that we follow to state that things are or are not the case. These rules establish which sentences are assertible and which are not, just as a theory establishes which sentences are theorems and which are not. Spelling out the assertibility rules that underlie discourse precisely would take an enormous amount of work and raise all kinds of objections. Since the enterprise falls outside the scope of this text, I will not row in waters so agitated. It suffices to say that we have an analogue of theoremhood available at hand, which is the notion of assertability. With it, we can escape the issue with interpretations, and claim that point (ii) of the definition of semantic closure can be written along the following lines (where the colloquial language in question is English):

(ii) *Truth predicate.* There must exist in English a predicate Tr, such that $Tr(x)$ is a sentence of English if and only if there exists a sentence λ of English such that x is the name of λ, that is, x is $\ulcorner\lambda\urcorner$. The Tr predicate must behave in such a way that all instances of the schema

$$Tr(\ulcorner\lambda\urcorner) \leftrightarrow \lambda$$

are assertible in English.

So far, I have proposed solutions to the objections that (a) the word "true" in languages such as English works nothing like the formal truth predicate and does not respect the T-schema; and that (b) the definition of semantically closed languages given before presupposes a model-theoretic interpretation of the language, and no such interpretation can be given for colloquial languages. Although these problems are important, they are not the main motivation logicians usually reject that everyday languages are semantically closed. By far, the main motivation is expressed in Tarski's quote in [51] (see section 4.7 of this text for the citation): he states that, because of semantic closure, everyday languages are *inconsistent*. However, this is not the only possibility – gappy or paracomplete logics bring one more possibility to the table: if colloquial languages are semantically closed, they are either inconsistent or contain gaps. That means

that either they contain at least a sentence which can be proven to be true *and* false, or which is neither true nor false. These are sentences such as the Liar or Grelling-Nelson's Paradox.[23] Both alternatives pose, of course, a huge threat to committed classical logicians who believe that non-formal languages are governed by classical rules, just as is reasoning itself. If they contain contradictions, by the classical ECQ principle,[24] they can be proven to be trivial. If they contain gaps, then they violate the LEM.[25] For classical logicians, both alternatives are hopeless.

I'm aware that the reader might be angry with my usage of the term *inconsistency* as referring to languages, and maybe even complaining about the awful imprecision that reigns in today's logic texts. To appease such justified tempers, I should specify what I mean by an inconsistent language.[26] The following definition will suffice to do so:

Definition. A language $\mathfrak{L}$ is said to be *inconsistent*$_{\mathfrak{L}}$ if and only if there is a sentence φ in $\mathfrak{L}$ such that both φ and $\neg\varphi$ are regarded as true.

This is quite similar to the definition proposed by Herzberger of *truth-conditional inconsistency*, by which he means a language that "incorporates at least one truth condition which can possibly obtain and which if it obtained would render true each member of an inconsistent set of sentences." ([19] p. 32). In Herzberger's paper, he argues that semantically closed languages are truth-conditionally inconsistent, and so non-formal languages cannot be semantically closed. To do that, he offers a proof that there can be no truth-conditionally inconsistent language, by appealing to the fact that, if $\mathfrak{L}$ is truth-conditionally inconsistent, by the *Ex Contradictione* principle, every sentence is *true,* and so the language is trivialized. Therefore, colloquial languages have to be semantically open. This argument may convince a classical logician: if everyday languages were semantically closed, they would contain true contradictions. As this is a logical impossibility, they cannot be semantically closed. But it certainly won't have much success with non-classical logicians, specially with those who believe that there are, in fact, true contradictions.

[23] See chapter two for proofs of the paradoxes.

[24] Defined in chapter 1.2.

[25] Defined in chapter 1.2.

[26] The worry is completely fair, but I should just add that Tarski himself, in [51], defends that everyday language is *inconsistent* without further explanation of the conclusion. We can only guess that he understands an inconsistent language much like I do in this text: as one that contains true contradictions.

One of these logicians is the *dialetheist* Graham Priest. Still in [35], Priest comments on Herzberger's proof:

> He [Herzberger] argues that there can be no such [truth-conditionally inconsistent] language, on the grounds that if there were, there would be true contradictions. Of course the argument works only if one rejects the view that there are true contradictions. This is precisely what I am denying. That there are true contradictions is an idea which is at the very root of the semantics of paraconsistent logics. In fact Herzbergers' argument is related to mine as *modus tollens* is to *modus ponens*. ([35], p. 120).

Priest's argument is not that Herzberger's proof does not work, but that he's only drawing the wrong conclusions from it. In fact, the proof most certainly does work, and its conclusion should be that everyday languages have a paraconsistent semantics. Since paraconsistent logics are those that do not accept the ECQ principle, if English is $inconsistent_{\mathfrak{L}}$, it does not follow that every sentence in English is true – that is, the language is not trivialized. As I showed in the last section, Priest's argument for defending that non-formal languages are semantically closed goes in the same direction of Tarski's analysis in [51]: there is no epistemic reason to consider that everyday languages are semantically open – to do so is a clear case of postulation *ad hoc*. However, I must remark that, to defend that semantically closed languages are inconsistent, Priest has to defend that they are *not* gappy, which he does in [35], section 5. I will not get into this point, though, because the central question here is solely if everyday languages are semantically closed or not, and not which is their actual background logic.

Surprisingly, the most frequent argument used to reject the semantic closure of everyday languages is the weakest, for it uses the fact that semantically closed languages are inconsistent (incomplete) to argue that everyday languages cannot be so. Of course, we must take care not to fall in anachronisms since, at the time most arguments (Tarki's, later Herzberger's etc) were written, non-classical logics were not as developed as they are today, and logicians commonly did not even take into account the possibility of having inconsistent (incomplete) but non trivial languages.

4.7.3 Consequences of Semantic Closure

If my defense accomplishes success, then we should conclude that everyday languages are semantically closed. With this comes the additional conclusion

that they are either inconsistent or incomplete, for they will contain paradoxical statements. So, what comes next? What consequences can be drawn from these facts?

Most importantly, this means that we would need non-classical semantics to model everyday languages such as Portuguese or English. This would open a whole new area of inquiry, that would ask what kind of structure would be needed to model those languages. Research could also be done to understand how contradictions work in everyday languages: if systems should be regarded as inconsistent but non-trivial (non-explosive), incomplete (with a restriction or abandonment of the LEM) or require yet another formalization.

4.8 Conclusion

The first part of this chapter is dedicated to exploring the concept of semantically closed languages and theories from a formal point of view. This investigation revealed some interesting, although unsettling, facts: one of those is that, to define semantically closed languages, we have to appeal to the model-theoretic concept of interpretation, to state that "all instances of the T-schema are true-in-the-intended-interpretation." Without a metalanguage of any sort, it would be impossible to state that "all instances of the T-schema are *true*," as is our aim to do. It's thus an unsettling result, for our aim was to characterize a language which has enough expressive power to provide its own semantics – *i.e.* interpret its own sentences – and, to do precisely that, we need to construct an interpretation outside the language itself. To use an expression coined by Kripke, the ghost of the Tarski hierarchy[27] is still among us. Still in the first part of the chapter, I provided the construction of a first-order bisorted semantically closed language and proved it to be semantically closed (according to the general definition in 3.3.1).

In the second part of the chapter, I relate the theme of semantic closure to everyday languages, by asking the question "are everyday languages semantically closed?" My take is that they should indeed be regarded as such. I defended this point at first indirectly, by arguing that it is implausible and erroneous to regard non-formal languages as semantically open, and thus they should be taken to be semantically closed. Next, I presented direct arguments for the view that everyday languages preserve items (i) and (ii) of the definition of semantically

[27] [25], p. 714.

closed languages. This thesis has quite a strong consequence: that colloquial language must be either $inconsistent_{\mathfrak{L}}$ or gappy, and so it should contain at least a sentence φ such that either φ and $\neg\varphi$ are both regarded as true, or neither φ nor $\neg\varphi$ is true. How a semantics to non-formal language would look like still remains to be seen, and it's certainly too large an enterprise to the scope of this text; my aim was solely to argue that they are semantically closed.

Appendix: Possible Solutions to the Liar Paradox

In this appendix, I will present an overview of the myriad of possible solutions to the Liar Paradox that have conquered an important spot in contemporary logic. Since solutions to the Liar are not the central theme of this text, this overview will be incredibly reductive and will unavoidably leave out prominent approaches – which is the reason why I left it to an appendix. Even if brief, the overview is necessary for a deeper understanding of the consequences of the Liar Paradox and of much of what I developed in this text.

One of the most fascinating aspects of the Liar is that anyone who wishes to provide a formal solution for it must create a *theory of truth,* or at least a theory of the *truth predicate.* The choice of such a theory involves both technical aspects, to determine the best possible way to dodge paradoxes in a formal language, and philosophical ones, to determine how truth evaluations work or should work. Such are the depths of the consequences of the Liar: it forces us to rethink the whole of the applications of the truth predicate.

Following the exact path of [5], I will present, in the next topics, three ways to solve the problem brought by the Liar's Paradox to the concept of truth. The first are heterodox logics solutions, which give up some principle of classical logic, for they identify a failure in classical reasoning. The second way are solutions of classical logic. There is more than one solution, but what they all have in common is that, in order to maintain the classical foundations, they restrict the applications of the capture-and-release principles. The third solution is given by substructural approaches: according to them, the problem that leads to the paradox is at the level of substructures, a deeper level than that of principles like the Excluded Middle or the T-schema. Instead, the problem lies in rules such as cut and contraction.

Heterodox Logics

The Liar Paradox undoubtedly shows us that there is a loose cog in our reasoning pattern. The disagreement between different theoretical currents is in deciding which are the problematic elements that should be modified. Those who advocate for the use of heterodox logics think that the problem lies in the reasoning established by classical logic: some classical principle must be abandoned or restricted in order for the paradox to be solved.

What many non-classical theories have in common is that they consider it necessary to maintain the full extent of the T-schema. In all non-opaque contexts, it must hold for a sentence A that "$Tr(\ulcorner A \urcorner) \leftrightarrow A$." This often comes together with a *deflationist* conception. There are many forms of deflationism about truth and many different theories that adopt it. It is not a particular well-defined thesis, but rather a certain *attitude*: that one must understand the concept of truth as a simple logical notion and not as a deep philosophical concept. In [6], the authors Burgess and Burgess stipulate three theses to which deflationists commit. Although they should not be taken as a rule, the theses illustrate the current well. The first thesis states that *applying the truth predicate to something is equivalent to simply saying it.* This is called the *equivalence principle*, a version of which is manifested in the T-scheme, but there are many others. The second thesis claims that *the equivalence principle is a sufficient notion of the meaning of the truth predicate.* The meaning of the truth predicate can be satisfactorily explained by what is said in the first thesis, which is expressed in the T-schema. The third thesis posits that *a notion of the meaning of 'true' is a sufficient notion of the nature of truth.* Together with the second thesis, the latter explains the explicit meaning of the term 'deflationism': we must deflate the concept of truth, because behind it there are no unexplored mysteries. To explain well the truth predicate, it is necessary to resort only to the principle of equivalence; to describe the very nature of truth, it is not necessary to go beyond explaining what the term 'truth' means.

Paracomplete logics are alternatives that aim to solve the Liar paradox by changing the logic. Liar sentences show that the Law of the Excluded Middle does not hold for all contexts; that is, it is not always valid that $\vdash A \vee \neg A$. In fact, looking at Proof 1 in section 2.1, if we do not take LEM as a logical truth, we do not reach the conclusion described in step 4, since we cannot, from 3, use the LEM to conclude $Tr(\ulcorner A \urcorner) \wedge \neg Tr(\ulcorner A \urcorner)$. What is such a view telling us? That the problem is not in the Tarskian T-schema, but rather in some instances of

LEM. This caveat is important: the law is inadequate for certain sentences; as Hartry Field puts it, paracomplete solutions claim that the "excluded middle fails in some sentences involving "true of." In particular, the assumption that 'not true of itself' is either true of itself or not true of itself is fallacious" ([17], p. 11). Overall, models of paracomplete logics are trivalued.

One of the most influential models of paracomplete logic is Kripke's *least-fixed point* model. Kripke uses the strong Kleene model: a pair $\langle D, I\rangle$ of a domain D (usually assumed to be non-empty) and interpretation function I. There are three valuations: $\{1, \frac{1}{2}, 0\}$. As in classical logic, a sentence has value 1 when it is true in the model and 0 when it is false. As for the $\frac{1}{2}$ value, its interpretation varies among theorists, and it is common to assign $\frac{1}{2}$ a value of 'neither true nor false'. Kripke, in [25], emphasizes that $\frac{1}{2}$ should be understood not as a third truth value, but only as indeterminate. This paracomplete logic takes 1 as the only designated value.

Another alternative to solve the Liar are paraconsistent logics. Like paracomplete logics, they also do not impose restrictions over the T-schema. Their models are trivalued, with 1 and $\frac{1}{2}$ being the designated values. For exponents of such lines, the Liar Paradox again shows us that the problem lies not in the Tarskian model of truth, but in classical reasoning: in paraconsistent logics, the *Ex Contradictione Quodlibet* (ECQ) is not a logical truth. Thus, $A, \neg A \nvdash B$ and, therefore, $A \wedge \neg A \nvdash B$. According to this view, contradictions are not necessarily problematic and do not lead to triviality.

In classical logic, proving a contradiction trivializes the system. The argument goes as follows: assume that A is a paradoxical Liar sentence. Thus, $T(\ulcorner A \urcorner) \wedge \neg T(\ulcorner A \urcorner)$ is a valid assertion. By T-schema, we are allowed to say that $A \wedge \neg A$ is also valid. Hence, A is valid. We are now authorized to assert the disjunction $A \vee B$. But we know that the sentence $\neg A$ is valid. Since the negation of one of the disjuncts is valid, we must assert the other disjunct: thus, B is valid, for any B. If a system proves any sentence B, it is a trivial system. What paraconsistent logics claim is that a contradiction does not trivialize the system, and so the Liar sentence does not cause any major problems. We can say that paraconsistent logics are defined negatively: any logics that *do not* accept ECQ are called paraconsistent.

The *dialetheist* approach, a logic that was mentioned several times in this text, is a type of paraconsistent logic that holds that there are true contradictions. In this alternative, the Liar sentence is seen as both true and false. It should

be stressed, however, that paraconsistent approaches do not require one to adopt a dialetheist approach, since, from the rejection of ECQ, no claim about truth necessarily follows; in particular, it does not necessarily follow that some contradictions are true. I will now list some objections to heterodox approaches and possible responses to defend them.

Let us begin with an unsatisfactory objection: one could argue that it is wrong to abandon classical logic, since it is the most successful system and the one that best expresses our way of reasoning. To this, a non-classical logician could easily reply that such a critique is akin to a petition of principle: well, non-classical approaches use paradoxes precisely to show that classical logic may not be the best option for reasoning; if the critique already starts from the maxim that classical logic is the best system, there seems to be no way to convince the partisan of such a view. A more prolific criticism points to the fact that, in order to build non-classical logics, one usually uses classical logic as a metalogic (in order for one to build the models of Kleene and Kripke, one employs set theory as a meta-theory). A classical logician might then ask: if heterodox logics are more appropriate, why do we usually use classical logics as a background?

Another pertinent criticism deals with the expressive power of paracomplete and dialetheist theories: how can such theories express disagreement? If I assert the Liar sentence, a paracomplete theorist will certainly not agree with me (he will not deem the sentence as true). It is unclear how he could express this: he cannot, of course, affirm the negation of the sentence, since he will also disagree with it. He can reject both options - but in that case, is he really expressing disagreement? The partisan of a dialetheist (glutty) theory, on the other hand, would have no such problem, since he would accept both the sentence itself and its negation. "Instead, the difficulty arises when you assert something they don't accept, such as "$2+2=5$." How can they express disagreement with that? Again, asserting the negation will not be enough" ([5], p. 59). Since the theory is based on accepting both the sentence and its negation, asserting the negation will not be understood as disagreement.

One possible answer to this would be to employ a different sense of disagreement than is commonly used: denying something is different from affirming the denial of that something. Thus, when a partisan of the paracomplete theory says that the Liar is not *either* false *or* true, he is not affirming that the sentence is neither, but rather denying that it is either. When, in turn, a partisan of the dialetheist theory says that it is *not* the case that all things are true, he *is not*

asserting that all things are not true, but *denying* that all things are.

In short, heterodox logics understand the Liar paradox as evidence of a flaw in classical reasoning. Therefore, we need a new logic that does not adopt one or another law of classical logic (ECQ; LEM; etc.), to escape the trivialization of the system and readjust it to our way of thinking. With this, we get to keep the full Tarskian T-schema, which, for non-classical logicians aligned with deflationism, is the complete expression of the concept of truth.

Classical Logic

To resolve the imbroglio generated by the Liar Paradox, some prefer to keep the classical logic as a panorama and modify some other element of sentence evaluation. Thus, many of these alternatives constrain the capture-and-release principles in different ways. Unlike deflationist approaches, the present routes in general do not identify in these principles the defining character of the concept of truth. Of course, this does not mean that classical approaches are opposed to them: they are useful and should be maintained, but not in every circumstance. In this case, what is most essential is to maintain the classical background.

Tarski's ([51], [52]) hierarchical approach fits into the picture drawn. In short terms, from the Liar sentence we conclude that "no language can contain its own truth predicate." If we go back to Proof 1 in section 2.1, we see that the error is not in any of the four steps, but in the fact that we define a truth predicate in the very language that expresses the sentence. Here is the way out: a language that does not contain its own truth predicate. But it is a fact that we need to assign values to the sentences of the language. Where do these values come from, if not from the language that contains the sentences? The answer lies in the hierarchy of languages and metalanguages: the predicate of a language $\mathfrak{L}_0$ is defined in its metalanguage, $\mathfrak{L}_1$. In section 4.7.2 of this text, I mentioned a current of thought that aims at transposing Tarski's hierarchy of formal languages to everyday languages, to argue that non-formal languages do not contain their own truth predicates. The idea here is exactly the same, but Tarski had only formal languages in mind.

More precisely, let's say that $\mathfrak{L}_0$ is to be interpreted by the standard model of arithmetic, $\mathbb{N}$. Since $\mathfrak{L}_0$ does not contain its truth predicate, we cannot tell in the language itself which sentences are true and which are false in the model. To do this, we select another language $\mathfrak{L}_1$, which contains $\mathfrak{L}_0$ and its truth predicate

– i.e. $\mathfrak{L}_1$ contains the truth predicate Tr_0 which applies only to sentences in $\mathfrak{L}_0$. Again, $\mathfrak{L}_1$ cannot contain its own truth predicate, so we select yet another language, $\mathfrak{L}_2$, which contains $\mathfrak{L}_1$ and Tr_1, which applies only to sentences in $\mathfrak{L}_1$. The order continues indefinitely, each language having its truth predicate defined always one level above itself. Finally, we may define compositional principles called CT-rules.

What prevents the Liar paradox is precisely the hierarchy of languages. This is because the predicate Tr_{n-1}, introduced in $\mathfrak{L}_n$, applies only to $\mathfrak{L}_{n-1}$, but not to $\mathfrak{L}_n$ itself. Thus, the capture and release principles only hold for Tr_{n-1} in $\mathfrak{L}_{n-1}$, and $Tr_{n-1}(\ulcorner A \urcorner)$ is always false if A is a sentence of $\mathfrak{L}_n$. Therefore, given the restriction of the capture-and-release principles, it is possible that, for a sentence A, we have A and $\neg Tr(\ulcorner A \urcorner)$ without a paradox arising from it, since we can evaluate A under the truth predicate that is in another language, and not the one containing A. In fact, the Liar sentence in the Tarskian hierarchy will be simply "$A \leftrightarrow \neg Tr(\ulcorner A \urcorner)$," when $A \in \mathfrak{L}_n$ and $Tr(\ulcorner A \urcorner)$ denotes the predicate Tr_m such that $n > m$. Since $n > m$, we have that Tr_m applies to a language that is one or more levels below $\mathfrak{L}_m$, then making it true to assert "$\neg Tr(\ulcorner A \urcorner)$," even if A is a valid sentence.

A first objection to the Tarskian hierarchy is that it does not address the concept of truth in natural languages. Kripke, in [25], follows this train of thought: for him, a solution to the Liar Paradox must also apply to everyday language, for it is an environment in which paradox might easily arise. We cannot assume that the speaker knows how to apply the truth predicate at the correct level to avoid the paradox – it seems unwise to require that, in uttering a sentence containing a truth predicate, he situate the predicate in the metalanguage of the language in which he speaks. Next, Kripke exposes a deep technical problem in the hierarchical theory, when applied to everyday languages as well. I explained this objection in 4.7.1; to refresh the reader's memory, here it goes again: if Dean says that *(a)* 'all of Nixon's statements are true,' this has to be stated at a level above everything Nixon says. On the other hand, Nixon himself states that *(b)* 'everything Dean says is true.' Well, for it to make sense, such a sentence also must be stated at a level above everything Dean says. Now, if both Dean's and Nixon's statements have to be above each other, it creates an unending hierarchical circle, which makes it impossible for situations like *(a)* and *(b)* to occur together.

A second point against the Tarskian hierarchy, as Richard Khirkam explains

in [24], is that it generates a new form of the Liar Paradox given by the following sentences:

(1) The following sentence is false-in-object-language.
(2) The previous sentence is true-in-metalanguage.

Given the powerful objections to Tarski's theory, it was necessary to think of alternatives that were less decisive with respect to self-referential sentences and that, as far as possible, avoided hierarchies. One of the alternatives that emerged in this picture was the proof theory approach: what it does is simply transform the CT-rules[28] into axioms and add them to the base theory, Peano Arithmetic. Again, we interpret $\mathfrak{L}$ by the standard model of arithmetic and add to it a truth predicate Tr, forming $\mathfrak{L}_+$. In this case there will be only one truth predicate, and it will apply to all the sentences of $\mathfrak{L}_+$ – so no more language levels are needed and Tr applies to the very language it belongs to; we finally have no hierarchy at all.

To strengthen the theory, one can add the rules of *necessitation* and *co-necessitation*. The first is expressed by $\frac{A}{Tr(\ulcorner A \urcorner)}$ and the other by $\frac{Tr(\ulcorner A \urcorner)}{A}$. Although, at first glance, they look like the capture and release principles, we must differentiate them: such rules are not principles, but *closure* conditions on theories. They are not principles because they depend on proofs: if there is a proof of A, then the rule allows us to prove $Tr(\ulcorner A \urcorner)$ and, conversely, if there is a proof of $Tr(\ulcorner A \urcorner)$, the rule allows us to prove A. With such an addition, the theory is called FS, Friedman-Sheard. One problem with this approach is that, while consistent, it is ω-inconsistent, "this means that there is an open sentence $A(x)$ such that FS proves $A(n)$ for every numeral n, and it also proves $\neg \forall x A(x)$" ([5], p. 73).

Another way of constructing a conception of truth based on proof theory is the KF, Kripke-Feferman theory. To construct it, the authors aimed at joining paracomplete dispositions to a classical setting. They wished to treat truth and falsity separately: in paracomplete logic, the $\frac{1}{2}$ value guarantees that the value 0 is different from the lack of the value 1. They needed to do something analogous in classical logic; however, because the latter is a bivalued logic, one must do so without the aid of the $\frac{1}{2}$ value. So we define 'falsity' – $F(\ulcorner A \urcorner)$ – as $Tr(\neg \ulcorner A \urcorner)$. This guarantees that $Tr(\ulcorner A \urcorner) \leftrightarrow \neg F(\ulcorner A \urcorner)$ does *not* hold, just as in the paracomplete theory.

[28] Rules formed from the usual compositional principles. For a more detailed exposition, see [5], p. 67.

There is, however, an important objection to the theory: with respect to the Liar, KF proves A and $\neg Tr(\ulcorner A \urcorner)$. This is precisely because of the definition of falsity described earlier. In this way, the theory succeeds in dodging the paradox: it is not valid that $Tr(\ulcorner A \urcorner) \wedge \neg Tr(\ulcorner A \urcorner)$, and from this B does not follow (the system is not trivialized). Still, that the theory proves A and $\neg Tr(\ulcorner A \urcorner)$ is controversial and unpleasant to classical intuitions.

Still in the classical setting, there is also a model-theoretic approach to solve the Liar Paradox, which is not axiomatic. Such an alternative aims at using the Kleene-Kripke model in classical logic. What makes the model non-classical is the Kleenean tri-valuation, and it is this that we will have to abandon if we want to build a classical Kleene-Kripke model. As with the alternative proof theory, we need to capture the effects of the heterodox $\frac{1}{2}$ value, but without really *using* it. To do this, we take the predicate Tr as partial: it *applies* to some sentences, does *not* apply to others, and *says nothing* about still others. We then define two sets: ξ, the extension of the predicate Tr, and ζ, the anti-extension of the predicate Tr as follows:

$\xi = \{(A) : I(A) = 1\}$
$\zeta = \{(A) : I(A) = 0\} \cup \{n : \neg sent_{\mathfrak{L}_+}(n)\}$

$(\mathbb{N}, (\xi, \zeta))$ is a partial model for $\mathfrak{L}_+$.

In this case, Liar sentences are neither in the set ξ nor in the set ζ; they are therefore in the gap between the two. Since a partial model is not exactly a classical model, one still has to work to transform it into one: Kripke's *closed-off* model abandons the set ζ and works with $(\mathbb{N}, \xi)$. This closes the gap between extension and anti-extension, since everything that would belong neither to ζ nor to ξ, comes to inhabit the category false'. Anyway, even in this case the Liar sentence A ends up in a kind of limbo, without belonging to either set. In this model, $\neg Tr(\ulcorner A \urcorner)$ will be true and so will $\neg Tr(\neg \ulcorner A \urcorner)$ be; thus, neither A nor $\neg A$ is true. The behavior of the Liar sentence in the closed-off model is subject to much criticism: since it is true that $\neg Tr(\ulcorner A \urcorner)$, by the nature of the sentence, this proves A (because A says that $\neg Tr(\ulcorner A \urcorner)$), then A turns out to be true, but the model itself told us that it is not. This anomalous behavior reveals that the closed-off construction is, in some sense, not well behaved.

Finally, the contextualist approach, the last classical alternative that will be examined, considers that the paradox reveals that sentences like the Liar depend on contexts, even if their components do not seem to depend on them. The idea,

basically, is that Liar sentences have no truth value. Since this is a classical (and therefore bivalent) theory, the sentences will not be assigned some other truth value; so when we say that they have no truth value at all, we should understand that they have no truth-bearing elements. If propositions are the truth-bearing components of sentences, then the Liar sentence does not express a proposition. From this perspective, the capture and release principles are restricted: they apply only to those things that qualify as truth-bearers, and since Liar sentences are not, then the principles do not apply to them.

According to contextualists, saying that the Liar is in the gap between truth and falsehood is an unstable proposition. The Liar, since it does not express a proposition, does not express a true proposition; but, since that is what the sentence says, it is true. First, we say that it has no truth value (or that it is in the gap, or at least that it does not have a "well-formed" truth value); from this, we prove that it is not true; since this is precisely what the sentence says, we prove that it is indeed true, and, hence, that it is not in the gap, but has a well-formed truth value after all. Again and by new means, we entangle ourselves in the paradox. The very status of the Liar as a 'defective sentence' is unstable, and this is, according to contextualists, the basic problem posed by such sentences. To answer it, they will have to turn to contexts: there must be some effect of context that makes the sentence, from one perspective, defective and, from another perspective, well-formed.

At last, a brief formal overview of the gears of contextualist theory. We add to the theory a predicate Tr_i, with the subscript 'i' marking the context: it influences how the predicate is interpreted in each situation. This is reminiscent of Tarski's hierarchy: each context can be seen as a level in the chain. Interestingly, we said earlier that one of the criticisms of the Tarskian model was that there would be no reason for truth to behave in levels. Contextualism offers, to some extent, a reason for this: each level represents a different context.

Overall, the contextualist theory has the advantage of trying to explain sentence instability, a major problem that many theories fail to address. But there are many fair objections to be made to such an approach: it is not clear in what way the Liar sentence depends on context. It is granted that it is possible to explain and formalize the problem in this way, but this does not guarantee that there is context-dependence indeed. Moreover, contextualism is not immune to objections to hierarchies in general. Finally, another kind of paradox seems to be formed in contextualist theory (if we assume that it is possible to quantify in

all contexts): "to maintain consistency, contextualists must apply constraints on quantifiers for quantifiers like "all contexts." To achieve this, presumably they must deny that there are contexts" ([5], p. 92).

Substructural Approaches

While the heterodox and classical approaches seek to solve the problem brought by the Liar by modifying either the truth predicate, the LEM, or the ECQ principle, and so on, the substructural approaches prefer to modify the deep structure of logic: the idea of a valid argument. Each alternative does this differently: among them, the *non-transitive* approach allows the notion of logical consequence to be non-transitive in certain cases, and the *non-contractive* approach allows it to be non-contractive in certain cases. To describe the constructions, we use sequent calculus, following Gentzen's [18] formalization.[29]

More specifically, the non-transitive alternative guarantees the possibility that, having the formulas A, B, and C such that A implies B and B implies C, it is not necessarily the case that A implies C. In practice, the approach rejects the cut rule of the calculus of sequents, which is described as follows:

$$\frac{\Gamma \vdash A,\ \Delta \qquad \Gamma',\ A \vdash \Delta'}{\Gamma,\ \Gamma' \vdash \Delta,\ \Delta'}$$

This rule tells us that if we derive a sequent that contains A as a conclusion and another that contains A as a premise, we can "cut" the A's and join the two sequents. As for the non-contractive alternative, it allows that there are A and B formulas for which, if two occurrences of A prove B, it is not the case that one occurrence of A proves B. For this, it rejects the *contraction rules* (WL and WR), described as

$$\textbf{WL: } \frac{\Gamma,\ A,\ A\colon \Delta}{\Gamma,\ A\colon \Delta} \qquad\qquad \textbf{WR: } \frac{\Gamma\colon A,\ A,\ \Delta}{\Gamma\colon A,\ \Delta}$$

Contraction rules make it possible for us to collapse more than one occurrence of a formula into just one. By abolishing such rules, therefore, each of the approaches can restrict either transitivity or contraction.

At first glance, it may seem strange to abolish such basic rules of reasoning. What would be the advantage of doing so? Well, the first big advantage is that

[29] Weir [55] warns us, however, that the boundaries between operational and substructural logic rules are not at all clear.

we can unite two things that, in the approaches we have studied so far, could not be united: classical logic and unrestricted applications of the T-schema. This is what the *non-transitivist* Alan Weir suggests, in [55]. He proposes a program named neoclassical logic, which consists of restricting cut to maintain a naive conception of truth (the deflationist conception).

With respect to the Liar sentence, these alternatives behave as follows: in the case of the non-transitive one, it is possible to derive $\vdash A$ and $A \vdash$ from the system, i.e., the sentence is proved from anything (hence, it is theorem), and anything follows from it (hence, it is refutable). Since there is no cut rule, this does not lead to the problems it would normally lead to in a classical situation. In the case of the non-contractive alternative, it is possible to derive $\vdash A, A$ and $A, A \vdash$. Since we cannot collapse the occurrences into a single A, the Liar is exactly neither theorem nor refutable: only the pair of the sentence with itself implies anything and follows from anything. Since the sentence alone cannot be contracted, we can accurately say that, on this theory, the Liar ceases to be a paradox.

Besides the advantage I have described, substructural approaches also have the following: there is an idea (of an almost intuitive glimmer) that semantic paradoxes have something in common in their structure. The substructuralists provide a reason for this – according to non-transitivists, what they have in common is the false assumption that the cut rule holds for all cases; according to non-contractivists, it is the false assumption that the contraction rule always holds. Another advantage is that, in substructural approaches, we can say that valid arguments preserve truth and are expressed in the object-language itself.

Perhaps the most important objection to substructuralists is with respect to the nature of logical consequence: this being, for many, the *defining* factor of logic (it is very common to say that logic is that which deals with the consequence relation), both non-transitivists and non-contractivists have to answer what it is, if it does not obey the rule of cut or contraction? On the one hand, the consequence relation is commonly considered transitive, and this is a useful feature for derivations; on the other hand, it is difficult to conceive of consequence without contraction, for "how can one occurrence of A be true without all other occurrences also being true?" ([5], p. 103).

Another important problem with substructuralist solutions is the possible flaw they introduce in cumulative reasoning. By this term, Beall, Glanzberg and Ripley mean the kind of basic reasoning that works as follows: From a set of premises, we

arrive at a conclusion ([5], p. 105). We add the conclusion to the set of premises, and from this new larger set we arrive at other conclusions. Finally, we say that the final conclusion follows from the original premises. Intuitively, the reasoning is perfectly valid (provided, of course, that the conclusions really derive from the premises). However, for both the non-transitivist and the non-contractivist approaches, this appears as a problem. For the former, the problem is that if A can be validly concluded, it does not follow that A can be introduced as a premise (the argument for transitivity assumes that this is possible; since this approach rejects transitivity, it must also reject cumulative reasoning). In the case of the non-contractivist, the problem is different: if we derive a conclusion from a set of premises, and use that conclusion alongside the premises to derive a new conclusion, we end up using the original premises twice – but that is precisely what non-contractivists are opposed to.

Thus, substructural approaches must overcome a great challenge: convincing logicians that they should be adopted even without the tool of cumulative reasoning. To do this, the non-transitivist logician may point to all the arguments that can be derived without cut – that basically form the whole of classical logic. The non-contractivist logician, on the other hand, cannot resort to all classical logic since, without the contraction rule, some classical arguments cannot be derived. In particular, the LEM and the ECQ cannot be derived in their full form. With a non-contractivist logic, we can only derive $A \wedge \neg A, A \wedge \neg A : B$ (the "double" ECQ principle) and $B : A \vee \neg A, A \vee \neg A$ (the "double" LEM). In view of such difficulty, partisans of this approach offer a variation of cumulative reasoning that excludes premises that have already been used, but allows valid conclusions to be added to the premises (see [5], p. 107).

Synthetically, substructural approaches solve the Liar Paradox through changes in the substructure of logical reasoning, a layer deeper than the one in which axioms dwell. Two such approaches are the non-transitivist and the non-contractivist ones: according to the former, the idea that the rule of cut can be applied in all contexts is not valid; for the latter, it is the contraction rule that is not valid for any and all contexts. Such solutions, although they force us to abandon or modify tools as useful as cumulative reasoning, have the advantage of making it possible to unite classical logic with the unrestricted T-schema.

Bibliography

[1] Aczel, Peter. Introduction to Inductive Definitions in J. Barwise, editor, *Handbook of Mathematical Logic,* p. 739-782. 1997.

[2] Aczel, Peter. *Non-well-founded Sets.* CSLI, 1988.

[3] Aristotle. *Metaphysics*. Hackett Publishing Company, 2016. Translation by C.D.C. Reeve. Original written sometime between 370 and 322 BC.

[4] Beall, J. C. Is Yablo's paradox non-circular?. *Analysis*, v. 61, n. 3, p. 176-187, 2001.

[5] Beall, J. C.; Glanzberg, Michael; Ripley, David. *Formal Theories of Truth.* Oxford University Press, 2018.

[6] Burgess, Alexis; Burgess, John P. *Truth*. Princeton Foundation of Contemporary Philosoph, 2001.

[7] Bolander, Thomas, Self-Reference, The Stanford Encyclopedia of Philosophy (Fall 2017 Edition), Edward N. Zalta (ed.), URL = ¡https://plato.stanford.edu/archives/fall2017/entries/self-reference/¿.

[8] Berto, Francesco; Nolan, Daniel. *Hyperintensionality*, The Stanford Encyclopedia of Philosophy (Summer 2021 Edition), Edward N. Zalta (ed.), Link: click here.

[9] Carroll, Lewis. *Alice in wonderland*. Children's Press, 1947.

[10] Carroll, Lewis. *Through the looking glass and what Alice found there*. Penguin UK, 2010. Original written in 1987.

[11] Cook, Roy. *The Yablo Paradox: An Essay on Circularity*, Oxford, 2014.

[12] Davidson, Donald. Truth and Meaning. *Synthese* Vol. 17, number 3: *On saying that.* Pp. 304–323, 1967.

[13] Deleuze, Gilles. *Différence et Répétition*. Presses Universitaires de France, 1985 (written in 1968).

[14] Deleuze, Gilles. *Logique du Sens*. Presses Universitaires de France, 1969.

[15] Dorandi, Tiziano (Ed.). Diogenes Laertius: *Lives of Eminent Philosophers.* Cambridge University Press, 2013.

[16] Frege, Gottlob. On Sense and Reference. *The philosophical review*, v. 57, n. 3, p. 209-230, 1948.

[17] Field, Hartry. A revenge-immune solution to the semantic paradoxes. *Journal of Philosophical Logic*, v. 32, p. 139-177, 2003.

[18] Gentzen, Gerhard. Investigations into logical deduction. *American philosophical quarterly*, v. 1, n. 4, p. 288-306, 1964.

[19] Herzberger, Hans G. The Truth-Conditional Consistency of Natural Languages. *The journal of philosophy,* p. 29-35, 1967.

[20] Hodel, Richard E. *An introduction to mathematical logic*. Courier Corporation, 1995.

[21] Horsten, Leon. *The Tarskian turn:* Deflationism and axiomatic truth. MIT press, 2011.

[22] Huggett, Nick, 'Zeno's Paradoxes", The Stanford Encyclopedia of Philosophy (Winter 2019 Edition), Edward N. Zalta (ed.), Link: Click here.

[23] Ketland, Jeffrey. Yablo's Paradox and ω-inconsistency. *Synthese*, v. 145, n. 3, p. 295-302, 2005.

[24] Kirham, Richard L. *Theories of truth*: A critical introduction. MIT Press, 1992.

[25] Kripke, Saul. Outline of a Theory of Truth. *The journal of philosophy*, v. 72, n. 19, p. 690-716, 1976.

[26] Koyré, Alexandre. Épiménide, Le Menteur. In: *Histoire de la Pensée*. Hermann et Cie Editeurs, 1947.

[27] Leitgeb, Hannes. What is a Self-Referential Sentence? Critical remarks on the alleged (non-) circularity of Yablo's paradox. *Logique et Analyse*, v. 45, n. 177/178, p. 3-14, 2002.

[28] Martin, Robert L. Are natural languages universal?. *Synthese*, v. 32, p. 271-291, 1976.

[29] Miller, David. Russell, Tarski, Gödel: um Guia de Estudos. *Ciência e Filosofia* n. 5, p. 67-105, 1996.

[30] Mualem, Shlomy. Nonsense and irony: Wittgenstein's strategy of self-refutation and Kierkegaard's concept of indirect communication. *Tópicos (México)*, n. 53, p. 203-227, 2017.

[31] Moss, Lawrence S., Non-wellfounded Set Theory, *The Stanford Encyclopedia of Philosophy* (Summer 2018 Edition), Edward N. Zalta (ed.), link: Click Here.

[32] Parmenides. On Nature. In: John Burnet's *Early greek philosophy*. London, AC Black, 1908; original text from around 475 B.C.

[33] Plato. Meno. In: *Laches, Protagoras, Meno and Euthydemus*. Loeb Classical Library; Harvard University Press, 1952. Original written around 387 B.C.

[34] Picollo, Lavinia. Alethic Reference. *Journal of Philosophical Logic*, v. 49, n. 3, p. 417-438, 2020.

[35] Priest, Graham. Semantic Closure. *Studia Logica*, v. 43, n. 1-2, p. 117-129, 1984.

[36] Priest, Graham. Yablo's Paradox. *Analysis*, v. 57, n. 4, p. 236-242, 1997.

[37] Priest, Graham. Inconsistent models of arithmetic part I: Finite models. *Journal of Philosophical Logic*, p. 223-235, 1997.

[38] Priest, Graham. Inconsistent models of arithmetic part II: The general case. *The Journal of Symbolic Logic*, v. 65, n. 4, p. 1519-1529, 2000.

[39] Priest, Graham. *Beyond the limits of thought.* Oxford University Press, 2002.

[40] Priest, Graham. *In Contradiction*. Oxford University Press, 2006.

[41] Pleitz, Martin. *Logic, language, and the Liar paradox*. Brill Mentis, 2018.

[42] Quine, Willard Van Orman. *Word and object.*MIT press, 2013 (original text from 1960).

[43] Quine, Willard V. The ways of paradox. In: *The Ways of Paradox and Other Essays,* pp. 3 - 20, Random House, New York, 1966.

[44] Rescher, Nicholas. *Paradoxes:* Their Roots, Range, and Resolution. Studia Logica, v. 76, n. 1, 2004.

[45] Rescher, Nicholas. *Aporetics: Rational deliberation in the face of inconsistency.* University of Pittsburgh Press, 2009.

[46] Russell, Bertrand. Mathematical logic as based on the theory of types. *American journal of mathematics*, v. 30, n. 3, p. 222-262, 1908.

[47] Russell, Bertrand. *An Inquiry Into Meaning and Truth.* George Allen and Unwin LTD, 1940.

[48] Smullyan, R. *Diagonalization and Self-Reference*, Oxford Logic Guides, 1994.

[49] Shoenfield, Joseph R. *Mathematical Logic.* Addison-Wesley Publishing Company, 1967.

[50] Soames, Scott et al. *Understanding Truth.* Oxford University Press on Demand, 1999.

[51] Tarski, Alfred. The Concept of Truth in Formalized Languages. *Logic, Semantics, Metamathematics*, v. 2, n. 152-278, p. 152, 1936.

[52] Tarski, Alfred. The semantic conception of truth: and the foundations of semantics. *Philosophy and phenomenological research,* v. 4, n. 3, p. 341-376, 1944.

[53] Urbaniak, Rafał. Leitgeb, 'About," Yablo. *Logique et Analyse*, p. 239-254, 2009.

[54] Wittgenstein, Ludwig. *Tractatus Logico-Philosophicus.* Routledge, 1974 (original written in 1921). Translated by D. F. Pears and B. F. McGuinness.

[55] Weir, Alan. Naive truth and sophisticated logic. *Deflationism and Paradox*, ed. por JC Beall e Bradley Armour-Garb. p. 218-249, 2005.

[56] Yablo, S. Paradox Without Self-Reference. *Analysis*, v. 53, n. 4, 1993.

[57] Yablo, Stephen. Truth and Reflection. *Journal of Philosophical Logic,* v. 14, n. 3, p. 297-349, 1985.

[58] Yablo, S. Paradox without self-reference. *Analysis*, v. 53, n. 4, 1993.