THE PARADOXES OF SELF-NEGATION

MSc Thesis *(Afstudeerscriptie)*

written by

Albert Janzen
(born February 27th, 1989 in Sibirskij, Russia)

under the supervision of **Prof. Francesco Berto** and **Dr. Luca Incurvati**, and submitted to the Board of Examiners in partial fulfillment of the requirements for the degree of

MSc of Logic

at the *Universiteit van Amsterdam*.

Date of the public defense:
November 27th, 2017

Members of the Thesis Committee:
Dr. Maria Aloni
Prof. Francesco Berto
Dr. Peter Hawke
Dr. Luca Incurvati
Dr. Katrin Schulz

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

Abstract

In *Beyond the Limits of Thought* Graham Priest presents the Inclosure Schema as the underlying structure of the paradoxes of self-reference. I argue that while the paradoxes fit the Inclosure Schema, (i) in case of Burali-Forti, Mirimanoff, 5th Antinomy, Richard, König and Berry one premise of the schema is not true and (ii) in case of the Liar, Russell and Grelling the schema does not capture what is essential to the paradoxes. For the Liar, Russell and Grelling I construct a new schema that reveals their highly analogous structure and proof of contradiction. At the heart of this schema is the notion of self-negation, a statement of the form A iff $\neg A$. By classical logic, self-negation always leads to contradiction and therefore describes a structure that can be found only in paradoxical cases. In this respect, the schema describes what is essential to the paradoxes. As a result, if one were to give up self-negation to solve the paradoxes one would not give up non-paradoxical cases. Finally, I provide an analysis of No-no and Yablo that also involves self-negation. Based on this analysis, I establish a second schema involving self-negation that captures No-no and Yablo.

Acknowledgements

This thesis is a first attempt to bring together ideas about paradoxes that I have entertained since I first encouterned paradoxes. The Master of Logic has not only given me the theoretical background and methodology to evaluate my ideas but has also shown me that these ideas are still new and need to be written down. I am filled with gratitude for my two supervisors Franz Berto and Luca Incurvati who have put so much time and effort in evaluating and improving my ideas. I learned a lot from them and this thesis would not have reached its current state without their help. I want to thank my friends Richard, Katharina, Julian, Melf, Anna, Laura and all the ones I am forgetting now for listening to my ideas for hours and offering their smart brains to give me the critical feedback I needed. Finally, I want to thank my wonderful family that supports me in everything! Thank you all so much!

Contents

Chapter 1

Introduction

1.1 What is a paradox?

A paradox is a statement derived by apparently sound reasoning from apparently true premises which states something that goes against what is commonly believed. A paradox is therefore a statement A that satisfies the following two conditions:

(a) It seems rational not to accept A.

(b) It seems rational to accept the argument for A.

What counts as rational can surely be debated. For the paradoxes mentioned in this thesis, however, the reasons for not accepting A but the argument for A will be made explicit. To solve a paradox A, one has to either deny (a) or deny (b). The best way to achieve that is to either show that it is rational to accept A or that there is a flaw in the argument for A, respectively. However, both options are difficult to argue for as they challenge our intuitions that lead to accept (a) and (b). This problem becomes particularly apparent with the so called paradoxes of self-reference.

In this group of paradoxes A is a plain contradiction, i.e. a statement of the form 'a and $\neg a$'. Accepting such a statement seems absurd since the rule of explosion (EXP) tells us that anything can be derived from a contradiction. To deny (b) is not much less of a challenge as the argument for contradictions in this group of paradoxes involves merely fundamental concepts and logical rules. The oldest example of this group of paradoxes is the Liar, a statement that says of itself that it is false. Given certain intuitive assumptions, it can be shown that it is both true and false. Let us have a look at the argument to see what we need to give up in order to avoid contradiction. Let $\mathcal{L}$ be a language such that the following premises and inference rules are given:

(P1) We can form sentences in $\mathcal{L}$ that express that sentences of $\mathcal{L}$ are true and we can form sentences in $\mathcal{L}$ that express that sentences of $\mathcal{L}$ are false.

(P2) We can form sentences in $\mathcal{L}$ that contain their own name.

(P3) Law of Excluded Middle (LEM): Let L be a sentence in $\mathcal{L}$. The claim that L is true or L is false is a logical truth.

(I1) T-schema: Every sentence in $\mathcal{L}$ is true iff what it states is the case.

(I2) Reasoning by Cases: Let Γ be a set of sentences in $\mathcal{L}$, and A, B and C sentences in $\mathcal{L}$. If Γ plus A imply C and Γ plus B imply C as well, then Γ plus the claim that A or B imply C.

This list of premises and inference rules is not exhaustive but seems to be sufficient for understanding the argument. It further highlights some of the premises and inference rules that have been discussed the most in order to solve the Liar, e.g. (P3). The argument goes as follows.

(1) By (P1), let 'L is false' be a sentence in $\mathcal{L}$ where L is a sentence of $\mathcal{L}$.

(2) By (P2), let L be the name of 'L is false'.

(3) By (P3), the claim that L is true or L is false is a logical truth.

(4) By (I1), L is true iff L is false.

(5) The claim that L is true iff L is false and the claim that L is true together imply the claim that L is true and L is false.

(6) The claim that L is true iff L is false and the claim that L is false together imply the claim that L is true and L is false.

(7) By (5), (6) and (I2), the claim that L is true iff L is false and the claim that L is true or L is false together imply the claim that L is true and L is false.

(8) By (3), (4) and (7), L is true and L is false.

To deny any of the premises or inference rules involved in this argument comes with the task to either dispose of them or replace them with acceptable alternatives.

1.2 How can a paradox be solved?

In case of the Liar, there are four main solution routes (Field 2008):

(I) Give up or replace the notion of truth

(II) Prevent the construction of the Liar sentence

(III) Restrict the T-schema

(IV) Restrict classical logic

(I)-(III) represent solutions to the Liar that deny (b), i.e. the argument of contradiction is not accepted. Alfred Tarski (Tarski, 1935) proposed that truth of statements in a language can only be stated in a metalanguage. This solution, however, is not applicable to natural language and therefore does not qualify as a version of (I). Due to its importance in the discussion of how to solve the liar paradox it is still worth mentioning. Initially, Tarski defined the T-schema for first-order arithmetic. In his *Undefinability Theorem* he proved that any theory extending arithmetic which contains the T-schema is inconsistent. The crucial part of the proof for this theorem is to formalise the liar sentence within the formal theory to derive a contradiction. Arithmetic truth is thus arithmetically not definable. Instead of giving up the T-schema, Tarski develops what is known as *Tarski's hierarchy of languages* where no language contains its own truth predicate. Each language of the hierarchy of languages, $L_0, L_1, L_2, \ldots$, contains a truth predicate for sentences at lower levels, i.e. for each L_{i+1} there is a truth predicate T_{i+1} applicable to sentences of L_j where $j \leq i$. In this hierarchy one cannot form the liar sentence since it contains a truth predicate that applies to the liar sentence itself. However, the expressive limitations of this hierarchy are subject to many criticisms. As we will see in this section, one cannot, for instance, formulate claims within the hierarchy that are not paradoxical.

Solution route (II) demands a dismissal of *all* forms of self-reference that can be used to construct the Liar. That means giving up sentences that contain their own name will only prevent the Liar sentence as stated above. As we will see in chapter 2, other ways of achieving self-reference in the liar cannot be given up without severe expressive limitations of the language. Even if we find a way to restrict the expressive capacities of the language to prevent self-referential constructions as in the liar, other paradoxes involving indirect self-reference like the liar chain[1] remain unsolved.

Many philosophers including Field (2008) and Priest (1987) deny solution route (III) due to the importance of the T-schema for the meaning of truth. For instance, Priest takes it to be necessary if we want to endorse the words of another, especially if we do not know what is said. 'For suppose the Pope utters α. We would like to assert α' (the Pope will, of course, speak in Latin), but we cannot. Instead, we form a noun phrase which refers to α, i.e. α, and assert $T\alpha$. Clearly, this construction will fail if $T\alpha$ does not imply α' or vice versa. In other words, the T-schema holds' (Priest, 1987, 55). For Field the truth predicate is not just a means of expressing agreement or disagreement. Since it is used inside conditionals it should be intersubstitutable with the statement it is applied to, i.e. 'if A is true then B' should be equivalent to 'if A then B'. More generally, Field endorses what he calls the Intersubstitutivity Principle (IP):

[1] In this paradox, two sentences refer of each other: A='B is true' and B='A is false'. A thus indirectly refers to itself as it refers to B which in turn refers to A. The same holds for B. Contradiction arises as one can show that A is true iff B is true iff A is false. It follows that A is true iff A is false (or, likewise, B is true iff B is false) which is analogous to step (4) of the argument of the liar and therefore leads to contradiction.

> If C and D are alike except that (in some transparent context) one has a sentence ‘A’ where the other has ‘it is true that A’, then one can legitimately infer D from C and C from D.

IP together with the validity of $A \rightarrow A$, which Field endorses, implies the T-schema. Moreover, Field is a deflationist on truth, i.e. he supports the view the T-schema captures everything significant that one can say about truth. Priest, on the other hand, neither endorses IP nor is he a deflationist on truth.

Restricting classical logic, solution route (IV), has dominated the discussion about how to solve the liar paradox. It may consist of a denial of (b), e.g. by giving up LEM[2], or a denial of (a), e.g. by giving up EXP. The former falls in the category of paracomplete solutions as it allows for some sentences to be neither true nor false. The latter is labeled paraconsistent as according to this solution some sentences are both true and false.

1.3 A paracomplete solution

In *Outline of a Theory of Truth* (1975) Kripke presents a truth theory that characterizes the Liar as a semantically pathological case. Any truth theory, he argues, has to cancel out such pathological cases by means other than syntactical rules. For the Contingent Liar shows that a Liar paradox can arise only by virtue of unfavorable empirical conditions. Consider the following two assertions:

(1) Jonas: Most of Nixon’s assertions about Watergate are false.

(2) Nixon: All of Jonas’ assertions about Watergate are true.

Now, suppose that the following unfavorable empirical conditions are true:

(a) (1) is Jonas’ only assertion about Watergate

(b) Apart from (2) half of Nixon’s assertions about Watergate are true and the other half are false

Now, if (1) is true then (2) must be false so that most of Nixon’s assertions about Watergate are false. But then it is not the case that all of Jonas’ assertions about Watergate are true, i.e. (1), so (1) is false. On the other hand, if (1) is false then (2) must be true so that it is not the case that most of Nixon’s assertions about Watergate are false. But (2) being true means that (1) is true after all. So given empirical assumptions (a) and (b) we have a statement (1) such that (1) is true iff it is false, a so called Contingent Liar. The same holds for (2) which is implicit in the reasoning above. The point of this example is to show that any solution to the Liar that involves an alteration of syntactical rules,

[2] One might instead give up the semantic version of LEM, called the Law of Bivalence. LEM states that $A \vee \neg A$ is logically valid while Bivalence states that every sentence is either true of false. There are semantics in which LEM and the Law of Bivalence do not coincide. For instance, supervaluationists like Shapiro (2003), respond to vagueness by leaving LEM logically valid but not endorsing that every sentence is either true or false.

e.g. Tarski's hierarchy of languages, either fails to cancel out all problematic cases or cancels out unproblematic cases. For the Contingent Liars (1) and (2) are contradictory only when the empirical conditions (a) and (b) are met. A syntactic sieve, however, would cancel out (1) and (2) regardless of whether (a) and (b) are met. In Tarski's hierarchy, (1) and (2) cannot even be formulated as one cannot consistently assign a level to them: Since (1) is a statement about the truth of Nixon's utterances, including (2), (1) would have to be on a higher level than (2). Yet, (2) is a statement about the truth of (1) and therefore (2) would have to be on a higher level than (1) which is not possible. Now, in case the conditions (a) and (b) are not met, (1) and (2) still cannot be formulated although there is no problem with them. On the other hand, if a syntactic sieve did not declare (1) and (2) ill-formed, it would not cancel out the problematic case in which (a) and (b) are met.

Kripke proposes a hierarchy of interpretations rather than one of languages. He starts with a first order language L to which he adds a monadic truth predicate T which gives us $\mathcal{L}$. Truth of sentences in this language is defined relative to the extension of the truth predicate, i.e. the class of sentences that satisfy T, and the extension of the truth predicate is different at each level of the hierarchy. At the first level $\mathcal{L}_0$, truth is not defined for sentences that contain the truth predicate. At this level the extension of the truth predicate is empty. At the next level the extension of the truth predicate contains all sentences that are true relative to the extension of the truth predicate at the first level. So at $\mathcal{L}_1$, '$T(\phi)$' is true when ϕ is a true sentence at $\mathcal{L}_0$[3]. '$T(T(\phi))$' then is true at level $\mathcal{L}_2$ given that $T(\phi))$ is a true sentence at $\mathcal{L}_1$ and so on. At each step of this procedure more and more sentences that contain the truth predicate are declared true. At some unique (infinite) level of the hierarchy, $\mathcal{L}_\sigma$, Kripke shows that the interpretation of the truth predicate cannot be extended further, i.e. no more new sentences satisfy T in the next levels. T is now a truth predicate for its own level, i.e. a sentence ϕ is true at $\mathcal{L}_\sigma$ iff $T(\phi)$ is true at $\mathcal{L}_\sigma$. A sentence is called *grounded* if it has a truth value at $\mathcal{L}_\sigma$. Kripke shows that the Liar sentence is not grounded and therefore has no truth value. It is neither true nor false.

Kripke noted himself that his language is too weak to express that sentences like the liar are ungrounded. To overcome this weakness by adding a predicate expressing that a sentence is ungrounded comes with a new challenge. One could then formulate the so called Strenghtened Liar:

(SL) This sentence is not true.

If SL is true then what its says is true and thus SL is not true which is a contradiction. So it is not true. That means it is either false or ungrounded, i.e. neither true nor false. In both cases it is not true which is precisely what SL states. So SL is true after all which is a contradiction. Thus SL is true and

[3] For '$T(\phi)$' to be a sentence of the formal language one needs to show that it is possible to refer to statements of that language within that language. One way to achieve this is by Gödel numbering which will be introduced in the next chapter.

not true. If our language is capable of expressing SL it would not be free of contradictions after all.

Laurence Goldstein offers a solution to this *revenge problem* following Strawson's view that '[w]e cannot talk of *the sentence* being true or false, but only of its being used to make a true or false assertion' (Strawson, 1950):

> A sentence may be meaningful or significant but it does not say or state anything and *a fortiori* does not say anything true and does not say anything false. It is statements (what sentences are used to make) that are true or false. [For reasons like the one given by Kripke] sentences in the Liar family, while perfectly meaningful, *cannot* be used to make statements, cannot be used to say anything true or anything false (Goldstein, 2009, 382).

Hence one cannot conclude from the assumption that SL has no truth value that it is not true. The fact that it has no truth value means that it fails to make any statement and therefore fails to state of itself that it is not true.

Another notable feature of Kripke's theory is that it does not validate the T-schema but the slightly weaker Intersubstitutivity Principle. To get the T-Schema one needs to add to IP the validity of $A \rightarrow A$. Kripke's theory, however, is based on a K_3 logic, a logic in which $\neg A \vee A$ is not valid, with a conditional that behaves like the material one. With such a conditional $A \rightarrow A$ is equivalent to $\neg A \vee A$ and therefore $A \rightarrow A$ is not valid in Kripke's theory. In *Saving Truth from Paradox* Field 2008 presents a formal language based on a K_3 logic with its own truth predicate that not only validates the Intersubstitutivity Principle but also the T-schema by showing how to define a conditional that respects many classical truths such as $A \rightarrow A$. Further, in the language one can express that certain sentences such as the Liar are defective, i.e. one can say of such a sentence that it is neither determinately true nor determinately false. Field even claims to avoid the Strengthened Liar and other paradoxes such as Grelling's and paradoxes of definability like Berry's and König's[4]. Yet, among other points of criticisms, Priest (2010) has argued that in Field's language one cannot both express that certain sentences like the Liar are defective and talk about such sentences in general. The latter, according to Priest, is necessary for formulating the essential idea of Field's solution. Further, he argues that Field fails in showing that the paradoxes of definability are avoided.

1.4 A paraconsistent solution

Graham Priest (1987) and JC Beall (2009) are the most prominent defenders of dialetheism, the view that there are true contradictions (dialetheia). According to this view, contradictions do not imply everything. The problem of the Liar is thus solved by denying the rule of explosion, $\{A, \neg A\} \vDash B$. Yet, there are paradoxes which are not solved after giving up EXP. In the Curry paradox,

[4] These paradoxes will be introduced in chapter 3.

an arbitrary statement (and thus everything) can be derived without EXP but rather by using the principle of absorption (or Contraction) $A \rightarrow (A \rightarrow B) \vDash A \rightarrow B$ (Priest, 1987, 83-84). The logics underlying dialetheism are called paraconsistent. To deal with the Curry paradox such logics can be developed with a conditional that does not validate absorption (Priest, 1987; Beall, 2009).

Both dialetheic theories take their motivation from the plausibility of the arguments for contradiction. Recall that Priest argues that we need the truth predicate to express agreement or disagreement to what is said especially when we do not know what is said and therefore endorses the T-schema. In Priest's theory the truth predicate does not satisfy IP whereas this is the case in Beall's theory. The T-schema then follows from IP and the validity of $A \rightarrow A$. Further, dialetheists take the Strengthened Liar to show that self-reference is unavoidable and that hierarchical interpretations of truth like the one proposed by Kripke fail. The Liar paradox thus provides evidence that contradictions are provable.

The Strenghtened Liar further shows that non-dialetheic solutions to the Liar suffer from limitations that Armour-Garb and Woodbridge call the Dialetheic Conjecture. Any consistentist, i.e. non-dialetheic, solution to the Liar has one of the following four drawbacks (Armour-Garb and Woodbridge, 2006, 400):

(i) The consistentist's response to the liar paradox is too narrow, in the sense that, even if she responds to one version of the paradox, she seems powerless to respond to others (e.g., revenge problems).

(ii) The consistentist's response to the liar paradox is ad hoc; it may respond to all versions of the liar paradox, but in an unprincipled way (e.g., via imposing Tarskian hierarchies or claiming liar sentences to be meaningless).

(iii) The consistentist's response to the liar paradox is neither too narrow to respond to all versions of the liar paradox nor ad hoc in its responses to those paradoxes, but it fails to respond to semantic pathology, generally (e.g., Berry's, Curry's, Grelling's, etc.).

(iv) The consistentist's response to the semantic paradoxes is neither too narrow nor ad hoc, but it restricts, unduly (and, thus, unacceptably), the expressive capacities of the language in question.

In addition, Priest's PUS demands a unified solution for the paradoxes as they all have the same underlying structure and therefore denies solutions that apply to one paradox but not another. However, the underlying structure he ascribes to the paradoxes has been challenged and one major goal of the present thesis it to extend the criticism it has received.

1.5 The goal of the thesis

In this thesis I will not deal with the question whether a paracomplete or paraconsistent theory is the appropriate way to respond to the paradoxes of self-reference. Rather, I deal with the question of the correct analysis of the paradoxes. In particular, I present an analysis for some paradoxes that does not take self-reference at its core. So far, several notions of self-reference have been ascribed to the paradoxes. In addition, Stephen Yablo (1993) presents a paradox that challenges not only the notions of self-reference that are applied to the paradoxes but also the claim that self-reference is necessary to derive a contradiction. Even worse, Hannes Leitgeb (2002) points out that none of the notions of self-reference that are applied to the paradoxes are satisfactory.

The goal of this thesis is to present a notion that replaces the notion of self-reference in describing what is crucial to a group of paradoxes of self-reference. It is the notion of self-negation and it describes a structure detectable in the Liar, Russell and Grelling. A sentence A negates itself iff it is the case that A iff $\neg A$. I further present a new schema, the schema of self-negation, which shows how a self-negating sentence and the resulting contradiction in the three paradoxes can be derived uniformly.

The main three claims about self-negation that I want to show are:

(i) The Inclosure Schema does not capture what is essential to all paradoxes of self-reference. In particular, it does not capture the essentials of the Liar, Russell and Grelling which is the objective of the schema of self-negation. The other paradoxes of self-reference are well captured by the Inclosure Schema. However, in most of these paradoxes one premise of the Inclosure Schema is not true[5].

(ii) Self-negation can only be found in paradoxical cases[6]. In this sense, self-negation is essential to the paradoxes. Giving up self-negation would solve the paradoxes without giving up unproblematic cases.

(iii) Self-negation is involved not only in the three paradoxes above but also in a paradox whose self-referential status is highly controversial, Yablo, and in a paradox that is not captured by the Inclosure Schema, No-no.

I proceed as follows. First, in chapter **2**, I discuss the notions of self-reference available in the literature. In chapter **3** I present the Inclosure Schema, in particular, how it applies to the paradoxes and which notion of self-reference it ascribes to them. I will then argue that in most cases either (a) the Inclosure Schema does not capture what is essential to the paradox or (b) that one premise of the paradox, as presented in the Inclosure Schema, is false and therefore the derivation of contradiction is blocked. Chapter **4** is the heart of this work in which I present a schema for those paradoxes that fall under (a),

[5] The argument for the claim that one premise of the Inclosure Schema is not true applies to all paradoxes but the Knower, Gödel and Berkeley. These paradoxes are left for further studies.

[6] We will see that this holds only when classical logic is to be maintained. According to non-classical solutions, self-negation does not logically imply a contradiction.

the Schema of Self-Negation. In this schema a statement of the form 'A iff $\neg A$' can be derived in a uniform way that leads to the contradiction 'A and $\neg A$'. In chapter **5** and **6**, I show how self-negation can be found in the No-no paradox and Yablo's paradox and present a second schema involving self-negation that captures the two paradoxes. Chapter **7** concludes the present work with a recap of the results and an outlook on how to proceed with them.

Chapter 2

What is Self-Reference?

Since it is the underlying structure of the paradoxes of self-reference that is under investigation in this thesis it is worth discussing what self-reference is. According to Field (2008) there are three notions of self-reference of which two I have already mentioned. The first notion is the one detectable in the Contingent Liar and is therefore called *Contingent Self-Reference*. For this kind of self-reference is achieved by empirical description. Recall that in Kripke's example to argue against syntactical solutions to the Liar Jonas and Nixon assert:

(1) Most of Nixon's assertions about Watergate are false.

(2) All of Jonas' assertions about Watergate are true.

Due to unfortunate empirical facts assumed in this example it can be shown that (1) is true iff it is false which is also the case for (2). But Jonas being the one to assert (1) and Nixon being the one to assert (2) are themselves empirical assumptions. It is these two empirical assumptions that make (1) a statement (indirectly) referring to itself. For (1) is a statement referring to (2) because it is one of Nixon's assertions about Watergate. And this assertion refers to (1) as this is an assertion about Watergate made by Jonas. For the same reason (2) (indirectly) refers to itself as well.

The kind of self-reference used in the derivation of the Liar in the Introduction is the one in which sentences contain their own name. L is the name of the sentence 'L is not true'. Field calls it *Artificial Self-Reference* as self-reference here is achieved via naming sentences in the certain way. He prefers self-reference via descriptions as in the kind of self-reference presented in the following section.

2.1 Diagonalisation

The third kind of self-reference is called *Gödel-Tarski Self-Reference* as it has been developed by Kurt Gödel and Alfred Tarksi. In 1931, Gödel establishes a way for number theory to refer to its own syntax via natural numbers[1]. The idea is to first assign to each symbol of the vocabulary of a formal language, in this case number theory, some natural number. Using facts about the relations between numbers one can then assign to each formula a of number theory its own unique *Gödel number* $\langle a \rangle$. Crucially, one can assign a Gödel number also to sequences of formulas and proofs. The function that assigns a Gödel number to an expression of the formal language is bijective, i.e. each expression has a unique Gödel number. Moreover, given an expression a, one can effectively compute $\langle a \rangle$ and the other way round. Now, the method to achieve Gödel-Tarski self-reference is *diagonalisation* and is applied in the well-known Gödel-Tarski Diagonalisation Lemma. Before we look at a general definition of diagonalisation and how it can lead to self-reference, let us see how it is done in the Lemma. This is the formulation and proof of Gödel-Tarski Diagonalisation Lemma as presented in Field (2008, 26-27):

Gödel–Tarski Diagonalization Lemma (proof-theoretic version). *For any formula $C(v)$ in the language of arithmetic with 'v' as its only free variable, there is a sentence F in that language such that $F \leftrightarrow C(\langle F \rangle)$ is provable in the arithmetic theory*[2].

Proof. Let the self-application of a formula $B(v)$ (with a single free variable v) be the sentence $B(\langle B \rangle)$ that results from substituting the name $\langle B \rangle$ for all free occurrences of v in B. (When the syntax is done in arithmetic, there is a corresponding operation on Gödel numbers that takes the Gödel number of a formula to the Gödel number of its self-application; since it's possible to identify formulas with their Gödel numbers, I will call this numerical operation 'self-application' as well.) This operation is expressible in the arithmetic language[3]. Now let $C(v)$ be some formula:

(1) Let $D(v)$ be the formula

$$\exists x [x \text{ is the self-application of } v \wedge C(x)]$$

[1] In the same year Tarski offers a direct way of self-reference, i.e. a language that can refer to expressions of that language. Field rightly notes, that both versions are equivalent in the sense that both theories can be developed within each other.

[2] For purposes of simplicity C is a monadic predicate. The general version of this lemma uses diagonalisation in no different way.

[3] That is to say that the operation of substitution is (strongly) representable by a formula in arithmetic. An n-ary relation R of natural numbers is (strongly) representable in a system of arithmetic F if there is a formula $A(v_1, \ldots, v_n)$ of the language of F with n free variables $v_1, \ldots, v_n$ such that for all natural numbers $\mathbf{a}_1, \ldots, \mathbf{a}_n$:

if $\mathbf{a}_1, \ldots, \mathbf{a}_n \in R$ then $F \vdash A(a_1, \ldots, a_n)$ and

if $\mathbf{a}_1, \ldots, \mathbf{a}_n \notin R$ then $F \vdash \neg A(a_1, \ldots, a_n)$ where $a_1, \ldots, a_n$ denote $\mathbf{a}_1, \ldots, \mathbf{a}_n$, respectively.

It is in this way that the formula $A(v_1, \ldots, v_n)$ expresses the relation R in F. Now, if $A(v)$ is a formula in F with one free variable v then the operation of substituting n, denoting the natural number $\mathbf{n}$ in F, for v in $A(v)$ can be represented (expressed) by a formula $\phi(x, y, z)$ which is true of x, y, z iff $x = \langle A(v) \rangle, y = \mathbf{n}$ and $z = \langle A(n) \rangle$. z is the self-application of $A(v)$ iff $x = y$. In the proof of the Diagonalisation Lemma, Field writes 'z is the self-application of x' instead of '$\phi(x, x, z)$'.

(2) Let F be the self-application of the formula $D(v)$, i.e. the formula

$$\exists x[x \text{ is the self-application of } \langle D(v) \rangle \wedge C(x)].$$

(3) The following claim is a truth that is provable by means available in the arithmetic theory:

$$\langle F \rangle \text{ is the unique self-application of } \langle D(v) \rangle.$$

(4) From (3) and the nature of the formula F, it is clear that

$$F \leftrightarrow \exists x[x = \langle F \rangle \wedge C(x)]$$

(5) Hence

$$F \leftrightarrow C(\langle F \rangle)$$

is an obvious truth that is provable by these elementary means.

□

Field remarks that the achieved self-reference is not literal. For F does not say of its Gödel number that it has the property C, rather it is provably equivalent to a sentence that says that its Gödel number has property C.

Now, what is diagonalisation and how does it lead to a formula F that is equivalent to a formula referring to the Gödel number of F? There are two crucial steps in this proof that deserve to be highlighted.

Step 1
Define the formula $D(v)$ that translates

> 'The self-application of v has the property C'.

$D(v)$ is a diagonaliser, i.e. a formula referring to a self-application.

Step 2
Form the self-application of that formula, thus form the self-application of a formula that refers to a self-application. By doing so we get the formula F that translates

> 'The self-application of 'the self-application of v has the property C' has the property C'.

Now, since $D(v)$ is a formula referring to a self-application its own self-application will refer to itself. For F *is* the self-application of 'the self-application of v has the property C' and therefore states that F has property C.

In short, self-reference is achieved by forming the diagonaliser of a diagonaliser. Interestingly, the notion of a diagonaliser itself seems to contain some notion of self-reference, i.e. the notion of self-application.

In *Diagonalization and Self-Reference* (1994, 16-17) Raymond Smullyan generalizes the notion of a diagonaliser and shows how to achieve self-reference using a diagonaliser. Consider a language in which we have predicates, sentences and a way to apply predicates to sentences. In particular, for each predicate H and each sentence X the application of H to X, $H(X)$, is an expression of that language. Further, some equivalence relation between sentences is defined for that language. Let H be a predicate. The *diagonaliser* of H is a predicate H^* such that for all predicates K:

$$H^*(K) \text{ is equivalent to } H(K(K)).$$

A sentence X is called a *fixed point* of a predicate H if X is equivalent to $H(X)$. Thus a fixed point is a self-referential sentence in the Gödel-Tarski sense. For X is equivalent to a sentence that states of X that it has the property H.

Theorem 1. *If H has a diagonaliser, then there is a fixed point for H.*

Proof. Let H^* be the diagonaliser of H. Thus, by definition, for all predicates K:

$$H^*(K) \text{ is equivalent to } H(K(K)).$$

Since H^* is a predicate, we take it for K and get:

$$H^*(H^*) \text{ is equivalent to } H(H^*(H^*)).$$

Thus, $H^*(H^*)$ is a fixed point of H. □

What is done in the proof here is analogous to what is done in the proof of Gödel-Tarski Diagonalisation Lemma. In number theory, the equivalence relation between two sentences A and B is that $A \leftrightarrow B$ is provable. While for Smullyan a diagonaliser is a predicate, the diagonaliser in the Diagonalisation Lemma is a formula with a single free variable. Here, $D(v)$ is the diagonaliser of the formula $C(v)$ as, by the definition of $D(v)$,

$$D(v) \leftrightarrow C(\langle v(\langle v \rangle)\rangle)$$

is provable. By defining F as the diagonaliser of D we take D itself for v just like in the proof for Theorem 1 and get that

$$D(\langle D(v)\rangle) \leftrightarrow C(\langle D(\langle D(v)\rangle)\rangle).$$

Thus F is a fixed point of C. Self-reference in the Gödel-Tarski sense is captured by the notion of a fixed point.

This and a slightly more general notion of artificial self-reference are the ones most often used in discussions about the self-referentiality of paradoxes, according to Leitgeb. He shows how to define these notions in a formal way and why both of them are 'strongly deficient concerning what they should be explications for' (2002, 1).

2.2 Two notions of self-reference

The first notion of self-reference is defined for singular sentences, i.e. sentences that refer to all the referents of all of its singular terms, and only to them. Formally, he begins with defining a reference relation for sentences. Let ref be the reference relation for terms. He defines

> $ref_1(x,y) \leftrightarrow x$ is a sentence $\wedge \exists z(z$ is a singular term $\wedge x$ contains $z \wedge ref(z,y))$.

Self-reference is then defined as

> $selfref_1(x) \leftrightarrow ref_1(x,x)$.

Clearly, the liar is $selfref_1$ as it is a sentence containing a singular term referring to itself, e.g. 'this sentence' or 'L' if L is the name of the sentence. This also shows that Field's notion of Artificial Self-Reference is just a special case of $selfref_1$. To include the cases of indirect self-reference as in, for instance, the Liar chain he first defines ref_1^* as the transitive closure of ref_1[4] and then defines:

> $circular_1(x) \leftrightarrow ref_1^*(x,x)$.

The second notion is a more formal version of a fixed point:

> $circular_2(x) \leftrightarrow x$ is a sentence $\wedge \exists f(f$ is a syntactical mapping $\wedge f(x) = x)$.

Formalizing the Liar as $L :=$ '$\neg T(\langle L \rangle)$' where T is a truth predicate and $\langle L \rangle$ the Gödel number of the formula L, it is also $circular_2$. Take the syntactical mapping f that maps a formula A to the formula $\neg T(\langle A \rangle)$. As shown by the Diagonalisation Lemma, for the predicate $\neg T$ (provided it is added to the arithmetical language) there is a sentence L that is equivalent to $\neg T(\langle L \rangle)$. Thus we have that $f(L) = L$ *up to arithmetical equivalence*. Analogously to saying that a sentence X is a fixed point of predicate H if X is equivalent to $H(X)$, we can say that a formula A is a fixed point of the syntactical f if $f(A) = A$ *up to arithmetical equivalence*.

Both notions, however, have unacceptable drawbacks. $circular_1$ does not satisfy the Equivalence Condition (EC): if A is self-referential/circular, and if B is logically equivalent to A, then also B is self-referential/circular. But EC 'is plausible because logically equivalent sentences are not only extensionally equivalent in the actual world, but indeed in every logically possible world, and thus indistinguishable in terms of the semantics of first-order predicate logic. If self-reference is to be defined by extending the usual reference relation for terms, i.e., a semantical relation, it is certainly strange if EC is invalidated' (2002, 7). As a counterexample take $L' =$ '$(P(a) \vee \neg P(a)) \vee \neg T(\langle L' \rangle)$' which is logically equivalent to $(P(a) \vee \neg P(a))$. The former is $circular_1$ but

[4] The transitive closure of ref_1 is the smallest superset of the extension of ref_1 such that if $ref_1(x,y)$ and $ref_1(y,z)$ then also $ref_1(x,z)$.

the latter is not. Any alteration of the definition of $circular_1$ in terms of logical equivalence would have the effect that $P(a) \vee \neg P(a)$ becomes $circular_1$ which seems absurd. Even worse, any sentence might turn out $circular_1$. My concern to this criticism of $circular_1$ is the plausibility of EC. The only two solutions to the counterexample, i.e. to save EC, is to either (i) make both L' and $P(a) \vee \neg P(a)$ $circular_1$ or (ii) make both not $circular_1$. Option (i) seems absurd since $P(a) \vee \neg P(a)$ is not self-referential. So (ii) is the only option to maintain EC. However, this seems not less controversial to me than giving up EC. Rather than giving up the notion of $circular_1$ as a definition for self-reference, I take the example to be an indicator that something is wrong with EC. After all, Leitgeb's motivation for the plausibility of EC seems far from obvious. According to him, defining self-reference by 'extending the usual reference relation for terms' should make self-reference invariant concerning logical equivalence. He seems to rely EC on some connection between the reference relation for terms and logical equivalence. Yet, clearly, logical equivalence between two sentences A and B does not depend on A and B containing singular terms referring to the same object. For instance, let $A =$ '$\langle P(a) \rangle = \langle P(a) \rangle$' and $B =$ '$\langle \neg P(a) \rangle = \langle \neg P(a) \rangle$'. A and B refer to different objects as the Gödel number of $P(a)$ is different from the Gödel number of $\neg P(a)$, but they are equivalent as they both state a tautology. It comes with no surprise that sentences like $C =$ '$\langle C \rangle = \langle C \rangle$' which are $circular_1$ are equivalent to A and B which are not self-referential in any way.

The second notion, $circular_2$, is too broad because it makes every sentence self-referential. For any sentence is the fixed point of some syntactical mapping. Any sentence x can be disjoined with a logical truth y resulting in a sentence that is equivalent to x. For instance, $P(a)$ is the fixed point of $g($'$P(a)$'$) =$ '$\langle P(a) \rangle = \langle P(a) \rangle \vee P(a)$' since $\langle P(a) \rangle = \langle P(a) \rangle \wedge P(a) \leftrightarrow P(a)$. Leitgeb sees no obvious way to state 'a proper characterization of what a circular sentence is demanded to be a fixed point of' (2002, 9) which is needed for this notion of self-reference to work. However, Smullyan's notion of a fixed point, the one involving a predicate rather than a syntactical function, might be of help. For, arguably, not everything that counts as a syntactical function counts as a predicate. It is clear that $g(P(a)) =$ '$\langle P(a) \rangle = \langle P(a) \rangle \vee P(a)$' is a syntactical map where the argument of the function, $P(a)$, is equivalent to the result of applying the function on it. However, it seems, at least, debatable that g can also be interpreted as a predicate. For what does it mean that '$P(a)$' has the property that $P(a)$ or $\langle P(a) \rangle = \langle P(a) \rangle$? Taking Smullyan's definition of a fixed point might thus avoid declaring all sentences a fixed point because they are equivalent to some other sentence involving a logical truth.

Objections

To apply Smullyan's definition of a fixed point involving a predicate to avoid that every sentence is a fixed point, we need restrictions on what counts as a predicate. For instance, if we count the truth predicate in we face the same problem Leitgeb described for $circular_2$. For according to the T-schema the truth predicate T is a predicate every sentence is a fixed point of. Maybe it is enough to exclude the truth predicate because it is governed by the T-schema. Yet, further discussion about what notion of predicate needs to be developed is necessary to avoid more problematic cases. Such a discussion, however, goes beyond the scope of this thesis. For now, it suffices to say that talking about predicates rather than syntactical functions might avoid turning every sentence into a fixed point.

2.3 A different notion of Diagonalisation

There is another notion of diagonalisation that is connected to some form of self-reference we have not yet considered. In Cantor's paradox one finds a contradiction by applying Cantor's Theorem to the set of all sets V. Cantor's Theorem states that no set x can be put into a one-to-one correspondence with its powerset $\mathcal{P}(x)$, the set of all subsets of x. Yet, this seems not to be the case for V. V is the set of all sets and therefore must contain $\mathcal{P}(V)$. But $\mathcal{P}(V)$ contains V as well as V itself is a subset of V. Thus $\mathcal{P}(V) = V$. The identity function of V is a bijection between V and $\mathcal{P}(V)$ and therefore contradicts Cantor's Theorem.

To prove his theorem, Cantor discovered a method that is called diagonalisation (Priest, 2002, 118-119). Suppose there is a bijection f between some set X and its powerset $\mathcal{P}(X)$. Now, consider the subset of X, which therefore is a member of $\mathcal{P}(X)$, whose members are not members of the set that f assigns to them, $\{y \in X | y \notin f(y)\}$, call it z. Since f is a bijection, there is some $w \in X$ that z is assigned to by f, i.e. $z = f(w)$. A problem arises when we ask if w itself is in $f(w)$:

$$w \in f(w) \text{ iff } w \in \{y \in X | y \notin f(y)\} \text{ iff } w \notin f(w).$$

Thus we have that $w \in f(w)$ iff $w \notin f(w)$ from which, just as in the Liar paradox in the Introduction, we can deduce the contradiction $w \in f(w)$ and $w \notin f(w)$. Why this method is called diagonalisation can be illustrated if we take X to be a set whose members can be well-ordered: $x_0, x_1, x_2, \ldots$. Take the corresponding sequence $f(x_0), f(x_1), f(x_2), \ldots$ and consider the Table 2.1. T ('true') and F ('false') stand for yes and no concerning the question whether x_i is a member of $f(x_j)$, e.g. $x_0 \in f(x_1)$ and $x_0 \in f(x_2)$, but $x_0 \notin f(x_0)$ as in

Table 2.1:

	$f(x_0)$	$f(x_1)$	$f(x_2)$	$\ldots$
x_0	**F**	T	T	$\ldots$
x_1	F	**T**	F	$\ldots$
x_2	F	T	**F**	$\ldots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

Table 2.1. Now, to build z we apply the condition $y \notin f(y)$ by reversing each entry of the diagonal (highlighted with bold letters), e.g. since $x_0 \notin f(x_0)$, we put x_0 in z and since $x_1 \notin f(x_1)$, x_1 is not a member of z. By reserving the condition at each entry of the diagonal we ensure that the resulting set z is different from $f(x_0), f(x_1), \ldots$ and therefore a set that is *not* among the sequence $f(x_0), f(x_1), \ldots$. In short, z is created by 'diagonalising out' of a given list, hence the label 'diagonalisation'.

Circularity in this form of diagonalisation can be found in the way z is defined. The function f occurring in the definition of z is a function from X to the powerset $\mathcal{P}(X)$ which contains z. Thus the definition of z involves a set that contains z. Such definitions are called *impredicative*. They involve self-reference or circularity in the sense that they generalize over a set that contains the object that is to be defined. In particular, since $z = f(w)$ for some $w \in X$, the condition $w \notin f(w)$, i.e. $w \notin z$, has to be checked to determine whether $w \in z$ or $w \notin z$. That means that in order to fully determine set z, a condition referring to z has to be checked. According to Priest's analysis of the paradoxes of self-reference, as we will see in the coming chapter, this implicit form of self-reference can be found in all paradoxes of self-reference.

There is a way in which the two forms of diagonalisation come together. According to the Gödel-Tarski Diagonalisation, we can form the Liar sentence since for the predicate $\neg T$ (if added to the arithmetical language) there is a sentence L that is equivalent to $\neg T(\langle L \rangle)$, i.e. $L \leftrightarrow \neg T(\langle L \rangle)$ is provable in arithmetic. On the other hand, if Tarski's schema T holds in arithmetic, $L \leftrightarrow T(\langle L \rangle)$ is provable as well. By classical logic, $T(\langle L \rangle) \leftrightarrow \neg T(\langle L \rangle)$ is provable in arithmetic which gives us a sentence of the form A iff $\neg A$. The very same form of sentence is achieved by the other form of diagonalisation: $w \in f(w)$ iff $w \notin f(w)$. This shows that both forms of diagonalisation can lead to a statement of the form A iff $\neg A$[5]. Such a statement, I will argue in chapter 4, is the crucial feature of the paradoxes of self-reference.

2.4 Conclusion

We saw two notions of self-reference linked to the paradoxes of self-reference, $circular_1$ and $circular_2$, which, according to Leitgeb, are deficient. I have argued that the problem with $circular_1$ might be solved by questioning the plausibility of EC according to which sentences that are logically equivalent to self-referential sentences are themselves self-referential. A sentence is $circular_2$ if it is a fixed point of a syntactic mapping up to arithmetical equivalence. In non-trivial cases the arithmetical equivalence can be established by the method of diagonalisation which is essentially some form of self-application. However, any sentence x is equivalent to $x \vee y$ where y is some logical truth, and thus

[5] In *Naming and Diagonalization, from Cantor to Gödel to Kleene* (2006), Haim Gaifman claims that the relation between the two notions of diagonalisation is closer than that. He proposes a way to get from Cantor's diagonalisation to a construction that yields a fixed point in the Gödel-Tarski sense. The Gödel-Tarski Diagonalisation Lemma is a special case of that construction.

any sentence is the fixed point of some syntactic function, e.g. the function that maps 'x' to '$x \vee y$'. I proposed to bring the notion of a fixed point closer to the notion provided by Smullyan which involves a predicate rather than a function.

We also saw another notion of diagonalisation, defining a new member of a given list that is, however, not in that list. This method involves some form of (implicit) self-reference, impredicative definitions. An impredicative definition involves the totality that contains the object that is defined and is therefore implicitly circular. As we will see in the next chapter, according to Priest's analysis of the paradoxes, this form of self-reference can be found in each paradox.

Chapter 3

Priest's Account of the Paradoxes of Self-reference

In *Beyond the Limits of Thought* (2002) Graham Priest establishes an important step towards believing in true contradictions, the Inclosure Schema. It is the first attempt of a unified account of the paradoxes of self-reference. He claims that it holds the key to what generates the paradoxes. In addition, he argues that all previous attempts to solve the paradoxes case by case fail. Instead, given a uniform account of the paradoxes the natural response is a uniform solution. The Inclosure Schema gives us contradictory objects and the only solution is thus to take them to be truly contradictory. In this chapter I am looking closely at how the Inclosure Schema characterizes the paradoxes and what notion of self-reference is ascribed to them according to this characterization. Further, I discuss and extend the criticism the Inclosure Schema has received.

3.1 The Inclosure Schema

The Inclosure Schema goes as follows (Priest, 2002, 134). Let ϕ and ψ be two properties and δ a function that satisfy the following conditions.

(1) $\Omega = \{y|\phi(y)\}$ exists and $\psi(\Omega)$ (Existence)

(2) for all subsets x of Ω such that $\psi(x)$:

  (a) $\delta(x) \notin x$ (Transcendence)

  (b) $\delta(x) \in \Omega$ (Closure)

Now, Ω itself is a subset of Ω and by (1) $\psi(\Omega)$. So by (2) we get the contradiction that $\delta(\Omega) \notin \Omega$ and $\delta(\Omega) \in \Omega$. The Inclosure Schema goes back to Russell (1905) and is a generalization of Russell's schema in which there is no property ψ, i.e. the assumption that $\psi(\Omega)$ in (1) and the assumption that

$\psi(x)$ for all subsets x of Ω in (2) are left out. Priest added ψ to the schema to incorporate the paradoxes of definability, as we will see in due course.

The crucial elements of this schema are Ω and δ, the *diagonaliser*, since the contradiction ultimately appears when δ is applied to Ω. The 'diagonaliser need not be defined literally by diagonalisation; but, as we shall see, it is always defined systematically to ensure that the result of applying it to any set cannot be identical with any member of that set. Diagonalisation proper is the paradigm of such a procedure' (Priest, 2002, 130). What makes δ a diagonaliser is not that it literally 'diagonalises out' a member of a given list, but it gives an object that is among the members of Ω (Closure), yet which cannot be in any subset of Ω including Ω itself (Transcendence). The reason why Priest named δ a diagonaliser thus seems to be that it creates the same kind of contradiction as in Cantor's Theorem: Some object is and is not a member of a certain set.

Even though δ is not defined literally as a diagonaliser in each paradox, $\delta(\Omega)$ is defined using an impredicative definition. For, as we will see in each paradox, it is defined referring to Ω and at the same time supposed to be in Ω (Closure). Priest describes the self-referential structure in the schema as follows (Priest, 2002, 4):

> In general, the arguments both for Closure and Transcendence use some form of self-reference, a method that is both venerable and powerful. Closure is usually established by reflecting on the conceptual practice in question. [...] Arguments for Transcendence are of more varied kinds; often they involve applying a theory to itself. Some of them are more technical; a paradigm of these is diagonalisation; a technique familiar from the logical paradoxes. This construction is precisely a boundary-tearing heuristic which, given any boundary of a suitable kind, can be applied to violate it.

This suggests that self-reference can be found in different forms and sometimes in form of diagonalisation. Let us have a look at how the paradoxes fit in the schema and what notion of self-reference they involve.

3.2 How it captures the paradoxes and what makes them self-referential

Russell's Paradox

Let $\phi(y)$ be '$y \in V$' where V is the set of all (pure) sets so that $\phi(y)$ is saying that y is a set. Thus $\Omega = V$. The function $\delta(x)$ is ρ_x, where $\rho_x = \{y \in x | y \notin y\}$ and $\psi(x)$ is the property of self-identity, $x = x$. Now, for all $x \subseteq V$:

> **Transcendence**: $\rho_x \notin x$.
>
> *Proof.* Let $x \subseteq V$ and $\rho_x = \{y \in x | y \notin y\}$. Suppose $\rho_x \in \rho_x$, then by definition of ρ_x we get that $\rho_x \notin \rho_x$ which is a contradiction. So $\rho_x \notin \rho_x$.

But then $\rho_x \notin x$ or otherwise $\rho_x \in \rho_x$. So we have $\rho_x \notin x$.

Closure: $\rho_x \in V$.

Proof. By construction of V, since $x \subseteq V$ so is $\rho_x \subseteq V$ and hence $\rho_x \in V$.

Recall, that in Cantor's Theorem we defined $z = \{y \in X | \ y \notin f(y)\}$ where f is a bijection between X and its powerset $\mathcal{P}(X)$ to prove there cannot be such a bijection. Yet, we have that $V = \mathcal{P}(V)$, so there must a bijection between V and its powerset, e.g. the identity function, which is a contradiction. A contradiction can also be brought about when by applying the construction in the proof of Cantor's Theorem to V: f is the identity function and $z = \{y \in V | \ y \notin y\}$. Notice that now z is just $\delta(\Omega)$ in Russell's paradox as stated above, i.e. $\rho_V = \{y \in V | \ y \notin y\}$. One can then prove that $\rho_V \in \rho_V$ iff $\rho_V \notin \rho_V$ and hence derive the contradiction that $\rho_V \in \rho_V$ and $\rho_V \notin \rho_V$. To get the exact same contradiction in Russell's paradox one can put it in a slightly different way that still fits the Inclosure Schema: Take $\phi(y)$ to be '$y \in V \wedge y \notin y$'. Again, $\psi(x)$ is the property of self-identity $x = x$. Ω is just ρ_V and $\delta(x)$ the identity function. Now, for all $x \subseteq \rho_V$:

Transcendence: $x \notin x$.

Proof. Let $x \subseteq \rho_v$, i.e. $x = \rho_x$. Suppose $\rho_x \in \rho_x$ then $\rho_x \notin \rho_x$, contradiction. So $\rho_x \notin \rho_x$.

Closure: $x \in \rho_V$.

Proof. Suppose $\rho_x \notin \rho_V$, then $\rho_x \in \rho_x$. But then also $\rho_x \notin \rho_x$, contradiction. So $\rho_x \in \rho_V$.
Since $\rho_V \subseteq \rho_V$ we get that $\rho_V \in \rho_V$ and $\rho_V \notin \rho_V$.

This shows that Cantor's paradox and Russell's paradox are essentially the same in the sense that they both apply diagonalisation to prove a contradiction about V. Applying the construction in the proof of Cantor's theorem to V is another way of proving a contradiction. This way of proving a contradiction is just another way of stating Russell's paradox.

In both versions no sentence can be detected that is either *circular*$_1$ or *circular*$_2$. Self-reference can be found in form of an impredicative definition. In particular, the definition of $\delta(\Omega)$, i.e. $\rho_V = \{y \in V | y \notin y\}$, is impredicative as it involves the totality V to which ρ_V belongs.

The Liar

Let $\phi(y)$ be the truth predicate so that Ω is the set of all true sentences, *Tr*. $\psi(x)$ is 'x is definable' and δ is a function σ defined by some form of diagonalisation such that for any definable set x, $\sigma(x) = \alpha$ where $\alpha = \langle \alpha \notin x \rangle$. Thus the diagonaliser assigns to any definable set x a sentence that says of itself that it is not in x. Now, let a be definable, $a \subseteq Tr$:

Transcendence: $\sigma(a) \notin a$.

Proof. Suppose that $\sigma(a) \in a$. Then $\langle \alpha \notin a \rangle \in Tr$. By the T-schema, $\alpha \notin a$, i.e. $\sigma(a) \notin a$ which is a contradiction. Hence $\sigma(a) \notin a$.

Closure: $\sigma(a) \in Tr$.

Proof. From Transcendence follows, by definition of σ, $\alpha \notin a$. Again, by the T-Schema, we get $\langle \alpha \notin a \rangle \in Tr$, i.e. $\sigma(a) \in Tr$ which is Closure.

As we have seen in chapter 2, the Liar is both $circular_1$ and $circular_2$. Here, Priest chose to formulate the Liar as a fixed point of the syntactical map δ which is defined by some form of diagonalisation. At the same time $\delta(\Omega)$, here $\langle \alpha \notin Tr \rangle$, is impredicative as it refers to Tr and is supposed to be in Tr. The Knower paradox and Gödel's paradox can be established in almost exactly the same way. In the Knower paradox, the only difference is that $\phi(y)$ is 'y is known to be true' so that Ω is the totality of known things, Kn. Transcendence is established the same way with the extra premise that $\langle \alpha \notin a \rangle \in Kn$ implies $\langle \alpha \notin a \rangle \in Tr$ since knowledge implies truth. Since it is established that $\sigma(a) \notin a$, i.e. $\sigma(a)$ is not known which is precisely what $\sigma(a)$ states, it follows that $\sigma(a) \in Kn$. For what is proven must be known to be true. So we have Closure. In Gödel's paradox, $\phi(y)$ is 'y is provable' and Ω the set of provable things. Since, similarly to knowledge, provability implies truth we get Transcendence. Closure follows as Transcendence is proven, i.e. it is proven that $\sigma(a)$ is not provable which is precisely what $\sigma(a)$ states.

Grelling's Paradox

Here, $\phi(y)$ is '$\neg y$ *sat* y' where *sat* is the satisfaction relation. So $\phi(y)$ is the predicate referring to the property of not being true of oneself and Ω is the set of all predicates that are not true of themselves, *Het*[1]. $\psi(x)$ is 'x is definable'. For any definable set x, $\delta(x) = \langle v \in x \rangle$ where v is any new variable. The diagonaliser thus assigns to any definable set x the open sentence that translates 'v belongs to x'. Let a be definable and $a \subseteq Het$:

Transcendence: $\langle v \in a \rangle \notin a$.

Proof. The proof of Transcendence is analogous to the one in the Liar. Assume that $\langle v \in a \rangle \in a$. So also $\langle v \in a \rangle \in Het$ and, by definition of Het, $\neg(\langle v \in a \rangle \; sat \; \langle v \in a \rangle)$. By the Satisfaction Schema $\langle v \in a \rangle \notin a$. Again, we have a contradiction, thus $\langle v \in a \rangle \in a$.

Closure: $\langle v \in a \rangle \in Het$.

Proof. Applying the Satisfaction Schema on Transcendence, we get that $\neg(\langle v \in a \rangle \; sat \; \langle v \in a \rangle)$ which means that $\langle v \in a \rangle \in Het$.

The impredicative definition of $\delta(\Omega) = \langle v \in Het \rangle$ is again what constitutes self-reference in this paradox. For *Het* is the set of *all* sentences that are not

[1] One might also take Ω to be the set of all properties that do not apply to themselves as I will do in the next chapter.

true of themselves. So the sentence $\langle v \in Het \rangle$ is defined by (indirectly) referring to the set of all sentences.

Grelling's paradox, the Liar and the analogous Knower and Gödel's paradox can also be established without the premise of definability $\psi(x)$. If one takes not sentences but propositions or other semantic entities to be truth bearers, the existence of $\delta(x)$ does not require definability of x. For $\langle \alpha \notin x \rangle$ is then interpreted as something other than a sentence. In the Liar, the Knower and Gödel's paradox, $\langle \alpha \notin x \rangle$ is a proposition that exists for any (also not definable) set x. Likewise, in Grelling's paradox $\langle \alpha \notin x \rangle$ is a property that exists for any set x.

Note that Transcendence so far has been established by reductio from $\delta(x) \in x$ by virtue of what Priest calls a *bridge principle* (Priest, 2002, 146), a principle connecting semantic notions with the world. In the Liar it is the T-schema providing a bridge between the truth predicate and the world. In Grelling it is the Satisfaction schema that connects 'true of' with the world.

For the paradoxes of definability, i.e. Berry's Paradox, König's Paradox, Berkeley's Paradox[2] and Richard's Paradox, I will take only one example as the way they fit the Inclosure Schema is analogous[3].

König's Paradox

Let $\psi(x)$ be 'x is definable'. $\phi(y)$ is the property of being a definable ordinal '$y \in DOn$', so Ω is the set of all definable ordinals DOn. The diagonaliser $\delta(x)$ gives us the least ordinal that is not in x, $\delta(x) = \mu y(y \notin x)$. Let $x \subseteq DOn$.

> **Transcendence**: $\mu y(y \notin x) \notin x$.
>
> *Proof.* By definition, $\mu y(y \notin x) \notin x$ as $\mu y(y \notin x)$ is the least y not in x.
>
> **Closure**: $\mu y(y \notin x) \in DOn$.
>
> *Proof.* Consider that x is definable and that therefore $\mu y(y \notin x)$ is definable, i.e. $\mu y(y \notin x) \in DOn$.

Again, $\delta(\Omega)$ is defined impredicatively. For $\mu y(y \notin DOn)$ is defined in terms of 'DOn' but itself is in DOn.

There is another group of paradoxes that fit the Inclosure Schema in a similar way, i.e. Burali-Forti's Paradox, Mirimanoff's Paradox and Priest's Fifth Antinomy, of which I will consider the first one[4].

[2] It is highly controversial if Berkeley's paradox belongs to this group (Kroon, 1995; Tennant, 1998).

[3] In each case $\phi(y)$ is the property of being an object definable in the English language (in some particular way).

[4] Here the resemblance lies in the fact that Ω is the collection of all objects generated by δ.

Burali-Forti's Paradox

Here, $\psi(x)$ is not needed, so we can take it to be the property of self-identity. $\phi(y)$ is 'y is an ordinal' and $\delta(x)$ is $log(x)$, a function assigning to each sequence of ordinals x the least ordinal greater than all members of x. Ω is the set of all objects generated by δ, On. Let $x \subseteq On$.

> **Transcendence**: $log(x) \notin x$.
>
> *Proof.* By definition of $log(x)$.
>
> **Closure**: $log(x) \in On$.
>
> *Proof.* By definition of On.

Again, the notion of self-reference applied here is the impredicative definition of $log(On)$. It is defined as the least ordinal greater than all members of On and yet itself in On.

Note that in this group of paradoxes and in the paradoxes of definability, Transcendence and Closure can be obtained by virtue of how δ and Ω are defined, respectively, given that Ω and $\delta(\Omega)$ exist[5].

3.3 How it has been criticized

The Inclosure Schema has received criticism of various kinds. A seemingly serious attack has been made by Dümont and Mau (1998) who accused the schema of being not sound. They argue that the condition of Transcendence, $\delta(x) \notin x$, cannot be met for Ω. This is because δ is defined as $\delta : \mathcal{P}(\Omega) \mapsto \Omega$ and according to the set-theoretic concept of functions for every $x \in \mathcal{P}(\Omega)$ we get that $\delta(x) \in \Omega$. Thus by the definition of a function it is required that $\delta(\Omega) \in \Omega$ since $\Omega \in \mathcal{P}(\Omega)$. Therefore Transcendence for Ω, i.e. $\delta(\Omega) \notin \Omega$, contradicts the requirements δ has to satisfy as a function. Since the condition of Transcendence also appears in Russell's Schema, this criticism concerns even the very first unified account of paradoxes by Russell. Fortunately, the issue is not as serious as it seems. Neither Russell nor Priest have specified what sets δ is defined on. The way Dümont and Mau have defined δ indeed leads to the problem described above. However, one can easily change δ to be defined as $\delta : \mathcal{P}(\Omega) \mapsto X$ where X is some set such that $\Omega \subset X$. This way, Transcendence can still hold as long as $\delta(\Omega) \in X$ (while $\delta(\Omega) \notin \Omega$). Russell's Schema and Priest's Schema are therefore perfectly sound. The reason why they left δ undefined might be that it simply does not matter. It is only important for δ to be defined in a way that it satisfies Transcendence and Closure.

Grattam-Guinness (1998) doubts the generality and accuracy of the Schema as it incorporates non-paradoxes like the Barber paradox and misses to incorporate the Curry Paradox. In a reply to his paper, Priest stresses the fact that

[5] Transcendence in the Fifth Antinomy requires a proof by induction which, however, requires only the definition of δ. (Priest, 2002, 131)

the premises of the Inclosure Schema have to be true in order to obtain a (truly) contradictory object (Priest, 2002, 277). There is no such thing as the Barber who shaves all Barbers that do not shave themselves, as Priest argues. Therefore the Barber paradox is ruled out as an inclosure paradox. Concerning the Curry Paradox, Priest distinguishes two kinds (Priest, 2002, 168-169):

> Some of the paradoxes I have discussed proceed by establishing a sentence of the form $\alpha \leftrightarrow \neg\alpha$. [Russell, Liar and Grelling] For each paradox of this kind, we can form a new paradox by replacing $\neg\alpha$ uniformly with $\alpha \rightarrow \beta$, where β is an arbitrary formula; or, more simply, with $\alpha \rightarrow \perp$, where $\perp$ is some logical constant entailing everything. [...] Do such paradoxes fit the Inclosure Schema? Yes and no, depending on what $\rightarrow$ is. If it is a material conditional then, in most logics, $\alpha \rightarrow \perp$ is logically equivalent to $\neg\alpha$, and so the curried version of each paradox is essentially the same as the uncurried form. If, on the other hand, $\rightarrow$ is a non-material conditional (for example, a strict conditional), then $\alpha \rightarrow \perp$ and $\neg\alpha$ are quite different notions. [...] In this case, the curried versions of the paradoxes belong to a quite different family. Such paradoxes do not involve negation and, *a fortiori*, contradiction.

Priest makes clear that the version of the Curry paradox that leads to contradiction does indeed fit the schema as this version involves material implication and therefore is equivalent to the uncurried paradoxes. The versions that do not involve a non-material conditional do not lead to contradiction and therefore belong to a different category.

A similar attack has been made by Badici (2008) who argues that the Liar Paradox is not (properly) captured by the Schema. Unlike Russell's paradox, to obtain the Liar Paradox one does not need the condition of Existence, i.e. to assume that there is a set of all true sentences Tr. While Russell's Paradox is ruled out as soon as one drops the existence of V, the Liar Paradox remains unsolved after the existence of Tr is dropped. What is needed for the Liar Paradox is a truth predicate rather than Tr. Priest, however, claims that there is '[a] conceptual connection between satisfying a condition - being true - and being a member of a certain totality - being one of the totality of true things' (Priest, 2002, 279), i.e. saying that some sentence is true is equivalent to saying that that sentence is a member of the totality of true sentences. Thus one cannot state the Liar in terms of a truth predicate without implicitly making a statement about Tr. In other words, dropping the existence of Tr is to either drop the truth predicate as well or to just switch to a way of stating the paradox without explicitly mentioning Tr (but not saying that Tr does not exist). The connection between having a property and being a member of a certain totality follows from what Priest calls the Domain Principle according to which 'quantifying presupposes a corresponding totality of quantification' (Priest, 2002, 280). For a statement about some variable quantity to have determinate sense the domain of its variability has to be determinate, a definite set. The principle however is defended by Cantor only for mathematics (Hallett, 1984, 25):

> In order for there to be a variable quantity in some mathematical study, the 'domain' of its variability must strictly speaking be known beforehand through definition. However, this domain cannot itself be something variable, since otherwise each fixed support for the study would collapse.

Badici rightly notes that the alleged connection between having a property and being a member of a certain totality fails in certain (non-mathematical) cases, e.g. in case of vague predicates. Vague predicates have borderline cases, i.e. in some cases it is impossible to determine whether the predicate is true or not true of something. For instance, one cannot tell whether a man of 1,8m in height is tall or not. The predicate 'tall' is vague. In other words, there is no totality of things that are tall. While this seems to be a valid point, it only defends Badici's claim that what is needed for the Liar is a truth predicate rather than *Tr* if it can be shown that the truth predicate is vague. I will not try to answer this question here. More convincing arguments for the failure of the conceptual connection between properties and totalities *in the paradoxes* are ahead.

Tennant (1998) also takes *Tr* to be unnecessary for the Liar Paradox. But his main criticism is that the Inclosure Schema is best interpreted as a reductio from the existence of Ω. That is because there is no ground for the assumption of Ω in many paradoxes, in particular the Liar Paradox and Russell's paradox, unless one believes in the connection between conditions and totalities. This connection, according to Tennant, is only tenable if no restrictions are put on set existence. Yet, Russell's paradox seems to show that some restriction is needed. Why then base all paradoxes of self-reference on a connection that is questioned by one of the paradoxes? Another criticism concerning the essence of the paradoxes is that the structure of the Inclosure Schema abstracts too much from the argumentative details. The essential details of the paradoxes lie in the proofs of contradiction. In Russell, for instance, the proof can be put into a form such that it becomes a proof of non-existence, while this is not the case in the Liar. 'The proof of absurdity from the Inclosure Schema, as it stands, sweeps all the interesting logical details under the carpet, in pursuit of an overall structure that abstracts too much from the argumentative details, and also interpolates too much that is unnecessary' (Tennant, 1998, 30).

Kroon's (1995) criticism is basically the same as Tennant's. He also denies the alleged connection between sets and properties and attacks Priest's defense of it via the Domain Principle according to which talking about a variable that ranges over sets only makes sense when there is a set of all sets. Despite the fact that Kroon simply takes this principle to be unintuitive, he attacks the priorities Priest takes in the paradoxes. Why give up the Law of Non-Contradiction in favor of the Domain Principle? Further, he criticizes the Domain Principle as a condition for making sense of statements involving variables. For in the paradoxes one encounters inconsistent totalities. For instance, does talking about sets make sense because there is a set of all sets that does and does not include the Russel set?

The shared criticism made by Badici, Tennant and Kroon depends on the disbelief in the general connection between properties and totalities. In the next part of this chapter, I will support this disbelief by arguing that certain totalities do not exist.

3.4 Extended criticism on the Inclosure Schema

I am following Badici, Tennant and Kroon in denying that there is a conceptual connection between satisfying a property and being a member of a certain totality *in general*. The connection fails in cases where the totality involved is inconsistent *by definition*. That is to say that there are totalities defined in a way that contradicts facts and therefore do not exist. It is a major objective of this section to show that such inconsistent totalities can be found in the paradoxes of self-reference. The suspects are Burali-Forti, Mirimanoff, 5th Antinomy and the three paradoxes of definability. If we take the Inclosure Schema, according to which the existence of Ω is necessary to derive a contradiction in the paradoxes of self-reference, to be the correct analysis of these paradoxes, the proof of contradiction in theses paradoxes can be blocked. Moreover, we will see that some of these paradoxes in fact depend on the existence of Ω whether or not the Inclosure Schema is taken to be the correct analysis of them. Since Ω does not exist in these paradoxes, they can be considered solved. Further, I want to support Tennant's point that the essential details of the paradoxes do not lie in the structure of the Inclosure Schema. A comparison of the Transcendence proofs of the paradoxes captured by the schema reveals that the Liar, Russell and Grelling have an entirely analogous proof of Transcendence that is different from the rest. I suspect that these three paradoxes are not properly captured by the Inclosure Schema. Finally, I want to point out two weaknesses of the notion of self-reference ascribed to the paradoxes by the Inclosure Schema. To summarize, I want to argue for three claims:

(1) If the Inclosure Schema is the correct analysis of Burali-Forti, Mirimanoff, 5th Antinomy and the paradoxes of definability, these paradoxes can be considered solved.

(2) The Liar, Russell and Grelling are not properly captured by the Inclosure Schema.

(3) The notion of self-reference that the Inclosure Schema ascribes to the paradoxes has two weaknesses.

Claim 1

I will show case by case why the totalities are inconsistent by definition. The general idea is to spell out what it means to define a set of *all* objects of a certain kind.

(1) **Burali-Forti**: Here, Ω is the set of all ordinals generated by $log(x)$. First, consider how the generator $log(x)$ is defined: '[We] generate the next member after *any* sequence, x (whether or not it has a last member), by forming the least ordinal greater than all the members of x' (Priest, 2002, 120). In case x has no last member, the least greater ordinal is the union of all the members of x. In other words, no matter what unbounded sequence of ordinals we look at (that is generated by $log(x)$), there is a least greater ordinal containing all members of that sequence. Now, let us try to consider the unbounded sequence of *all* ordinals generated by $log(x)$, call it y. *By definition*, y is a sequence to which $log(x)$ cannot be applied further, i.e. to which there is no least greater ordinal. However, *by definition*, $log(x)$ can be applied to any unbounded sequence of ordinals (generated by $log(x)$) by building the union of all the members of the sequence. This definition is sound by the axiom of union in Zermelo-Fraenkel set theory. Thus, to say that there is the sequence y is to contradict the definition of $log(x)$. There is no unbounded sequence of *all* ordinals and therefore no least greater ordinal containing *all* ordinals, On. The contradiction of this paradox as stated above presupposes the existence of an object that is itself contradictory. This paradox does not prove a contradiction, it *assumes* a contradiction.

(2) **Mirimanoff**: This case is analogous to (1). Here, Ω is the set of all well-founded sets generated by $UP(x)$. By definition, $UP(x)$ can be applied to *any* unbounded sequence of well-founded sets generated by $UP(x)$ by taking the union of all members of the sequence. The unbounded sequence of all well-founded sets is, by definition, a sequence to which $UP(x)$ cannot be applied further and therefore that sequence is a construction that contradicts the definition of $UP(x)$. Thus, there is no such sequence and therefore no union R of all members of that sequence.

(3) **5th Antinomy**: Priest defines Ω, here the set of all objects generated by *thought of x* as follows (Priest, 2002, 100):

> Start with any object, say, the *Critique of Pure Reason*, and apply the generator [*thought of x*] iteratively, to produce, at each stage, the thought of the previous object: the *Critique*; the thought of the *Critique*; the thought of the thought of the *Critique*. And when we have an unbounded sequence of thoughts we next produce the thought of all of them, and then keep going. [...] Let this procedure be performed as often as possible [...]. Consider the totality of all thoughts generated in this way, T. [...] And *ex hypothesi*, the generator can be no further applied to T. So T cannot be thought of (Transcendence). But you *can* think of T: you have just done so (Closure). Contradiction.

The contradiction at the end of this quote is not proven, but simply *ex hypothesi*. The totality of *all* thoughts generated by *thought of x*, T, is, *by* *definition*, something that cannot be thought of. For it is the sequence of thoughts generated by *thought of x* to which the generator cannot be applied further to create a new thought. However, we can think of *any* sequence of thoughts generated in this way. So we can also think of T. Thus, again, to assume that the totality T, which the generator can no further be applied to, exists is to contradict the fact that the generator can be applied to any sequence of thoughts generated in this way.

(4) **König's**: The crucial notion of this paradox is the notion of definability. An ordinal is called *definable* if there is some non-indexical noun-phrase that (uniquely) refers to it, call it a definition. DOn is supposed to be the collection of all definable ordinals. Since there are only countably many noun-phrases to define uncountably many ordinals there must be ordinals that are not definable (Priest, 2002, 131). Since ordinals are well-ordered there is a least ordinal not in DOn, by the 'least ordinal principle' (Field, 2008). Now, the least ordinal not in DOn can be defined by 'the least ordinal not in DOn' which is a contradiction. It can be defined because the complement of DOn is itself well-ordered as all ordinals are well-ordered. However, how do we know DOn exists? Note, that any member of a well-ordered set x can be defined. The least member can be defined as 'the least member of x', the second least one as 'the second least member of x', etc. Likewise, no matter what subset y of x we consider, the least member of y (or its complement) can always be defined by 'the least member of (the complement of) y', the second least member (of its complement) as 'the second least member of (the complement of) y', etc. That means that in any well-ordered set x the members of (the complement of) a subset y can *always* be defined, provided x and y exist. Now, in case we take y to be the set of *all* definable objects of a certain kind, e.g. DOn, we define a subset of On whose compliment cannot contain definable objects. To assume that such a subset exists is to simply deny the fact that one can indeed define members of the compliment of DOn. Thus, there is no set of *all* definable ordinals and hence the proof of the existence of the least ordinal not definable via the least ordinal principle is blocked.

(5) **Berry's**: This is an analogous case to (4). Here, we consider the set of *all* natural numbers definable of a certain kind, i.e. definable in less than 99 words, DN_{99}. Now, since there are only finitely many noun-phrases with less than 99 words to define infinitely many natural numbers there must be natural numbers that are not definable in less than 99 words. Since natural numbers are well-ordered there is a least one not in DN_{99}, by the least number principle, which is definable by 'the least natural number not in DN_{99}'. Yet again, all definitions in less than 99 words of members of the complement of DN_{99}, e.g. 'the least natural number not in DN_{99}' or 'the second least natural number not in DN_{99}', prove that there cannot be DN_{99}, a set of natural numbers whose complement contains only undefinable natural numbers. For *any* complement of a

subset x of natural numbers has definable members, e.g. the member defined by 'the least natural number not in x' or 'the second least natural number not in x'.

(6) **Richard's**: Again, this is analogous to (4). Here, Cantor's diagonal argument is used to produce a new real number from a countably infinite set of real numbers between 0 and 1. However, Cantor's argument can be applied to *any* countable set of real numbers between 0 and 1 to define a new real number. It can certainly be applied to a countably infinite set of definable real numbers between 0 and 1. The crucial part of this paradox is to consider a countably infinite set of *all* definable real numbers between 0 and 1. Again, this set, by definition, excludes the possibility to apply Cantor's argument to define a new real number.

The argument for the inconsistency of Ω is basically the same in (1)-(6): To claim that there is a totality of objects that satisfy some property ϕ contradicts a fact about any set of objects that satisfy ϕ. In (1) and (2), the existence of Ω contradicts the fact that one can always unify the members of a sequence (of ordinals or well-ordered sets). In (3), the existence of Ω contradicts the fact that one can always apply $t(x)$ to (or think of) any sequence generated by $t(x)$. In (4) and (5), the existence of Ω contradicts the fact that one can always define the members of a well-ordered set. In (6), it contradicts the fact that one can apply Cantor's diagonal argument to any countable set of real numbers between 0 and 1.

Given that the totalities in (1)-(6) are defined in a way that contradicts facts, I conclude that these totalities do not exist. According to the Inclosure Schema, the paradoxes (1)-(6) depend on the existence of Ω. Thus, taken the Inclosure Schema as the correct analysis of these paradoxes, they can be considered solved. Even though the Inclosure Schema correctly analyzes the paradoxes, one cannot use the schema to prove that the contradictions are true and hence to support a dialetheic solution. The contradictions in (1)-(6) are not true since the existence premise of the Inclosure Schema is not satisfied.

Objections

The contradictions in (1)-(6) stem from how Ω is defined. The contradictions can be avoided by discarding the definitions of Ω, i.e. to deny that these definitions have a referent. Yet, one might object that defining the totality of objects in (1)-(6) seems intuitively unproblematic. For nothing seems wrong with forming a totality of things of a certain kind[6]. The paradoxes (1)-(6) can then be interpreted as cases in which two intuitions clash: The intuition that we can form the totality of objects of a certain kind and the intuition that no contradictions are true. Here is an attempt to show that in cases (1)-(6) it is in fact intuitively problematic to define the totality of all things satisfying ϕ.

In (1)-(3), the problem lies in the fact that totalities are involved in the definition of the objects that satisfy ϕ. In every limit case of the construction, one

[6] Let us leave out vague predicates for the moment.

takes the totality of all members of an unbounded sequence (of objects satisfying ϕ) and continues constructing new members out of it. Thus *each* totality of members of an unbounded sequence can be extended, i.e. one can define new members and build a greater set of objects satisfying ϕ. In other words, there are no absolute totalities of objects constructed in this way.

In (4)-(6), it seems absurd to hold on to the existence of totalities of definable objects of a certain kind. For defining something is always to also define the opposite of that something. For instance, if we define a house to be something that satisfies a conjunction of properties $\bigwedge_i \phi_i$, e.g. ϕ_1 is to have walls, ϕ_2 is to have a roof on top etc., we also define the objects not satisfying this conjunction as objects that are not a house. Now, consider a proper subset Y of some well-ordered set X. Then $X \setminus Y$ is non-empty and we can refer to specific members of it, e.g. the least member of $X \setminus Y$, the second least member of $X \setminus Y$, etc. Again, by defining Y we also define things that are not Y. Here, among the things that are not Y are the least member of $X \setminus Y$, the second least member of $X \setminus Y$, etc. Thus, by defining a proper subset Y of some well-ordered set X, we also define specific members of $X \setminus Y$ (as things that are not Y). In case X is On and Y is the set of all ordinals that satisfy some ϕ, we also defined the least member of $On \setminus Y$ as something that is not ϕ. Now, if we take ϕ as 'is definable', i.e. Y is DOn, we immediately define the least member not in DOn as something that is not DOn. In this respect, it seems unintuitive to assume there *is* DOn.

More on the totalities

In the previous section I argued that the totalities involved in the paradoxes (1)-(6) do not exist. According to the Inclosure Schema, these paradoxes depend on the existence of the totality. Taking the Inclosure Schema as the correct analysis of these paradoxes, they can be considered solved. Are there ways to prove the contradiction in these paradoxes without assuming the existence of the totalities?

In the case of Burali-Forti (1), Mirimanoff (2) and 5th Antinomy (3), it is easy to see that Ω is indispensable. Here, the contradiction cannot be *stated* without explicitly referring to the totality involved. Since the contradiction cannot be stated without referring to a totality, proving the contradiction depends on assuming that such a totality exists. The contradiction in the three paradoxes is stated as follows.

(1) $On \in On$ and $On \notin On$.

(2) $R \in R$ and $R \notin R$.

(3) $t(T) \in T$ and $t(T) \notin T$.

Even if we take the conceptual connection between satisfying a property $\phi(x)$ and being a member of a certain totality $\Omega := \{y| \ \phi(y)\}$ to hold, we cannot avoid reference to Ω. Applying the conceptual connection means to replace

'$x \in \Omega$' by '$\phi(x)$'. If we were to perform the replacement in these cases we would still end up with a statement involving the totality Ω:

(1) *On* is an ordinal and *On* is not an ordinal.

(2) R is well-founded and R is not well-founded.

(3) $t(T)$ is a thought (generated by t) and $t(T)$ is not a thought (generated by t).

Richard (6) also depends on the existence of Ω since the argument for contradiction essentially involves applying Cantor's diagonal argument on the totality of definable real numbers between 0 and 1, call it DR, to define a real number between 0 and 1 that cannot be in this totality[7].

Thus, in (1)-(3) and (6) the proof of contradiction depends on the premise that the totalities exist. Since this premise is false, as shown in claim (1), these paradoxes may be considered solved.

König (4) and Berry (5) are presented in claim 1 in a way that the existence of Ω is necessary. For the existence of $\delta(\Omega)$ is deduced by applying the least ordinal/number principle on the well-ordered set Ω. The least ordinal/number principle, however, does not need to be applied on a definite well-ordered set. The principle states that for any *property* of ordinals/natural numbers there is a least ordinal/natural number with that property. Certainly, there are ordinals with the property of not being definable and natural numbers with the property of not being definable in less than 99 words, as shown in claim 1 (4) and (5). Thus, if (4) and (5) are restated in terms of properties rather than totalities, one could still apply the least ordinal/number principle and deduce the existence of the least ordinal not definable (that is definable) or the least natural number not definable in less than 99 words (that is definable in this way). So after all, there are versions of (4) and (5) that do not depend on the existence of Ω. These versions do not apply the existence premise of the Inclosure Schema and therefore cannot be solved by denying the existence premise.

Can they be solved otherwise? Field mentions that there might be a problem with applying the least ordinal/number principle to vague or otherwise indeterminate concepts (2008, 100-101). The principle can be reduced to properties that satisfy LEM: If up to some ordinal n all smaller ordinals k satisfy LEM concerning some property F, i.e. $F(k)$ or $\neg F(k)$, then there is a smallest ordinal satisfying F. Since, as he argues, the predicate 'definable' does not satisfy LEM, the restriction of the least ordinal/number principle blocks the argument for contradiction.

[7] Alternatively, one could talk about the phrases defining real numbers between 0 and 1 instead of the real numbers defined by those phrases. One then applies Cantor's argument on the totality of all phrases that define a real number between 0 and 1 to get a defining phrase that cannot be in that totality.

Claim 2

Another criticism I want to add to is Tennant's point that the essential details of the paradoxes do not lie in the structure of the Inclosure Schema. Looking at the proofs of Transcendence one encounters two different groups. Transcendence, i.e. $\delta(x) \notin x$ for some $x \subseteq \Omega$, in the Liar, Russell (second version) and Grelling requires a proof by reductio ad absurdum. In particular, it is assumed that $\delta(x) \in x$ from which $\delta(x) \notin x$ can be derived using a bridge principle. Thus, in all three cases a contradiction can be derived due to the fact that one particular statement, i.e. $\delta(x) \in x$, implies its negation, $\delta(x) \notin x$. Transcendence in the other cases can be derived directly by virtue of how δ is defined. The proofs for Transcendence can therefore be put in two groups, one in which the result is proven by reductio using a statement that implies its negation and one in which the result is proven directly by virtue of the definition of δ. On this ground, I conclude that the essentials of the paradoxes of self-reference are not captured by the Inclosure schema.

The claim that the essentials of the paradoxes are not captured receives further support in chapter 4. We will see that the entire proof of contradiction in the Liar, Russell and Grelling can be carried out by assuming a statement that implies its own negation. That makes the three paradoxes entirely analogous in the way the contradiction can be proven. Moreover, we will see in chapter 4 how the Liar, Russell and Grelling can be stated and proven without assuming the existence of a totality. This constitutes another important difference to the paradoxes (1)-(6).

Claim 3

The self-referential nature of the paradoxes, as presented in the Inclosure Schema, lies in the impredicative definition of $\delta(\Omega)$. Accepting the Inclosure Schema as the correct uniform analysis of the paradoxes of self-reference is therefore to accept that what makes the paradoxes self-referential is that they involve impredicative definitions. This, however, constitutes a weakness of the Inclosure Schema for two reasons. First, it seems to distract from what makes the liar genuinely self-referential. Even though $\langle \alpha \notin Tr \rangle$ is an impredicative definition, there are versions of the liar that do not depend on this definition as argued by Badici, Tennant and Kroon. What one cannot escape in constructing the liar is a sentence that is $circular_1$ or $circular_2$. Second, even though Priest does not suggest giving up impredicative definitions as a uniform solution to the paradoxes[8], this solution would leave the Liar unsolved and declare non-paradoxical cases of impredicative definitions as problematic. Take, for instance, the definition of the smallest member of a set X, $min(X)$: $y = min(X)$ iff $y \in X$ for all $x \in X$ it holds that $y \leq x$. Since this definition generalizes over all $x \in X$ it is impredicative. Yet it is harmless[9].

[8] Russell introduced the Vicious Circle Principle to declare impredicative definitions unacceptable. However, it is questionable that this is the key to the paradoxes (Giaquinto, 2002, 69-84).

[9] Why this example is harmless is a matter of further discussion (Bernays, 1935; Gödel, 1944).

The second criticism concerns all notions of self-reference one can ascribe to the paradoxes. Giving up *circular*$_1$ and *circular*$_2$ would also declare non-paradoxical problematic. Is it possible to find a structure in all paradoxes of self-reference that could be given up to solve the paradoxes *without* declaring non-paradoxical cases problematic? The notion of self-negation, that will be introduced in chapter 4, is such a structure.

3.5 Conclusion

I argued that the totalities involved in Burali-Forti, Mirimanoff, 5th Antinomy and the paradoxes of definability are inconsistent by definition. The definition of the totalities contradicts facts about any set of objects that are members of the totalities. The existence premise of the Inclosure Schema is therefore not satisfied in these paradoxes. Taking the Inclosure Schema as the correct analysis of the paradoxes, it follows that they can be considered solved. The Inclosure Schema properly captures the structure of these paradoxes, yet, taking it as the underlying structure of the paradoxes, one fails in proving the contradiction and therefore fails in showing that these paradoxes are about truly inconsistent objects. We also saw that in some cases the totalities are indispensable for proving a contradiction. The claim that the totalities involved are inconsistent by definition may then be taken as a solution to these paradoxes. On the other hand, the Inclosure Schema does not capture what is essential to the Liar, Russell and Grelling, where the proof of Transcendence employs a distinct argument that is not made explicit by the schema: A proof by reductio using a statement that implies its negation. In the next chapter, I will show that the proofs for Russell, Liar and Grelling can be formulated and proven uniformly in what I call the Schema of Self-Negation. The notion of self-negation will prove a better alternative to the notion of self-reference in describing what is essential to the paradoxes of self-reference. For, unlike the notion of impredicativity or other notions of self-reference one might ascribe to the paradoxes, the notion of self-negation describes solely paradoxical cases, if one commits to classical logic, and can therefore be given up as a solution to the paradoxes without also dragging down non-paradoxical cases.

Chapter 4

A new Account of the Paradoxes of Self-Reference

We saw that among the paradoxes captured by the Inclosure Schema there are some that involve a totality that is inconsistent by definition, i.e. to assume that the totality exists is to contradict facts (see Ch. 3, claim 1). I conclude that the totalities in these cases do not exist. The existence premise of the Inclosure Schema is therefore not satisfied. If the Inclosure Schema is the correct analysis of these paradoxes, they can be considered solved. The Liar, Russell and Grelling, on the other hand, can be formulated and proven without reference to a totality. Whether or not the assumption of a totality is implicit due to a conceptual connection between satisfying a property and being a member of a certain totality *à la* Priest is irrelevant. What distinguishes this group of paradoxes from the other paradoxes of self-reference is that they are brought about in a highly analogous way. In particular, there are strong similarities between the assumptions and the derivation of the contradiction. The major objective of this chapter is to present a schema underlying the three paradoxes that captures the analogy between them. Before making the analogy explicit let us have a look at how the paradoxes are usually presented, here by Field (2008, Ch. 1):

Grelling's Paradox

Grelling's paradox is commonly presented as a paradox involving a predicate. For any predicate of our language that one substitutes for the letter '*F*', it seems that it holds that:

(TO) '*F*' is true of all and only the things that are *F*.

Consider the predicate 'is not true of itself' which is abbreviated by 'heterological'. According to (TO), 'heterological' is true of all and only those things that are heterological. Is 'heterological' true of 'heterological'? 'heterological' is heterological iff 'heterological' is not true of itself, i.e. 'heterological' is *not* heterological. This yields:

> ‘heterological’ is heterological iff ‘heterological’ is not heterological.

Having established a statement of the form ‘A iff $\neg A$’ we can show, analogously to the Liar in the introduction, that A and $\neg A$. Field (2008, 7-8) provides a general proof from ‘A iff $\neg A$’ to ‘A and $\neg A$’ and calls it the *Central Argument from Equivalence to Contradiction*.

> Step 1: The claim A iff $\neg A$ and the claim A together imply the claim A and $\neg A$.
> Clearly, the claim A iff $\neg A$ and the claim A together imply $\neg A$. Since A follows from A itself, we have A and $\neg A$.
>
> Step 2: The claim A iff $\neg A$ and the claim $\neg A$ together imply the claim A and $\neg A$.
> This is entirely analogous to Step 1.
>
> Step 3: The claim A iff $\neg A$ and the claim A or $\neg A$ together imply the claim A and $\neg A$.
> This follows from Step 1 and Step 2 and the rule of Reasoning by Cases: If assumptions Γ plus A imply C and Γ plus B imply C as well, then Γ plus the claim that A or B imply C.
>
> Step 4: The claim A iff $\neg A$ implies the claim A and $\neg A$.
> This step follows from Step 3 and LEM according to which the claim A or $\neg A$ is a logical truth.

By the Central Argument from Equivalence to Contradiction (CAEC), ‘heterological’ is heterological and not heterological.

Russell’s Paradox

There are two versions of Russell’s paradox, one that concerns properties (Russell 1) and the more familiar one that concerns sets (Russell 2). For **Russell 1**, consider the following apparently true principle about predicates. For any intelligible predicate of our language that one substitutes for the letter ‘F’ there is the property of being F such that:

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

Consider the property of not instantiating itself, the ‘Russell property’. According to (INST), the Russell property is instantiated by all and only those things that do not instantiate themselves. Does the Russell property instantiate itself? Again, we get:

> The Russell property instantiates itself iff it does not instantiate itself.

By CAEC, the Russell property instantiates itself and does not instantiate itself.

For **Russell 2**, consider the following apparently true principle governing the notion of set. For any intelligible predicate of our language that one substitutes for the letter '*F*' there is the set of things that are *F*, call it *y*, such that:

> (MEMB) The set *y* contains all and only the things that are *F*.

Consider the set of things that do not contain themselves, the Russell set. According to (MEMB),

> the Russell set contains itself iff it does not contain itself.

By CAEC, a contradiction follows.

The Liar

I have presented a version of the Liar paradox in the Introduction. Let me restate it more simply to reveal its analogy to the paradoxes just presented. If sentences are truth bearers, truth of sentences can be characterized by the T-schema. For all sentences *A*:

> (T) '*A*' is true iff *A*.

Now, consider a sentence *L*, the Liar, that states its own untruth, $L :=$ '*L* is not true'[1]. According to (T), *L* is true iff what *L* states is the case. Thus,

> *L* is true iff *L* is not true.

By CAEC, a contradiction follows.

The Analogy

Three major similarities between the four paradoxes (including Russell 1 and Russell 2) can be detected:

(1) Each paradox involves a fundamental concept of language and thought: Grelling and the Liar involve truth, applied to predicates and sentences, respectively. Russell 1 involves property instantiation and Russell 2 involves set membership[2].

(2) The fundamental concept involved in each paradox is governed by a rule that tells us what condition has to be satisfied for the concept to apply to something:

> For all intelligible predicates '*F*' and for all *x*:
>
> > (TO) '*F*' **is true of** *x* iff *x* is *F*.
>
> For all sentences *A*:
>
> > (T) '*A*' **is true** iff *A*.

[1] Note, that this is the strengthened Liar.

[2] We will see at the end of this chapter why the observation that each paradox involves a somehow fundamental concept is relevant.

For all intelligible predicates 'F' there is the property of being F such that for all x:

(INST) x **instantiates the property** of being F iff x is F.

For all intelligible predicates 'F' there is the set $y := \{z| \, z \text{ is } F\}$ such that for all x:

(MEMB) x **is a member of the set** $y := \{z| \, z \text{ is } F\}$ iff x is F.

(3) There is an object, to which the fundamental concept can be applied, that is defined such that the concept applies to that object iff it does not apply to that object:

'heterological' is true of 'heterological' iff 'heterological' is not true of 'heterological'.

L is true iff L is not true.

The Russell property R instantiates R iff R does not instantiate R.

The Russell set S is a member of $S := \{z| \, z \text{ is not a member of } z\}$ iff S is not a member of S

By CAEC, a contradiction follows.

4.1 Grelling's paradox reinterpreted

While Russell 1 and Russell 2 are considered two versions of the same paradox and Grelling is presented as a paradox involving linguistic objects, I proceed the discussion about the paradoxes by excluding Grelling, as presented above, from the discussion and use the label 'Grelling's Paradox' for Russell 1 instead. In this section I will argue for this decision. However, the results presented in this chapter are applicable to Grelling as well, if one insists on including the paradox.

First, Russell 1 seems to be entirely equivalent to Grelling. While Russell 1 involves properties, Grelling involves the linguistic objects referring to properties, predicates. To put it in Field's words, 'Grelling's paradox concerns linguistic expressions, but it is arguable that underlying it is a more basic paradox concerning properties, one that is in essence due to Russell' (Field, 2008, 1). Second, the concept of truth already appears in the Liar. If the paradoxes are understood as arguments that challenge the concepts involved in it, one might be inclined to think that either the Liar or Grelling is somehow redundant as they both challenge the same concept, i.e. the concept of truth. In what follows, I propose an interpretation of Grelling that reflects the two concerns just mentioned and is based on the view that truth is a concept primarily applicable to sentences (or what they express) rather than predicates. Obviously, this view can be questioned. Yet, let us assume that it holds for the sake of argument. The view implies that an expression of the form

(1) 'A predicate 'F' is *true of* x',

which appears in Grelling, is an abbreviation or different way of saying that

(2) '*x* instantiates the property of being *F*' *is true*.

Applying the T-schema on (2) amounts to saying that *x* instantiates the property of being *F* which is identical to the statement that appears in Russell 1. Given that (1) is to be interpreted as (2),

(3) For all intelligible predicates '*F*' and for all *x*:

> (TO) '*F*' is true of *x* iff *x* is *F*.

can be interpreted as

(4) For all intelligible predicates '*F*' there is the property of being *F* such that for all *x*:

> '*x* instantiates the property of being *F*' is true iff *x* is *F*.

Observe that (4) is just (INST) where '*x* instantiates the property of being *F*' is replaced by "*x* instantiates the property of being *F*' is true'. For (4) to hold, we need to have (T) as well. Thus, (4) is like (INST) with (T) applied on sentences of the form '*x* instantiates the property of being *F*'. I therefore call (4) '(T)+(INST)'.

Thus, Grelling can be interpreted as Russell 1 where every sentence of the form '*x* instantiates the property of being *F*' is replaced by "*x* instantiates the property of being *F*' is true' and then abbreviated by using the expression 'true of'. Assuming that (T) holds, the reinterpreted Grelling is just Russell 1 together with (T), let me call it Truth-Russell 1 and fully state it:

Truth-Russell 1

(i) For any intelligible predicate '*F*' there is the property of being *F* such that for all *x*:

> (T)+(INST) '*x* instantiates the property of being *F*' is true iff *x* is *F*.

(ii) Consider the Russell property *R*: *x* is *R* iff '*x* instantiates the property of being *x*' is not true.

(iii) It follows that

> 'the Russell property instantiates the Russell property' is true iff 'the Russell property instantiates the Russell property' is not true,

or the abbreviated version:

> 'heterological' is true of the Russell property iff 'heterological' is not true of the Russell property.

Truth-Russell 1 has no premises that do not appear in the other three paradoxes and therefore seems redundant. Obviously, this conclusion rests on the questionable view that (1) can be interpreted as (2). Accepting this view, however, would not only expose Grelling as Truth-Russell 1 but also expose another paradox as redundant:

Paradox 5

(i) For any intelligible predicate 'F' there is the set $y := \{z|\ z \text{ is } F\}$ such that for all x:

> The membership relation is true of (x, y), where $y := \{z|\ z \text{ is } F\}$, iff x is F.

(ii) Consider the set $R := \{z|$ the membership relation is not true of $(z, z)\}$.

(iii) The membership relation is true of (R, R) iff the membership relation is not true of (R, R).

Following the view that truth is primarily applicable to sentences, one could argue that sentences of the form

(5) 'the membership relation is *true of* the pair (x, y), where $y := \{z|\ z \text{ is } F\}$'.

are also just a different way of expressing sentences of the form

(6) "x is a member of the set $y := \{z|\ z \text{ is } F\}$' is true'.

If so, then (i) of Paradox 5 can be interpreted as

(7) For any intelligible predicate 'F' there is the set $y := \{z|\ z \text{ is } F\}$ such that for all x:

> 'x is a member of $y := \{z|\ z \text{ is } F\}$' is true iff x is F,

which is just (MEMB) where

(8) 'x is a member of the set $y := \{z|\ z \text{ is } F\}$'

is replaced by (6) using (T) and then replaced by (5). Let us call (7) '(T)+(MEMB)'. Thus, Paradox 5 is just Russell 2 with (T) applied on sentences of the form 'x is a member of the set $y := \{z|\ z \text{ is } F\}$'. I therefore call Paradox 5 'Truth-Russell 2':

Truth-Russell 2

(i) For any intelligible predicate 'F' there is the set $y := \{z|\ z \text{ is } F\}$ such that for all x:

> (T)+(MEMB) 'x is a member of the set $y := \{z|\ z \text{ is } F\}$' is true iff x is F.

(ii) Consider the set $R := \{z|$ 'z is a member of z' is true $\}$.

(iii) 'R is a member of R' is true iff 'R is a member of R' is not true .

Analogously to Russell 1 and Grelling, Russell 2 seems to be what is underlying Paradox 5. While Russell 2 concerns set membership, Paradox 5 concerns the linguistic object 'membership relation' referring to it. Moreover, both Grelling and Paradox 5 concern the concept of truth which also appears in the Liar. The idea that both paradoxes are somehow redundant is reinforced when one assumes that truth is primarily applicable to sentences or what they express rather than predicates or relations. According to this view, both paradoxes can be reinterpreted as what I call Truth-Russell 1 and Truth-Russell 2. The reinterpretation shows that these paradoxes use premises of the other three, Russell 1, Russell 2 and the Liar. If one is not convinced by this view and insists on treating Grelling and Paradox 5 like the other three, one can apply all coming results to all five paradoxes. For simplicity, however, I will take only Russell 1, Russell 2 and the Liar into consideration. Further, I will call Russell 1 'Grelling's Paradox' from now on to stress the point that it involves a concept different from Russell 2 and the Liar. Grelling, Russell and the Liar are three highly analogous paradoxes involving the fundamental concepts of property instantiation, set membership and truth (for sentences), respectively. The next section explores the schema underlying the three paradoxes.

4.2 The Schema of Self-Negation

Let ϕ be the concept of truth (for sentences), set membership or property instantiation. For the Liar ϕ semantically functions like a predicate, i.e. '$\phi(x)$' translates 'x is true'. For the other two paradoxes, ϕ functions like a two-place relation, i.e. '$\phi(x, y)$' translates 'x is a member of y' in Russell and 'x instantiates the property y' in Grelling. For the paradoxes, however, it suffices to treat ϕ as a predicate. I will therefore write '$\phi(x)$' to abbreviate '$\phi(x, x)$', so that '$\phi(x)$' translates 'x is a member of itself' in Russell and 'x instantiates itself' in Grelling. Before we arrive at a unified schema for the three paradoxes, I will first show how to derive the paradox in each case.

The Liar

(1) Let ϕ be the concept of truth:
Let X be a sentence. Then $\phi(X) := t(X)$.

(2) Let ψ be a closed formula. Then $X :=$ 'ψ' is a sentence where ψ is called the truth condition of X[3].

(3) A sentence is true iff its truth condition holds:
Let X be a sentence and ψ its truth condition. Then ψ iff $t(X)$.

(4) Let $\psi_S :=$ '$\neg t(\psi_S)$'[4].

[3] I do not intend to make a statement about truth-conditions, I rather just *call* a sentence its truth condition to stress the analogy to Russell and Grelling (see point (2) in Russell and Grelling).

[4] Note that the truth condition is a sentence and can be defined circularly by some form of self-reference, e.g. as a fixed point of '$\neg t(x)$'.

(5) By (2), there is a sentence with the truth condition ϕ_S, call it $S :=$ '$\neg t(S)$'.

(6) claim: $t(S)$ iff $\neg t(S)$
Proof. $t(S) \overset{(5)}{\text{iff}} t(\psi_S) \overset{(4)}{\text{iff}} t(\text{'}\neg t(S)\text{'}) \overset{(3)}{\text{iff}} \neg t(S)$.

(7) By (6) and (CAEC): $t(S)$ and $\neg t(S)$.

Russell's Paradox

(1) Let ϕ be the concept of membership:
Let X be a set. Then $\phi(X) := X \in X$.

(2) Let $\psi(y)$ be a formula with 'y' as its only free variable. Then $X := \{y|\psi(y)\}$ is a set. Let $\psi(y)$ be called the membership condition of X.

(3) Y is a member of set X iff Y satisfies X's membership condition:
Let $X := \{y|\psi(y)\}$ be a set. Then for all Y: $Y \in X$ iff $\psi(Y)$.

(4) Let $\psi_S(y) := y \notin y$.

(5) By (2), there is a set with the membership condition ψ_S, call it $S := \{y|y \notin y\}$.

(6) claim: $S \in S$ iff $S \notin S$
Proof. $S \in S \overset{(5)}{\text{iff}} S \in \{y|\psi_S(y)\} \overset{(4)}{\text{iff}} S \in \{y|y \notin y\} \overset{(3)}{\text{iff}} S \notin S$.

(7) By (6) and (CAEC): $S \in S$ and $S \notin S$.

Grelling's Paradox

(1) Let ϕ be the concept of instantiation:
Let X be a property. Then $\phi(X) := X(X)$.

(2) Let $\psi(y)$ be a formula with 'y' as its only free variable. Then $X(y) := \psi(y)$ is a property. Let ψ be called the instantiation condition of X.

(3) Y instantiates the property X iff Y satisfies X's instantiation condition:
Let $X(y) := \psi(y)$ be a property. Then for all Y: $X(Y)$ iff $\psi(Y)$.

(4) Let $\psi_S(y) := \neg y(y)$.

(5) By (2), there is a property with the instantiation condition ψ_S, call it $S(y) := \neg y(y)$.

(6) claim: $S(S)$ iff $\neg S(S)$
Proof. $S(S) \overset{(5)}{\text{iff}} \psi_S(S) \overset{(4)}{\text{iff}} \neg S(S)$[5] $\overset{(3)}{\text{iff}} \neg S(S)$[6].

(7) By (6) and (CAEC): $S(S)$ and $\neg S(S)$.

[5] This is saying that S satisfies the instantiation condition of S.

[6] This is saying that S instantiates the property S.

There is a clear analogy between the assumptions of the three paradoxes and the way to prove a contradiction. More strikingly, the way to construct and prove the claim in step (6) of each paradox is analogous. We can therefore give the unified Schema for the paradoxes by making the analogy explicit.

The Schema of Self-Negation

(1) Let ϕ be a fundamental concept.

(2) For any condition ψ there is an object x, s.t. $\phi(x)$ is well defined, that is defined by ψ.

(3) For all x s.t. $\phi(x)$ is well-defined: x is ϕ iff x satisfies its condition ψ.

(4) Let ψ_S be a condition defined in terms of '$\neg\phi$'.

(5) By (2), there is an object, call it S, that is defined by ψ_S.

(6) claim: $\phi(S)$ iff $\neg\phi(S)$
Proof.

$$\phi(S)$$

$\overset{(5)}{\text{iff}}$ $\phi(S)[\psi_S]$ (where S is stated in terms of its defining condition ψ_S)

$\overset{(4)}{\text{iff}}$ $\phi(S)[\neg\phi]$ (where the definition of ψ_S is applied)

$\overset{(3)}{\text{iff}}$ $\neg\phi(S)$

(7) By (5) and (CAEC): $\phi(S)$ and $\neg\phi(S)$.

Let me clarify what it means that ψ_S is defined *in terms of* '$\neg\phi$' in step (4). Note, that $\phi(y)$ is a formula with 'y' as its only free variable. For ψ_S, however, we have to distinguish the two cases (i) where it is a formula with 'y' as its only free variable, $\psi_S(y)$, as in Russell and Grelling, and (ii) where it is a sentence, ψ_S, as in the Liar. In case (i), since both $\psi_S(y)$ and $\phi(y)$ are formulas with 'y' as its only free variable, to define ψ_S in terms of '$\neg\phi$' means that $\psi_S(y) := \neg\phi(y)$. It is not enough for $\psi_S(y)$ to logically entail $\neg\phi(y)$. Consider, for instance, the set $S' := \{z \mid z \neq z\}$ which is the empty set. Here, $\psi'_S(y) = y \neq y$ logically entails $\neg\phi(y)$, i.e. $y \notin y$, since anything follows from a contradiction. Yet, we cannot infer $\phi(S')$ iff $\neg\phi(S')$ for $\neg\phi(S')$, i.e. $S' \notin S'$, does not entail $S' \in S'$. $S' \notin S'$ is true since S' is the empty set while $S' \in S'$ is false since the empty set has no members. In case (ii), to define ψ_S in terms of $\neg\phi$ means that ψ_S is (a) a self-referential sentence and (b) stating (of itself) that it is $\neg\phi$, $\psi_S :=$ '$\neg\phi(\psi_S)$', e.g. as a fixed point of $\neg\phi$. Here again, it is not enough to require ψ_S to be (a) and to state something (of itself) that logically entails that it is $\neg\phi$. Consider, for instance, the sentence $S' :=$ '$t(S') \wedge \neg t(S')$' which is self-referential and implies '$\neg t(S')$'. Yet, '$\neg t(S')$' does not imply '$t(S') \wedge \neg t(S')$' and hence we do not have $\phi(S')$ iff $\neg\phi(S')$.

Why is it called the Schema of Self-Negation? Steps (1)-(3) are to the effect that for any object x, s.t. $\phi(x)$ is well defined, there is a ψ s.t. $\phi(x)$ iff $\psi(x)$ or, in case of the Liar, $\phi(x)$ iff ψ. In step (4), we define such a ψ in terms of '$\neg\phi$' so that we can prove in (6) that

$\phi(S)$ iff $\neg\phi(S)$.

The fact that $\phi(S)$ iff $\neg\phi(S)$ is what makes $\phi(S)$ a *self-negating* statement. In general, a statement A negates itself iff it is the case that A iff $\neg A$. Here, A is about the condition ϕ. 'A iff $\neg A$' amounts to saying that S satisfies ϕ iff it does not satisfy ϕ. In other words, ϕ is a condition that negates itself when applied to S. The satisfaction of $\phi(S)$ consists of its non-satisfaction. In particular, the truth condition of the Liar paradox negates itself as it is satisfied by not being satisfied. Likewise, Russell's set satisfies its membership condition by not satisfying it and the property labeled 'heterological' satisfies its instantiation condition by not satisfying it. Step (4) is the crucial part of the schema as it is the step where S is defined in a way that $\phi(S)$ becomes a self-negating statement.

Note that self-negation can also be achieved with a weaker version of (4): Let ψ_S be a condition defined in terms of 'ϕ'' where $(*)$ ϕ' is logically equivalent (in classical logic) to $\neg\phi$. For to prove $\phi(S)$ iff $\neg\phi(S)$ would require only one extra step:

$\phi(S)$

$\overset{(5)}{\text{iff}}$ $\phi(S)[\psi_S]$ (where S is stated in terms of its defining condition ψ_S)

$\overset{(4)}{\text{iff}}$ $\phi(S)[\phi']$ (where the definition of ψ_S is applied)

$\overset{(3)}{\text{iff}}$ $\phi'(S)$

$\overset{(*)}{\text{iff}}$ $\neg\phi(S)$

In case of the Liar, for instance, take $\phi'(x) :=$ '$\neg\phi(x) \wedge (P(x) \vee \neg P(x))$', which is logically equivalent to '$\neg\phi(x)$', so that we get the sentence S':='$\neg t(S') \wedge (P(S') \vee \neg P(S'))$'. By (3)-(5) of the Liar, $t(S')$ iff ($\neg t(S')$ and ($P(S')$ or $\neg P(S')$)) which is equivalent to $\neg t(S')$. In case of Russell, as another example, we could take $\phi'(x) := \phi(x) \rightarrow x \neq x$, where '$\rightarrow$' is the material implication which means that $\phi'(x)$ is logically equivalent to $\neg\phi(x)$, to define the set $S' := \{z| \, z \in z \rightarrow z \neq z\}$. By (3)-(5) in Russell, $S' \in S'$ iff (if $S' \in S'$ then $S' \neq S'$) which is equivalent to $S' \notin S'$.

In (4) and (5) of the schema of self-negation, there is some form of self-reference involved. The Liar is a self-referential sentence. In Russell and Grelling, S is defined impredicatively. However, the schema shows that it is not self-reference on its own that leads to contradiction. The Liar sentence, the Russell set and the Russell property (Grelling here) are not just *some* circularly defined objects. It is the fact that they are defined in such a way that

$\phi(S)$ becomes a self-negating statement which leads to contradiction, i.e. they are objects with a condition ψ that is defined in terms of $\neg\phi$.

An evaluation of the schema

Unlike the Inclosure Schema, the schema of self-negation captures the striking similarity between the three paradoxes that distinguishes them from the other paradoxes of self-reference. Thereby it captures every step towards the contradiction and therefore provides are more fine-grained analysis of the paradoxes. The schema does not explicitly mention the premise that we can construct self-referential sentences as in the liar or use impredicative definitions as in Russell and Grelling, but it is implicit in step (4).

Yet, it is not self-reference alone but self-negation that, given CAEC (or, broadly speaking, classical logic), leads to contradiction. The schema thus pinpoints a particular structure in the paradoxes that cannot be found in non-paradoxical cases and thereby captures what is essential to them. This is an advantage over analyses of the paradoxes according to which it is self-reference that characterizes the paradoxes such as the Inclosure Schema. For self-reference can also be found in non-paradoxical cases and therefore does not suffice to characterize the paradoxes. If one were to solve the paradoxes by giving up self-negation, one would not have to give up non-paradoxical cases.

Giving up self-negation means to give up one of the premises (1)-(4) of which premise (4) seems to be particularly contentious. For it allows to *construct* the object S in a way that $\phi(S)$ iff $\neg\phi(S)$. Alternatively, one could say that rather than the construction of S it is premise (2) that is contentious for it says that whatever we choose ψ_S to be there *is* an object that is defined by ψ_S. Taking a stand on the question which premise is the trouble maker goes beyond the purpose of this paper.

Of course one does not have to deny self-negation to solve the paradoxes. The other solution routes presented in the introduction can also be applied. One can follow solution route (I), i.e. give up or replace the concept of truth, by denying step (1) of the schema. Denying step (3), the T-schema, is to follow solution route (III). More importantly, CAEC of step (7) involves classical logic, mainly LEM, which includes paracomplete solutions like the one by Kripke or Field, see solution route (IV), in the list of possible responses to the three paradoxes.

Obviously, the Inclosure Schema captures more paradoxes. Most of these paradoxes, however, as argued in the previous chapter, have a false premise. The Knower, Gödel and Berkeley, however, have not been shown to involve self-negation. Whether they involve self-negation or have false premises, when analyzed by the Inclosure Schema, is left for further studies. Yet, there are paradoxes that are either not captured by the Inclosure Schema, No-no, or controversial concerning whether they are self-referential, Yablo. In the next two chapters, I present an analysis of No-no and Yablo that involves self-negation.

Before we arrive there, let us have a look at a case of self-negation that is easily solvable and should be separated from the paradoxes.

4.3 Another case of Self-Negation

There are other well-known paradoxes that involve self-negation. The most prominent one is about a barber who shaves all and only those men that do not shave themselves. Does he shave himself? Just by how he is defined, he shaves himself iff he does not shave himself, self-negation. Does this paradox fit the schema of self-negation? The standard and uncontroversial solution to the barber is to say that he does not exist. Can we apply a similar solution to the Liar, Russell and Grelling? To answer these questions, let me reveal the structure of the Barber paradox by comparing it to other structurally similar examples:

(1) The barber, who shaves all and only those men that do not shave themselves, shaves himself iff he does not shave himself.

(2) The list, that lists all and only those lists that do not list themselves, lists itself iff it does not list itself.

(3) The website, that links all and only those websites that do not link themselves, links itself iff it does not link itself.

(4) The killer, who kills all and only those people that do not kill themselves, kills herself iff she does not kill herself.

The underlying structure consists of four steps. Let ι be the description operator so that $\iota x[\psi(x)]$ translates 'the/a thing that is ψ'.

Schema 2

1. Let $\phi(x, y)$ be a binary concept:

   (1) $\phi(x, y)$ is 'x shaves y'.

   (2) $\phi(x, y)$ is 'x lists y'.

   (3) $\phi(x, y)$ is 'x links y'.

   (4) $\phi(x, y)$ is 'x kills y'.

2. We define an object $S := \iota x[\text{ for all } y : \phi(x, y) \text{ iff } \neg\phi(x, x)]$.

3. By 2. and the Characterization Principle[7], CP: for all y : $\phi(S, y)$ iff $\neg\phi(S, S)$.

4. By 3. and $y = S$: $\phi(S, S)$ iff $\neg\phi(S, S)$.

[7] This is a principle according to which an object that is defined to be a or the thing that has such and such properties has indeed those properties. The principle was named 'Characterization Principle' by Routley (1980).

For a paradox to fit this schema it is enough to show that there is some binary concept and an object, such that the concept can be applied to that object, defined as in 1. By CP and a little bit of classical logic we can deduce self-negation, here $\phi(S,S)$ iff $\neg\phi(S,S)$.

The main difference to the schema of self-negation is that the self-negating statement is derived using CP rather than principles governing the concept ϕ. According to the schema of self-negation, we get $\phi(x)$ iff $\neg\phi(x)$ since (i) there are principles (the steps (1)-(3) of the schema) telling us that there is a ψ such that $\phi(x)$ iff ψ and (ii) we can choose ψ accordingly (step (4)). In Schema 2, there is no such ψ. In the Barber, for instance, we have no ψ, such that x shaves y iff ψ, which we can choose as we wish. On the contrary, for x to be in a set y or to instantiate some property z, as in Russell and Grelling for instance, x has to satisfy y's membership condition or z's instantiation condition, respectively, of which both depend on how we define them.

The difference between the schema of self-negation and Schema 2 can also be brought about by the fact that the concepts involved in the former seem to be of a special kind. Whatever this kind might be characterized by, e.g. being fundamental or general, there is no restriction on what concepts are allowed in the Schema 2. We could take ϕ in schema 2 to be '$x \in y$' or '$x(y)$' and show that Russell and Grelling do also fit Schema 2. However, this does not work the other way around. The examples (1)-(4) for Schema 2 do not fit in the schema of self-negation since the concepts involved are not of the nature described above: there is no condition ψ such that $\phi(x,y)$ iff ψ.

4.4 Conclusion

In this chapter I presented a schema, the Schema of Self-Negation, for the Liar, Russell and Grelling that reveals the highly analogous proof of contradiction via a *self-negating* statement. The schema also reveals that what distinguishes the three paradoxes is mainly that they involve different, yet arguably fundamental concepts, i.e. truth, membership and instantiation. From a classical point of view, self-negation leads to contradiction and therefore the schema captures what is essential to the three paradoxes.

Despite the fact that self-negation describes what is essential to the paradoxes from a classical point of view, it is not clear what the consequences of the new analysis are. Does self-negation suggest a solution different from the Inclosure Schema? What does self-negation tell us about the paradoxes that is not captured by the Inclosure Schema? Further, it is obvious that the Inclosure Schema captures more paradoxes. Most of these paradoxes as argued in the previous part, have a false premise. The Knower, Gödel and Berkeley, however, have not been shown to involve self-negation. Whether they involve self-negation or have false premises, when analyzed by the Inclosure Schema, is left for further studies.

On the other hand, there are paradoxes which are either not captured by the Inclosure Schema, No-no, or controversial concerning whether they involve self-reference, Yablo. In the next two chapters, I will examine in how far No-no and Yablo involve self-negation.

Chapter 5

The No-no Paradox

In the previous chapter I established a schema for the Liar, Russell and Grelling that involves self-negation. The schema thereby captures an essential detail of the three paradoxes that is not captured by the Inclosure Schema. Yet, there are paradoxes that fit in the Inclosure Schema but have not yet been shown to involve self-negation, e.g. the Knower, Gödel and Berkeley. On the other hand, there is a parodox, the No-no paradox, that does not fit the Inclosure Schema. In this chapter I will present an analysis of this paradox that involves self-negation. I begin with an informal presentation of the paradox.

5.1 The paradox

John Buridan (Hughes, 1982, 73-79) presents a variant of the Liar paradox, called No-no paradox (Sorensen, 2001) or the open pair (Armour-Garb and Woodbridge, 2006), that involves two sentences A and B stating of each other that they are false:

> A:= 'B is false.'
>
> B:= 'A is false.'

In contrast to the Liar, there is a consistent way to assign truth values to A and B, as long as A and B have different truth values. For instance, if A is true and B is false then A truly states that B is false and B falsely states that A is false. The contradiction arises when we assign the same truth values to A and B. For both cannot truly state of each other that they are false and both cannot falsely state of each other that they are true. For this reason we can establish a principle that Armour-Garb and Woodbridge call DA:

> (DA) If each statement in the open pair has a unique truth-value, then each has the opposite value of the other.

Note, that the antecedent of (DA) is only for those important who question either (the semantic version of) the law of Non-Contradiction (LNC) or Bivalence. For both principles together entail that every sentence has a unique truth value. According to Bivalence, each sentences has at least one of the truth values 'true' or 'false' and LNC states that it can only be at most one of the two. I am citing Goldstein's (2009) version of (DA) who, as we have seen in Ch. 1.3, endorses the view that not every sentence is a statement, i.e. not every sentence has a truth value. However, I thereby do not intend to take a stand on the issue.

So far, there seems to be no problem. We can avoid contradiction by assigning different truth values to A and B. The paradox arises when we consider that both A and B are in a symmetric relation to each other. 'There is no more reason why [A] should be true, or false, than [B], or vice versa, since they stand in an exactly similar relation to each other. So I shall assume that if either is true, so is the other, and if either is false, so is the other' (Hughes, 1982, 73)[1]. We can therefore establish a second principle:

> (SA) If each statement in the open pair has a unique truth-value, then each has the same value of the other.

SA is not a principle that can be verified by logical inspection as DA, but rather 'a principle of reasonableness' (Goldstein, 2009, 378). DA and SA together entail that if A and B have a unique truth value then they both have different truth values and the same truth values. So either (i) they each do not have a unique truth value or (ii) they both have different truth values and the same truth values. Option (i) entails that either LEM or Bivalence does not hold. Option (ii) means that (iia) either (A and B are true) or (A and B are false) and (iib) (A is true and B is false) or (A is false and B is true). In any case of (ii) we can derive that A and B are both true and false[2]. Thus, option (ii) entails the contradiction that A and B are both true and false and hence goes against LNC. Summing up, if both A and B have a unique truth value, then the contradiction of the No-no paradox is that A and B are both true and false.

The paradox partly rests on a symmetry assumption which is not only stated vaguely but is also defended poorly. What does it mean that A and B are symmetric and why does this imply that if they each have a unique truth value then they have the same truth value? Let me try to answer this question by reinterpreting the symmetry assumption.

[1] A similar situation arises when we consider the so-called truth teller: A:= 'A is true.' There is a consistent way to assign either truth value to A. If A is true, it is true. If A is false, it is false. Yet, there seems to be no reason to favor one option over the other. Armour-Garb and Woodbridge (2006) therefore distinguish between two cases of *semantic pathology*: inconsistency as in the Liar and indeterminacy as in the truth-teller. Krike's solution (see. Ch. 1.3) applies to all cases, the Liar, the truth teller and No-no, for all sentences involved are ungrounded. However, in the same paper of 2006, Armour-Garb and Woodbridge argue that there are variants of No-no that remain unsolved and problematic not only for Kripke's solution but also for dialetheists.

[2] *Proof.* By (iia), we have that either (A and B are true) or (A and B are false). If A and B are true then, by (T), A and B are false. If A and B are false then, by (T), A and B are true.

5.2 The Symmetry Assumption

The symmetry between A and B can be stated as follows. Let R be a binary relation such that

for all $x, y : (x, y) \in R$ iff $x :=$ 'y is false'.

It is easy to see that both (A,B)$\in R$ and (B,A)$\in R$ which makes A and B symmetric concerning the relation R. If A and B have a unique truth value then they state of each other that they are false[3]. Thus, the symmetry between A and B consists of the fact that, if they each have a unique truth value, they both *make the same statement* about each other, i.e. they both assign the truth value 'false' to each other[4]. If one would argue for the symmetry assumption, one could go as follows.

(P1) If A and B each have a unique truth value, they make the same statement about each other.

(P2) If A and B make the same statement about each other, they have the same truth value.

(C) If A and B each have a unique truth value, they have the same truth value.

The crucial premise of this argument is (P2). Why is it plausible that A and B have the same truth value even though they refer to different objects? Of course there can be cases in which two equivalent sentences refer to different objects, e.g. '$\langle P(a)\rangle = \langle P(a)\rangle$' and '$\langle\neg P(a)\rangle = \langle\neg P(a)\rangle$' (see 2.2) since they are both logically true. The explanation for the case of No-no must have something to do with the fact that A and B *make the same statement* about each other. Is there a way to make the plausibility of (P2) more explicit?

In the following I propose an equivalent version of No-no that makes the symmetry argument more plausible by examining what it means for two sentences to make the same statement about each other. For simplicity, I will consider only languages without indexicality. Two sentences X and Y make the same statement about each other iff

(i) X and Y have a unique truth value and

(ii) there is a formula $\phi(x)$ with 'x' as its only free variable such that

(a) 'X' and 'Y' do not occur in $\phi(x)$ and

(b) $X :=$ '$\phi(Y)$' and $Y :=$ '$\phi(X)$'.

Condition (iia) ensures that not both X and Y refer to X or Y. In other words, of the two sentences it is only X that refers to Y and only Y that refers to X.

[3] The antecendent of this sentence ensures that A and B make a statement at all. Following Strawson (1950), sentences make a statement only if they have a unique truth value.

[4] I will give a formal definition of what it means for two sentences to make the same statement about each other in due course.

Now, let X and Y be two sentences satisfying (i) and (ii) with some formula $\phi(x)$ with 'x' as its only free variable. Let $S := \{X, Y\}$. Since S has only two members it is easy to see that

(iii) $\phi(Y)$ iff for all $Z \in S$: if $Z \neq X$ then $\phi(Z)$, and

(iv) $\phi(X)$ iff for all $Z \in S$: if $Z \neq Y$ then $\phi(Z)$.

> *Proof.* It suffices to show (iii). Let $\phi(Y)$. Then, if $X \neq Y$ then $\phi(Y)$. Further, we have that (if $X \neq X$ then $\phi(X)$) since the antecedent is false. Thus we get for all $Z \in S$: if $Z \neq X$ then $\phi(Z)$. Now, let for all $Z \in S$: if $Z \neq X$ then $\phi(Z)$. Then, if $X \neq Y$ then $\phi(Y)$. Since the antecedent if true, we get $\phi(Y)$.

Each sentence in S, a set of two sentences satisfying (ii), is equivalent to saying that all sentences in S other than itself are ϕ. Thus, all sentences U in S are equivalent to a sentence of the form

(v) 'For all $Z \in S$: if $Z \neq U$ then $\phi(Z)$'.

That means that in case two sentences make the same statement about each other we can replace both sentences by a sentence of type (v) which is equivalent to both sentences and translates 'all sentences in S other than me are ϕ', assuming there is a set S containing the two sentences.

Let us rephrase A and B of the No-no paradox in terms of (v). Let $S := \{A', B'\}$ be the set of two sentences such that

> A$'$:='For all $Y \in S$: if $Y \neq$ A$'$ then Y is false' and
>
> B$'$:='For all $Y \in S$: if $Y \neq$ B$'$ then Y is false'.

The crucial feature of this version of No-no is that A$'$ and B$'$ can be defined *uniformly*. For all $X \in S$:

> X:='For all $Y \in S$: if $Y \neq X$ then Y is false'.

Given this uniform definition of the sentences of S, they satisfy two interesting conditions. There is a formula $\phi(x)$ with 'x' as its only free variable such that

(1) for all $X \in S$: '$X := \phi(X)$' and

(2) 'A$'$' and 'B$'$' do not occur in $\phi(x)$.

Here, $\phi(x)$='For all $Y \in S$: if $Y \neq x$ then Y is false'. As we can see, (1) and (2) are satisfied. These two conditions allow to conclude that both A$'$ and B$'$ have the same truth value: By (1), we know that all X in S are syntactically the same except for the name of each sentence that is substituted for the free variable in $\phi(x)$. At first sight, it seems that this already justifies the conclusion that all X have the same truth value, yet it is not enough. Consider the case in which some of the names that we choose for the members S occur in $\phi(x)$. Let, for instance, 'A$'$' but not 'B$'$' occur in $\phi(x)$. Then *both* A$'$='$\phi(A')$' and B$'$='$\phi(B')$' refer to A$'$. But only B$'$ refers to B$'$. In this case, we cannot assure that both A′ and B′ have the same truth value[5]. By (2), we exclude this case and ensure that among the sentences in S only A′ refers to A′ and only B′ refers to B′. By (1) and (2), whether we choose ‘A′’ or ‘B′’ for X in ‘$X :=$ ‘$\phi(X)$’’, X is the *same self-referential sentence*, i.e. it is a token of the same type of (self-referential) sentence. In No-no, the type is X:=‘For all $Y \in S$: if $Y \neq X$ then Y is false’ which translates ‘All sentences of S other than *me* are false’. Since both sentences in S are the same self-referential sentence, they must have the same truth value, if they each have a unique truth value.

The claim that A′ and B′ have the same truth value can thus be based on the fact that they are the same self-referential sentence. The argument for A′ and B′ having the same truth value is then:

(P1′) If A′ and B′ each have a unique truth value, they make the same self-referential statement.

(P2′) If A′ and B′ make the same self-referential statement, they have the same truth value.

(C′) If A′ and B′ each have a unique truth value, they have the same truth value.

To sum up, the claim that A and B have the same truth value, if they each have a unique one, becomes obvious by the observation that they are equivalent to A′ and B′, respectively. For A′ and B′ are equivalent, i.e. they have the same truth value, since they are both the same self-referential sentence. In particular, both sentences in No-no are equivalent to the sentence translating ‘All sentences of S other than me are false’, where S is the set containing both sentences. From now on, when talking about No-no, I refer to the set N of two sentences such that for all $X \in N$:

X:=‘For all $Y \in N$: if $Y \neq X$ then Y is false’.

The equivalence between both A and B and the self-referential sentence saying that all sentences of S other than itself are false, where S is the set containing A and B, only stands if we assume that there is set S, or alternatively, any kind of unity that contains A and B, e.g. a bunch, etc. It therefore has to be admitted that the argument for the plausibility of the symmetry assumption requires an additional existence assumption. However, the additional assumption is fairly reasonable. Further, it comes with the advantage that the symmetry assumption can be made obvious.

5.3 Self-Negation in No-no

The new interpretation of No-no, the set N of two sentences each stating of all other sentences in N that they are false, does not only explain why both

[5] Consider, for instance, $\phi(x)$:=‘The name of x is ‘A′’’. Then A′:=‘The name of A′ is ‘A′’’ and B′:=‘The name of B′ is ‘A′’’ have different truth values. For A′ is true since this is how we named the sentence ‘The name of A′ is ‘A′’’, while B′ is false since we did not name it A′.

sentences in No-no have the same truth value. It also allows to discover a self-negating statement in this paradox. In particular, the claim that both sentences of S have the same truth value is one of two steps towards self-negation.

Step 1

Recall that both sentences of N have the same truth value since they are the same self-referential sentence, i.e. they both translate 'All sentences of N other than me are false'. The fact that they are the same self-referential sentence is due to two conditions that both sentences of N satisfy. However, these two conditions can be met by more than two sentences.

In general, let $\phi(x)$ be a formula with 'x' as its only free variable and S a set of sentences such that for all $X \in S$:

(1) X:='$\phi(X)$' and

(2) 'X' does not occur in $\phi(x)$.

These two conditions ensure that all sentences of S are the same self-referential sentences. For (2) ensures that the only distinct reference in each sentence is the one to itself while (1) ensures that all sentences of S are syntactically the same apart from what is substituted for the free variable in $\phi(x)$.

Let S be a set of sentences satisfying (1) and (2). Then all sentences of S have the same truth value, i.e. for all $X, Y \in S$: $t(X)$ iff $t(Y)$. This means that *either* (a) for all $X \in S$: $t(X)$ *or* (b) for all $X \in S$: $\neg t(X)$ and there is no third option. Then (a) is equivalent to the negation of (b) and vice versa, i.e.

> (i) for all $X \in S$: $t(X)$ iff not for all $X \in S$: $\neg t(X)$ and
> (ii) for all $X \in S$: $\neg t(X)$ iff not for all $X \in S$: $t(X)$.
>
> *Proof.* It suffices to show (i). Let for all $X \in S$: $t(X)$. This trivially implies that not for all $X \in S$: $\neg t(X)$. Conversely, assume that not for all $X \in S$: $\neg t(X)$. Then (b) cannot be the case. Since we have that either (a) or (b), we must have (a), i.e. for all $X \in S$: $t(X)$.

To capture the fact that (a) and (b) are each other's negation, let us define :

> $t(S)$ iff for all $X \in S$: $t(X)$ and
>
> $\neg t(S)$ iff for all $X \in S$: $\neg t(X)$.

'$t(S)$' is not supposed to explain what it means for a set of sentences to be true. It is rather to be understood as an abbreviation for the fact that all sentences of S, which are either all true or all false, are true. However, given that all sentences of S in No-no satisfy (1) and (2), it seems reasonable to define a truth value for S. For, by (1) and (2), all sentences of S make the same self-referential statement. They only differ in their name and hence in what name they use to refer to themselves. Thus, the sentences of S are just different tokens of the same type, e.g. in No-no the type is Y:='For all $X \in N$: if $X \neq Y$ then X is false', and the truth value for S is the truth value for that type of sentence. Summing up, for any set S of sentences that satisfy (1) and (2) we have that

(3) either $t(S)$ or $\neg t(S)$

Since N satisfies (1) and (2), we have either $t(N)$ or $\neg t(N)$.

Step 2

In case of No-no there is another important feature. Each sentence of N states of all other sentences of N that they are *false*. Thus, for each sentence X of N there is a subset of N containing sentences of which X is saying that they are false. Let for each $X \in N$, F_X be the subset of N that contains all and only those sentences of N of which X states that they are false. Then for all $X \in N$:

(4) X iff for all $Z \in F_X : \neg t(Z)$.

> *Proof.* Let $X \in N$ and let Y be the sentence in N s.t. $Y \neq X$. By the definition of X, X iff for all $Y \in N$: if $X \neq Y$ then Y is false. Now, assume X. Then Y is false. By (iv) of section 5.2, the set of sentences of N of which X states that they are false contains only Y, $F_X = \{Y\}$. Thus, we have that for all $Z \in F_X : \neg t(Z)$. Conversely, assume for all $Z \in F_X : \neg t(Z)$. Then Y is false. Since $X = X$ we have that for all $Y \in N$: if $X \neq Y$ then Y is false, i.e. X.

Now, if *all* sentences of N are false, then what each sentence of N states is true since they state of all other sentences that they are false. Conversely, if all sentences of N are true, then what each sentence of N states cannot be true since they state of all other sentences that they are false. In general, let S be a set of sentences such that for all $X \in S$ there is a $F_X \in \mathcal{P}(S)$ such that (4). Then:

(5) For all $X \in S : t(X)$ iff for all $X \in S : \neg t(X)$

> *Proof.* Let for all $X \in S : t(X)$. Then for all $X \in S$: there is a $Z \in F_X$: $t(Z)$. By (4) and (T), for all $X \in S : \neg t(X)$.
> Conversely, let for all $X \in S : \neg t(X)$. Then for all $X \in S$: for all $Z \in F_X$: $\neg t(Z)$. By (4) and (T), for all $X \in S : t(X)$.

Since N satisfies (4), we have that for all $X \in N : t(X)$ iff for all $X \in N : \neg t(X)$. Given that N *also* satisfies (1) and (2) of Step 1, we also have that either for all $X \in N : t(X)$ or for all $X \in N : \neg t(X)$. In terms of the definition of Step 1, we have either $t(N)$ or $\neg t(N)$ *and* $t(N)$ iff $\neg t(N)$. Thus, '$t(N)$' is a self-negating statement, since by Step 2 '$t(N)$' is equivalent to '$\neg t(N)$' which by Step 1 is equivalent to the negation of '$t(N)$'.

Summing up, we saw that the No-no paradox partly rests on the claim that both sentences in No-no must have the same truth value, if they each have a unique one, the symmetry assumption. So far this assumption is taken to be true because it seems plausible. In 5.2, I presented an attempt to explain the symmetry assumption by showing that, given the set N containing both sentences of No-no, they both are equivalent to the same self-referential sentence, i.e. 'all other sentences in N are false'. In 5.3 Step 1, I showed that if the sentences of a set S of sentences are either (a) all true or (b) all false, as in No-no, then (a) and (b) are logical negations of each other. In 5.3 Step 2, I showed that for all sets of sentences S where each sentence is equivalent to stating that some subset of S is false, as in No-no, have the property that all its members are false iff they are all true, i.e. (a) iff (b). Since in No-no both (a) and (b) are logical negations of each other and (a) iff (b) is the case, we established a self-negating statement. According to this analysis of No-no, the contradiction is (a) and (b), i.e. both sentences of No-no are both true and false. Alternatively, one can deny that each sentence in N has a unique truth value. Then one cannot derive (a) and (b), but instead denies LEM or LNC. The case is therefore the same as in the other version of No-no where either one denies that A and B each have a unique truth value or both are true and false. According to the results of 5.2 and 5.3, we can establish another schema that involves self-negation and captures No-no.

Schema 3

(1) Let ϕ be the concept of truth, i.e. $\phi(x)$ translates 'x is true':
Let X be a sentence. Then $\phi(X) := t(X)$.

(2) A sentence is true iff its truth condition holds:
Let X be a sentence and ψ its truth condition, i.e. X='ψ'[6]. Then ψ iff $t(X)$.

(3) There is a formula $\psi_S(x)$ with 'x' as its only free variable and a set of sentences $S := \{X|\ X = `\psi_S(X)'\}$[7] such that

(i) for all $X \in S$ there is a $F_X \in \mathcal{P}(S)$ such that $F_X \neq \emptyset$ and

X iff for all $Z \in F_X : \neg t(Z)$, and

(ii) either for all $X \in S : t(X)$ or for all $X \in S : \neg t(X)$.

(4) claim: For all $X \in S$: $t(X)$ iff for all $X \in S$: $\neg t(X)$.

Proof. Let for all $X \in S : t(X)$. Then for all $X \in S$: there is a $Z \in F_X$: $t(Z)$. By (3i) and (2), for all $X \in S : \neg t(X)$.
Conversely, let for all $X \in S : \neg t(X)$. Then for all $X \in S$: for all $Z \in F_X$: $\neg t(Z)$. By (3i) and (2), for all $X \in S : t(X)$.

(5) By (4), (3ii) and (CAEC)[8]: For all $X \in S$: $t(X)$ and for all $X \in S$: $\neg t(X)$.

[6] Just as in the Schema of Self-Negation for the Liar, I do not intend to make a statement about truth-conditions but rather call a sentence its truth condition.

[7] Similar to step (4) in the Schema of Self-Negation for the Liar, we assume some form of self-reference, here each X is $circular_1$.

[8] Recall, that CAEC is an argument from 'A iff $\neg A$' to 'A and $\neg A$' applying LEM, 'A or $\neg A$'. Alternatively, one can say that CAEC is an argument from 'A iff B' to 'A and B' with the extra assumption of the claim that A or B. Assume the claim that A or B.

Step 1: The claim A iff B and the claim A together imply the claim A and B.
Clearly, the claim A iff B and the claim A together imply B. Since A follows from A itself, we have A and B.

Schema 3 has profound similarities with the Schema of Self-negation for the Liar, but also profound differences. On the one hand, there is the concept of truth (1), the T-schema (2) and one derives a self-negating statement (4) from which a contradiction follows (5). On the other hand, in (3) one does not define a particular truth condition ψ_S. In the Schema of Self-Negation for the Liar, one defines the truth condition $\psi_S = `\neg t(\psi_S)'$ and deduces with step (2) of the schema that there is a sentence with that truth condition. In Schema 3, one simply assumes that there is some formula $\psi_S(x)$ with '$x$' as its only free variable and a set $S = \{X|\ X = `\psi_S(X)'\}$ that satisfies two conditions. First, all sentences of S are equivalent to saying that there is a nonempty subset of S with only false members. Second, all sentences of S have the same truth value. The self-negating statement is that all sentences of S are true iff they are all false. It is self-negating since by (3ii) 'for all $X \in S$: $t(x)$' and 'for all $X \in S$: $\neg t(x)$' are each other's negations. The contradiction is that all sentences of S are both true and false.

For a set S of sentences to fit in Schema 3, we must have $S = \{X|\ `X = \psi_S(X)'\}$ for some formula $\psi_S(x)$ with 'x' as its only free variable and we must show (3). Whether S satisfies (3) depends on how $\psi_S(x)$ is defined. In No-no, $\psi_S(x)$='For all $Y \in N$: if $Y \neq x$ then Y is false' where N is the set containing both sentences. In 5.3 Step 2, I have shown that N satisfies (3i). In 5.3 Step 1, I have shown that N satisfies (3ii) and hence (4). Therefore N fits Schema 3.

In fact, even the Liar fits Schema 3 if we assume there is a set containing only the Liar sentence, call it S_L. Obviously, all sentences in S_L are of the form $X = `\phi(X)'$ for some $\phi(x)$, here $\phi(x) = `\neg t(x)'$. All sentences in S_L are equivalent to saying that some subset of S_L contains only false sentences, for the Liar says of itself that it is false. And all sentences in S_L have the same truth value, i.e. the Liar has a unique truth value. Therefore, all sentences in S_L are true iff they are false, i.e. the Liar is true iff it is false.

A clear weak spot of Schema 3, however, is that it does not provide criteria for how $\psi_S(x)$ is to be defined in order for a set $S = \{X|\ X = `\psi_S(X)'\}$ to satisfy (3). For one constructs the No-no paradox not by simply assuming that there is a set of sentences with a certain property. Rather, one defines two truth conditions and applies (2) of the Schema of Self-Negation to deduce that there are two sentences with these truth conditions. On this basis one can deduce the existence of the set containing these sentences. However, (3i) indicates that $\psi_S(x)$ must involve '$\neg\phi(x)$', i.e. '$\neg t(x)$'. Otherwise it seems unlikely that it is, as according to (3i), equivalent to saying that there is a subset of S containing only *false* sentences. This suggests that there is a close link between step (3) of

Step 2: The claim A iff B and the claim B together imply the claim A and B.
This is entirely analogous to Step 1.

Step 3: The claim A iff B and the claim A or B together imply the claim A and B.
This follows from Step 1 and Step 2 and the rule of Reasoning by Cases: If assumptions Γ plus A imply C and Γ plus B imply C as well, then Γ plus the claim that A or B imply C.

Step 4: The claim A iff B implies the claim A and B.
This step follows from Step 3 and the claim that A or B which is given.

Therefore we can apply CAEC on (4) together with (3ii).

Schema 3 and step (4) of the Schema of Self-Negation for the Liar. In both cases the truth condition involves '$\neg t(x)$'. Another link is that in both schemas we have a self-referential sentence, a fixed point of $\neg t$ in the liar and a fixed point of ψ_S in No-no. While in the former case it is clear in how far self-reference is linked to self-negation, for the truth condition of the liar to negate itself it has to refer to itself, it remains an open question in how far self-reference is linked to self-negation in No-no. Further, it remains an open question whether the missing criteria for $\psi_S(x)$ can be defined so that the link between Schema 3 and the Schema of Self-Negation for the Liar can be further investigated. In total, Schema 3 needs further refinement but it shows that No-no is another paradox involving self-negation and hints that this in turn involves self-reference and a truth condition in which '$\neg t$' occurs, as in the Liar.

5.4 Conclusion

In this chapter I presented an interpretation of the No-no paradox according to which self-negation is involved in the paradox. The interpretation is motivated by an attempt to give an explanation for the symmetry assumption, the claim that both sentences in No-no have the same truth value, if they each have a unique one. The explanation for this claim is that both sentences are equivalent to one and the same self-referential sentence, 'all other sentences of S are false' where S is the set containing both sentences of No-no. Since both sentences are equivalent to the same self-referential sentence, they are equivalent to each other, i.e. they must have the same truth value. If No-no is stated in terms of a set N containing two types of that self-referential sentence then, we have that the sentences in N are either both true or both false (5.3 Step 1), which makes 'for all $X \in N : t(X)$' and 'for all $X \in N : \neg t(X)$' negations of each other. Further, I showed that both sentences in N are equivalent to stating that there is a subset of N of which all its members are false (5.3 Step 2). Step 1 and Step 2 lead to the self-negating statement that both sentences in N are true iff they are both false.

Based on these results, I then established a schema that involves self-negation and captures No-no, Schema 3. Schema 3 has strong similarities with the Schema of Self-Negation for the Liar and therefore shows the strong connection between the Liar and No-no. However, the schema does not reveal the conditions under which a set like the one in No-no exists. Unlike the Schema of Self-Negation that facilitates solution route (II) to the paradoxes, since it provides a finer analysis of what kind of sentence could be given up to solve the Liar, i.e. sentences with a truth condition defined in terms of '$\neg t$', Schema 3 lacks such an ability. Nevertheless, it suggests that the sentences of a set that satisfy the schema have properties similar to the ones of the Liar sentence, they refer to themselves and involve an occurrence of '$\neg t$'. In the final chapter of this thesis, I will show that this schema also captures Yablo. However, we will also see that there is a crucial difference between Yablo and No-no that is not captured by Schema 3.

Chapter 6

Yablo's Paradox

In the previous chapter I presented a schema involving self-negation that captures the No-no paradox, a paradox not captured by the Inclosure Schema. In this chapter I will show that this schema also captures Yablo's paradox, a paradox whose self-referential status is highly debated. However, we will see that the schema does not capture a distinct property of the paradox. First of all, let us have a look at Yablo's paradox and the discussion about its self-referential status.

6.1 The paradox

In 1993, Yablo presents a paradox that supposedly does not trade on self-reference. To put it in his own words (1993, 1):

> Imagine an infinite sequence of sentences S_1, S_2, S_3,..., each to the effect that every subsequent sentence is untrue:
>
> (S_1) for all $k > 1$, S_k is untrue
>
> (S_2) for all $k > 2$, S_k is untrue
>
> (S_3) for all $k > 3$, S_k is untrue
>
> $\vdots$
>
> Suppose for contradiction that some S_n is true. Given what S_n says, for all $k > n$, S_k is untrue. Therefore (a) S_{n+1} is untrue, and (b) for all $k > n+1$, S_k is untrue. By (b), what S_{n+1} says is in fact the case, whence contrary to (a) S_{n+1} is true! So every sentence S_n in the sequence is untrue. But then the sentences subsequent to any given S_n are all untrue, whence S_n is true after all! I conclude that self-reference is neither necessary nor sufficient for Liar-like paradox.

The contradiction is that there is an n such that $t(S_n)$ and for all k: $\neg t(S_k)$. Since each sentence of the sequence refers only to subsequent sentences, no sentence is self-referential. So it seems we have a paradox without self-reference. The matter, however, is a bit more complicated if one looks at alternative ways to phrase the paradox which will be explored in the next section.

6.2 Is it self-referential?

Priest (1997) was quick in replying to Yablo's paradox and claimed that (i) it not only involves self-reference but (ii) also fits the Inclosure Schema. For (i) Priest presents a formalization of the paradox (1997, 237-238): Let S be the two place satisfaction relation between numbers and predicates. Let $\dot{s}$ be the predicate $\forall k > x, \neg S(k, \dot{s})$, translating 'no number greater than x satisfies this property'. By the Satisfaction Principle, a natural number n satisfies $\dot{s}$ iff $\forall k > n, \neg S(k, \dot{s})$, i.e. for all $n \in \mathbb{N}$:

(*) $S(n, \dot{s})$ iff $\forall k > n, \neg S(k, \dot{s})$.

This is simply saying that each natural number n satsifies $\dot{s}$ iff no greater number than n satisfies $\dot{s}$. This is equivalent to saying that for each natural number n, $t(S_n)$ iff $\forall k > n, \neg t(S_k)$ which is just the T-schema applied to S_n of the sequence in the quote above. However, as Priest notes, we cannot apply the T-schema to formulas with a free variable such as $\forall k > n, \neg t(S_k)$. Now, suppose for contradiction that $S(n, \dot{s})$ for some $n \in \mathbb{N}$:

$$\begin{array}{lll} S(n,\dot{s}) & \Rightarrow \forall k > n, \neg S(k,\dot{s}) & (*) \\ & \Rightarrow \neg S(n+1,\dot{s}) \wedge \forall k > n+1, \neg S(k,\dot{s}) & \\ & \Rightarrow \neg S(n+1,\dot{s}) \wedge S(n+1,\dot{s}) & (*) \end{array}$$

Since $S(n, \dot{s})$ implies a contradiction, we have $\neg S(n, \dot{s})$. Since n was arbitrary we have, for instance, $\neg S(0, \dot{s})$ and, by Universal Generalization, $\forall n, \neg S(n, \dot{s})$. In particular, $\forall k > 0, \neg S(k, \dot{s})$, i.e. $S(0, \dot{s})$. Thus, we have the contradiction that $\neg S(0, \dot{s})$ and $S(0, \dot{s})$.

Clearly, in this version of Yablo we have a fixed point, $\dot{s} =$ '$\forall k > x, \neg S(k, \dot{s})$'. The fixed point exists either by the demonstrative ('no number greater than x satisfies *this property*') or, in an arithmetic setting, by the Gödel-Diagonalization Lemma (generalized to formulas with a free variable) according to which there must be a formula $\dot{s}$, such that $\dot{s} \leftrightarrow \forall n, \neg S(n, \langle \dot{s} \rangle)$. Thus, $\dot{s} =$ '$\forall k > x, \neg S(k, \dot{s})$' is *circular*$_2$.

Priest goes further and shows that Yablo's paradox is captured by the Inclosure Schema (1997, 240-242). Recall, that for a paradox to fit the Inclosure Schema, one has to show that there are two properties ϕ and ψ and a function δ such that:

(1) $\Omega = \{y | \phi(y)\}$ exists and $\psi(\Omega)$ (Existence)

(2) for all $x \subseteq \Omega$ such that $\psi(x)$:

(a) $\delta(x) \notin x$ (Transcendence)

(b) $\delta(x) \in \Omega$ (Closure)

Let $\Omega = \{\langle n, p\rangle |\, S(n, p)\}$ be the set of ordered pairs $\langle n, p\rangle$ where n is a number and p a predicate such that n satisfies p. Let ψ the property of being definable. Clearly, Ω is definable as it has just been defined. Let for all definable $X \subseteq \Omega$, $\delta(X) := \langle 0, r_X\rangle$, where r_X is the predicate

$$\dot{s} \wedge \forall k > 0, \langle k, \dot{s}\rangle \notin \bar{X},$$

where $\bar{X}$ is a name for X and $\dot{s} =$ '$\forall k > x, \neg S(k, \dot{s})$' as above[1]. Now, we prove Transcendence and Closure by reductio. Let $X \subseteq \Omega$ be definable.

Transcendence: $\delta(X) \notin X$

Proof. Suppose $\delta(X) \in X$.

$$\Rightarrow S(0, r_X)$$
$$\Rightarrow \forall k > 0, \neg S(k, \dot{s}) \wedge \forall k > 0, \langle k, \dot{s}\rangle \notin \bar{X}$$

The first conjunct entails $\neg S(1, \dot{s}) \wedge S(1, \dot{s})$ since, as we have already seen, for each n, $\forall k > n, \neg S(k, \dot{s})$ entails $\neg S(n+1, \dot{s}) \wedge S(n+1, \dot{s})$. Thus, we have a contradiction.

Closure: $\delta(X) \in \Omega$

Proof. We need to show:

$$\forall k > 0, \neg S(k, \dot{s}) \wedge \forall k > 0, \langle k, \dot{s}\rangle \notin \bar{X}$$

The first conjunct entails the second since if $\neg S(k, \dot{s})$ then $(k, \dot{s}) \notin \Omega$ and hence $(k, \dot{s}) \notin \bar{X}$. It therefore suffices to show the first conjunct. Suppose $\neg\forall k > 0, \neg S(k, \dot{s})$, then $S(n, \dot{s})$ for some n. Again, we get $\neg S(n+1, \dot{s}) \wedge S(n+1, \dot{s})$ and hence a contradiction.

The contradiction is $\delta(\Omega) \in \Omega$ and $\delta(\Omega) \notin \Omega$. Note that, $\delta(\Omega) \in \Omega$ is $S(0, r_\Omega)$ which is $\forall k > 0, \neg S(k, \dot{s}) \wedge \forall k > 0, \langle k, \dot{s}\rangle \notin \bar{\Omega}$. Since both conjuncts are equivalent, the contradiction can be stated as '$\forall k > 0, \neg S(k, \dot{s})$ and $\neg\forall k > 0, \neg S(k, \dot{s})$' which is just like the contradiction Yablo stated.

Priest ultimately concludes that Yablo does involve self-reference and that it is captured by the Inclosure Schema. 'However one formulates it, it has the characteristic fixed-point structure. Moreover, the paradox is an inclosure contradiction, as are all the other paradoxes of its kind' (1997, 242). However, there are other voices like Sorensen (1998) who claims that self-reference is inessential to not only Yablo but even all paradoxes of the Liar family. '[Priest's] conclusion is based on the premise that Yablo's *specification* of his sequence makes indispensable use of self-reference' (1998, 144). He argues that

[1] Since δ is only defined for definable subsets of Ω, the first conjunct of r_X implies the second. For $\neg S(k, \dot{s})$ implies $\langle k, \dot{s}\rangle \notin \bar{X}$. Priest is fully aware of this fact. The reason for setting up δ this way is to ensure that it 'depends genuinely on X' (1997, 241).

as a finite thinker we cannot list all infinite sentences of the Yablo sequence. We, finite thinkers, therefore require the formula

(S_n) For all $k > n$, S_k is untrue

to describe the Yablo sequence. This formula 'is self-referential in the sense that [(S_n)] uses its own location in the sequence as a reference point to specify which statements are not true i.e. the statements after [(S_n)]' (1998, 144). However, Sorensen argues, an *infinite* being could enumerate the infinite Yablo sequence, i.e. by specifying it demonstratively instead of descriptively. 'Since we finite beings know that the Yablo sequence is paradoxical for the infinite being, we know that the Yablo sequence is paradoxical *simpliciter*'(1998, 145). Moreover, Sorensen points out that using a self-referential description to refer to something does not make the referent self-referential. For instance, the description 'any sentence that has fewer words than this very sentence' is self-referential and refers to descriptions like 'the cat is on the mat' which is not self-referential. 'The architecture of a description does not mold the structure of what it describes'(1998, 148).

Beall (2001) steps in and argues in favor of Priest. He grants Sorensen that an infinite being could enumerate the infinite Yablo sequence. However, we as finite beings can only fix the referent of 'Yablo's sequence' by a description involving self-reference. Therefore, if we ask an infinite being to enumerate the Yablo sequence, the sequence that is to be enumerated is inevitably the referent of our description involving self-reference. Given that the reference of 'Yablo's sequence' is fixed by *our* description, the infinite being would, after all, enumerate a *circular* sequence. He also grants Sorensen that a demonstration of an infinite being to refer to the infinite Yablo sequence *could* be used to fix the reference of 'Yablo's sequence'. Yet, since no one has seen such an infinite sequence, we are stuck with fixing 'Yablo's sequence' to a circular sequence. Concerning Sorensen's last point, Beall counters that it is true but irrelevant. Priest does not argue that the properties of a description are satisfied by the referent of that description. Rather, he argues that any description used to fix the referent of 'Yablo's sequence' is such that its satisfaction conditions require the satisfier to be circular.

The issue is not settled yet. Cook (2006, 2014) shows that there is in fact a way to construct Yablo without circularity, but instead with infinitary means, i.e. where each $S_n =$ '$\neg t(S_{n+1}) \wedge \neg t(S_{n+2}) \wedge \neg t(S_{n+3}) \ldots$' is an infinite conjunction of falsity statements of the subsequent sentences. However, Priest claims that such means are impossible to be applied. For no finite reasoner can deduce from an infinite conjunction. In the end, it is finite reasoning that grounds the conclusion made in Yablo's paradox.

Taking everything into account, Priest has demonstrated that there is a version of Yablo's paradox that involves self-reference, a fixed point, and even fits the Inclosure Schema. Whether or not there are versions of the paradox that do not involve circularity, there is still the problem that the notion of a fixed point needs fixing since, as Leitgeb points out[2], every sentence is a fixed point. In the next section, we will explore a version of the paradox in which each sentence is $circular_1$, i.e. each sentence contains a singular term referring to itself. However, a major objective of this thesis is to provide another structure that is detectable in the paradoxes, including Yablo, i.e. self-negation. In the next section, I will show how Yablo's paradox fits Schema 3.

6.3 Self-Negation in Yablo

To show that Yablo fits Schema 3, consider a formulation of the paradox that is closer to the formulation used by Yablo. First, consider the uniform description of the Yablo sequence: Let $>$ be the usual ordering of the natural numbers and let $S := \{S_i\}$ be a sequence of sentences such that for all $i \in \mathbb{N}$:

S_i:='For all $j > i$: $\neg t(S_j)$'.

Or, equivalently, we can say that for all $i \in \mathbb{N}$:

S_i:='For all $j \in \mathbb{N}$: if $j > i$ then $\neg t(S_j)$'.

Define for all $i, j \in \mathbb{N}$: $S_i >_S S_j$ iff $i > j$. In other words, for all $X, Y \in S$, $X >_S Y$ iff the index of X is larger than the index of Y. Then for all $X \in S$:

X iff for all $Y \in S$: if $Y >_S X$ then $\neg t(Y)$.

Thus, instead of letting each sentence of the Yablo sequence refer to the indexes of the other sentences of the sequence, each sentence can be formulated such that it refers to the sentences of the sequence directly, i.e. instead of saying 'any sentence with a ($>$ -)greater index than mine is untrue' they say 'any sentence that is $>_S$ - greater than *me* is untrue'. Let S be a well-ordered set of sentences such that for all $X \in S$:

X:= 'For all $Y \in S$: if $Y >_S X$ then $\neg t(Y)$'.

In this formulation, the self-referential nature of the sentences becomes evident. Each X in S is $circular_1$ as it contains a singular term referring to itself. Further, unlike the version presented by Yablo, all X in S are sentences (as opposed to formulas with a free variable) which means that we can apply the T-schema. By the T-schema, we have for all $X \in S$:

(1) $t(X)$ iff for all $Y \in S$: if $Y >_S X$ then $\neg t(Y)$

Now, to show that S fits Schema 3 we have to show that there is a formula $\psi_S(x)$ with 'x' as its only free variable such that S is of the form $\{X |\ X = `\psi_S(X)'\}$ and satisfies two conditions:

(i) for all $X \in S$ there is a $F_X \in \mathcal{P}(S)$ such that $F_X \neq \emptyset$ and

X iff for all $Z \in F_X : \neg t(Z)$, and

[2] See Ch.2.2.

(ii) either for all $X \in S : t(X)$ or for all $X \in S : \neg t(X)$.

First, it is easy to see that S is of the desired form with $\psi_S(x)$='for all $Y \in S$: if $Y >_S x$ then $\neg t(Y)$'. Second, we have that for all $X \in S$, 'X' does not appear in $\psi_S(x)$. Therefore, as shown in 5.3 Step 1, all sentences of S are the same self-referential sentence and hence have the same truth value, i.e. either for all $X \in S : t(X)$ or for all $X \in S : \neg t(X)$. Recall, that this makes 'for all $X \in S : t(X)$' and 'for all $X \in S : \neg t(X)$' negations of each other. Third, let for each $X \in S$, $F_X = \{Y \in S | \; X >_S Y\}$ be the set of all successors of X. Clearly, for all $X \in S$: $F_X \in \mathcal{P}(S)$, $F_X \neq \emptyset$ and X iff for all $Z \in F_X : \neg t(Z)$. So, by step (4) of Schema 3, we get

for all $X \in S : t(X)$ iff for all $X \in S : \neg t(X)$.

And by step (5), we get the contradiction that

for all $X \in S : t(X)$ and for all $X \in S : \neg t(X)$.

Thus, Yablo's paradox involves the self-negating statement that all sentences of S are true iff they are all false. The crucial steps are (i) that all sentences in Yablo, just as in No-no, are equivalent to saying that some subset of S contains only false sentences, and (ii) that they are (equivalent to) the same self-referential sentence, just as in No-no, and therefore all sentences have the same truth value.

One might object, that, strictly speaking, it is not Yablo's paradox that is captured. The contradiction captured here is that (i) for all $X \in S : t(X)$ and (ii) for all $X \in S : \neg t(X)$. Recall, that the contradiction in Yablo's original version is that (a) there is an $X \in S : t(S)$ and ($\neg$ a) for all $X \in S : \neg t(X)$. Thus, the characteristic contradiction of Yablo is not captured by Schema 3. However, as shown above, all sentences of Yablo satisfy condition (ii), i.e. they have the same truth value and *either* (i) *or* (ii) is the case. Then, by 5.3 Step 1, (i) is equivalent to (a), i.e. 'there is an $X \in S : t(X)$' is equivalent to 'for all $X \in S : t(X)$'. Therefore, by showing (i) and (ii) we have also shown (a) and ($\neg$ a). The characteristic contradiction of Yablo is thus captured by Schema 3.

However, one can phrase a more serious objection. The contradiction in Yablo does not *depend* on the assumption that all sentences of the sequence have the same truth value. In fact, one needs much less than what Schema 3 requires to derive a contradiction: Let for each $X \in S$, X' be the least greater sentence than X. By Bivalence, we either have that (a) there is a $X \in S$ s.t. $t(X)$ or ($\neg$a) for all $X \in S$: $\neg t(X)$. Suppose (a), i.e. $t(A)$ for some $A \in S$. Then we get a contradiction:

$$\begin{array}{lll} t(A) & \Rightarrow \text{for all } Y \in S : \text{if } Y >_S A \text{ then } \neg t(Y) & (1) \\ & \Rightarrow t(A') \text{ and for all } Y \in S\text{: if } Y >_S A' \text{ then } \neg t(Y) & \\ & \Rightarrow t(A') \text{ and } \neg t(A') & (1) \end{array}$$

However, assuming ($\neg$a), i.e. for all $X \in S$: $\neg t(S)$, trivially implies that for all $X \in S$: for all $Y \in S$ if $Y >_S X$ then $\neg t(Y)$. By (1) again, that is for all $X \in S$: $t(S)$ and thus $t(A)$ after all.

The fact that one does not need to assume that all sentences in Yablo's sequence have the same truth value to derive a contradiction is what distinguishes Yablo from No-no and what is not captured by Schema 3. Nevertheless, the assumption that all sentences have the same truth value is reasonable and shows that there is a strong connection between Yablo and No-no and the other paradoxes of self-negation. All these paradoxes (have versions that) involve self-negation and the two schemas that capture them have strong similarities. Yet, further work is required to fully explore these similarities. Moreover, the fact that (a version of) Yablo fits Schema 3 illustrates that the question of how to define the sentences of a set S so that it fits the schema is more complicated than one might expect. For in Yablo, the truth condition of each sentence X, 'for all $Y \in S$: if $Y >_S X$ then $\neg t(Y)$', reflects the fact that S is a well-ordered set. Further, the fact that each sentence is equivalent to saying that all members of some *non-empty* subset of S are false also depends on the fact that S is infinite. It is no surprise that facts about S are reflected in the definition of ψ_S since 'S' *occurs* in ψ_S.

6.4 Conclusion

Whether or not Yablo's paradox is self-referential is highly debated. On the one hand it is clear, as Priest shows, that there are descriptions of the paradox that involve circularity. On the other hand, it remains open whether there are other descriptions that get by without circularity. In line with the goal of the thesis, I presented a version of Yablo that involves not only circularity but also self-negation by showing that Yablo's paradox fits Schema 3, a schema involving self-negation that also captures No-no. The key to this result is to interpret Yablo's sequence as a well-ordered set of sentences S and that all sentences have the same truth value, i.e. either they are all true or all false. For, just as in No-no, all sentences of the Yablo sequence are (equivalent) to the same self-referential sentence. Further, it is obvious that each sentence in Yablo says of some non-empty subset of the sequence that it is false. These two observations lead to the self-negating statement that all sentences of S are true iff they are all false. Note again, that this is self-negating since *either* all sentences of S are true *or* all are false and therefore 'all sentences of S are true' and 'all sentences of S are false' are each other's negation.

Still, Yablo fits Schema 3 only if the claim that all sentences of the sequence have the same truth value is taken into account. While this may be a reasonable claim, we have seen that the contradiction in Yablo can also be brought about without taking it into account. In fact, going this way towards the contradiction requires much less than Schema 3. Nevertheless, the fact that Yablo fits Schema 3 shows that the paradox *can* be proven via self-negation, just like No-no. The next chapter concludes the thesis by reviewing the main results and exploring what to make of them.

Chapter 7

Conclusion

The goal of this thesis is to provide an analysis of a group of paradoxes of self-reference according to which it is self-negation rather than self-reference that describes what is essential to the paradoxes.

To achieve this goal, I started in Chapter 1 with a short introduction to the paradoxes of self-reference and discussed the wide range of solutions to them by looking at the oldest example of paradoxes, the Liar. The most discussed solutions are the ones that consist in a restriction of classical logic. Here, I have focused on Kripke's highly influential idea to restrict LEM. Next to the challenges that Kripke's theory has to face, we also discussed the dialetheic solution to the paradox which states that there are true contradictions. A prominent defender of dialetheism, Priest, criticizes paracomplete solutions for not applying to all paradoxes. In *Beyond the Limits of Thought*, Priest argues that all paradoxes have the same underlying structure, the Inclosure Schema. He further argues, that a uniform structure demands a uniform solution, in this case, the dialetheic one. The important lesson of Chapter 1 is that depending on what structure one ascribes to the paradoxes of self-reference different solutions seem appropriate.

In Chapter 2, we have looked at the definitions of self-reference that can be found in the literature. According to Hannes Leitgeb, there are two major definitions to consider. A sentences is *circular*$_1$ if it contains a singular term (directly or indirectly) referring to it. A sentence is *circular*$_2$ if it is a fixed point of a syntactical mapping up to arithmetical equivalence. In arithmetical settings, one can show via Gödel's and Tarki's diagonalisation method that such self-reference is possible. The second notion of self-reference is therefore closely linked to the notion of diagonalisation. Both notions of self-reference, as Leitgeb argues, are deficient. A third notion of self-reference is the one of an impredicative definition and is linked to another method of diagonalisation, the one developed by Cantor. According to Priest, it is this notion that describes the self-referential nature of all paradoxes of self-reference.

Chapter 3 handles Priest's account of the paradoxes of self-reference, the Inclosure Schema in particular. The essence of the schema is that in each paradox of self-reference, a function, the diagonaliser, gives an object that is among the members of a totality Ω, that, at the same time, is not in Ω. Most criticisms that the Inclosure Schema has received are to the effect that in case of the Liar the existence of the totality is not necessary to derive a contradiction. I added the criticism that in most paradoxes of self-reference, to assume that Ω exists is to contradict facts. I concluded that in theses cases one cannot use the Inclosure Schema to derive a true contradiction since the existence of Ω is not given. Even worse, in most cases the existence of Ω is indispensable for deriving a contradiction. In these cases, a necessary premise of the paradox is false and therefore the paradox can be considered solved. I further argued that the Inclosure Schema does not capture what is essential to the Liar, Russell and Grelling.

The essentials of the Liar, Russell and Grelling are discussed in Chapter 4. Here, I showed that there is a strong analogy between the three paradoxes. First, in all three cases a somewhat fundamental concept is involved. In the Liar, it is the concept of truth, in Russell the concept of membership (sets) and in Grelling the concept of instantiation (properties). Second, in all three cases one proves a statement of the form A iff $\neg A$ in a *uniform* way from which a contradiction follows almost immediately. I call such an A a *self-negating* sentence. In the Liar, for instance, one can prove that the Liar is true iff it is not true. The truth condition of the Liar therefore negates itself. Its satisfaction consists of its non-satisfaction. In the Schema of Self-Negation I captured the analogy between the three paradoxes. A great advantage of this schema is that at its core is a notion, self-negation, that, given classical logic, always leads to contradiction. The schema thereby captures what is essential to the three paradoxes. Moreover, following a classical solution to the paradoxes, i.e. a solution that retains classical logic, that gives up self-negation would not involve giving up unproblematic cases as it is the case with giving up self-reference. What it means to give up self-negation, however, is yet to be examined. Further, the new schema provides a more fine-grained analysis of the paradoxes as it captures every step towards contradiction. The possible consequences of the new schema are yet to be figured out. It remains open whether it is compatible with the Inclosure Schema or whether it suggests an alternative solution to the paradoxes.

In Chapter 5 and 6, I showed that there is a version of No-no, a paradox not captured by the Inclosure Schema, and a version of Yablo, a paradox with a controversial link to self-reference, that both involve self-negation. I further developed another schema involving self-negation, Schema 3, that captures the two paradoxes and that has strong similarities with the Schema of Self-Negation. In fact, the Liar fits the schema if one assumes that there is a set containing only the Liar sentence. Schema 3 captures the fact that in No-no and Yablo we have a set of sentences that have the same truth value, i.e. either all sentences are true or all are false, and the fact that all sentences are false iff they are all true. Since, by the former, the case that all sentences are true and the case that all sentences are false are each other's negation, the latter is an example of self-negation. However, in case of Yablo one can object that the contradiction can be obtained without considering the fact all sentences have the same truth value and therefore Schema 3 contains an unnecessary assumption.

In total, I provided a schema, the Schema of Self-Negation, that, unlike the Inclosure Schema, captures the fact that in case of the Liar, Russell and Grelling a contradiction can be derived in a highly analogous way via self-negation. While the Inclosure Schema captures many more paradoxes, it is highly controversial whether it supports Priest's dialetheic solution according to which the Existence premise of the Schema is true in all cases. I also provided a similar schema involving self-negation, Schema 3, that captures No-no and Yablo. Besides the benefits and difficulties of the two schemas, I have shown that (versions of) these paradoxes involve a structure, self-negation, that, from a classical point of view, always leads to contradiction. In this sense, the schema capture what is essential to the paradoxes.

These results suggest further study on what role self-negation plays in understanding and solving the paradoxes. Is there a schema that unifies the insights of the Schema of Self-Negation and Schema 3, i.e. is there a schema involving self-negation that captures the Liar, Russell, Grelling, No-no and Yablo? If No-no and Yablo have the same underlying structure as the Liar, are there No-no and Yablo versions of Russell and Grelling? How can a solution to the paradoxes that denies self-negation look? Finally, is self-negation involved in the Knower, Berkeley and Gödel?

Bibliography

Armour-Garb, B. and J. A. Woodbridge (2006). Dialetheism, semantic pathology, and the open pair. *Australasian Journal of Philosophy 84*(3), 395–416.

Badici, E. (2008). The liar paradox and the inclosure schema. *Australian Journal of Philosophy 86*(4), 395–416.

Beall, J. (2001). Is yablo's paradox non-circular? *Analysis 61*(3), 187–193.

Beall, J. (2009). *Spandrels of Truth*. Oxford: Oxford University Press.

Bernays, P. (1935). *On Platonism in Mathematics*. Cambridge: Cambridge University Press, Second edition. reprinted in Benacerraf, Paul and Putnam, Hilary, editors (1983).

Cook, R. T. (2006). There are non-circular paradoxes (but yablo's isn't one of them!). *The Monist 89*, 118–149.

Cook, R. T. (2014). *The Yablo Paradox: An Essay on Circularity*. Oxford: Oxford University Press.

Dümont, J. and F. Mau (1998). Are there true contradictions? a critical discussion of graham priest's "beyond the limits of thought. *Journal for General Philosophy of Science 29*(2), 289–299.

Field, H. (2008). *Saving Truth from Paradox*. Oxford University Press, USA.

Gaifman, H. (2006). Naming and diagonalization, from cantor to gödel to kleene. *Logic Journal of the IGPL 14*(5), 709–728.

Giaquinto, M. (2002). *The Search for Certainty*. Oxford University Press, USA.

Goldstein, L. (2009). A consistent way with paradox. *Philos Stud 144*, 377–389.

Grattan-Guinnes, H. (1998). Structural similarity or structuralism? comments on priest's analysis of the paradoxes of self-reference. *Mind 107*(428), 823–834.

Gödel, K. (1944). *Russell's Mathematical Logic*. Cambridge: Cambridge University Press, Second edition. reprinted in Benacerraf, Paul and Putnam, Hilary, editors (1983).

Hallett, M. (1984). *Cantorian Set Theory and Limitation of Size*. Clarendon Press.

Hughes, G. E. (1982). *John Buridan on self-reference: Chapter eight of Buridan's Sophismata*. Cambridge: Cambridge University Press.

Kripke, S. (1975). Outline of a theory of truth. *The Journal of Philosophy 72*(19), 690–716.

Kroon, F. (1995). Beyond belief? a critical study of graham priest's beyond the limits of thought. *Theoria 67*(2), 140–153.

Leitgeb, H. (2002). What is a self-referential sentence? critical remarks on the alleged (non-)circularity of yablo's paradox. *Logique et Analyse 177-178*(177), 3–14.

Priest, G. (1987). *In Contradiction*. Oxford: Oxford University Press, 2006. Dordrecht: Martinus Nijhoff. 2nd expanded edition.

Priest, G. (1997). Yablo's paradox. *Analysis 57*, 236–242.

Priest, G. (2002). *Beyond the Limits of Thought*. New York: Oxford University Press.

Priest, G. (2010). Hopes fade for saving truth. *Philosophy 85*(1), 109–140.

Russell, B. (1905). On some difficulties in the theory of transfinite numbers and order types. *Proceedings of the London Mathematical Society 2*(4), 29–53. reprinted in (1973).

Shapiro, S. (2003). *Vagueness and Conversation*. in Beall, Edited. *Liars and Heaps*. Oxford, England: Clarendon.

Smullyan, R. M. (1994). *Diagonalization and Self-reference*. Oxford: Clarendon.

Sorensen, R. (1998). Yablo's paradox and kindred infinite liars. *Mind 107*, 137–155.

Sorensen, R. (2001). *Vagueness and contradiction*. Oxford: Oxford University Press.

Strawson, P. F. (1950). On referring. *Mind 59*, 320–344.

Tarski, A. (1935). Der wahrheitsbegriff in den formalisierten sprachen. *Studia Philosophica 1*, 261–405.

Tennant, N. (1998). Beyond the limits of thought. *Philosophical Books 39*, 20–37.

Yablo, S. (1993). Paradox without self-reference. *Analysis 53*, 251–252.