Self-reference and Logic

Thomas Bolander

22nd August 2005

Self-reference is used to denote any situation in which someone or something refers to itself. Object that refer to themselves are called **self-referential**. Any object that we can think of as referring to something—or that has the ability to refer to something—is potentially self-referential. This covers objects such as sentences, thoughts, computer programs, models, pictures, novels, etc.

The perhaps most famous case of self-reference is the one found in the **Liar sentence**:

> "This sentence is not true".

The Liar sentence is self-referential because of the occurrence of the indexical "this sentence" in the sentence. It is also **paradoxical**.[1] That self-reference can lead to paradoxes is the main reason why so much effort has been put into understanding, modelling, and "taming" self-reference. If a theory allows for self-reference in one way or another it is likely to be inconsistent because self-reference allows us to construct paradoxes, i.e. contradictions, within the theory. This applies, as we will see, to theories of sets in mathematics, theories of truth in the philosophy of language, and theories of introspection in artificial intelligence, amongst others.

This essay consists of two parts. The first is called "Self-reference" and the second is called "Logic". In the first part we will try to give an account of the situations in which self-reference is likely to occur. These can be divided into situations involving *reflection*, situations involving *universality*, and situations involving *ungroundedness*.[2] In the second part we will turn to a more formal treatment of self-reference, by formalizing a number of the situations involving self-reference as theories of first-order predicate logic. It is shown that Tarski's schema T plays a central role in each of these formalizations.[3] In particular, we show that each of the classical paradoxes of self-reference can be reduced to

[1] If the sentence is true, what it states must be the case. But it states that it itself is not true. Thus, if it is true, it is not true. On the contrary assumption, if the sentence is not true, then what it states must not be the case and, thus, it is true. Therefore, the sentence is true iff it is not true.

[2] Often cases of self-reference will fit into more than one of these categories.

[3] Tarski's schema T is the set of all first-order logical equivalences

$$T(\ulcorner\varphi\urcorner) \leftrightarrow \varphi$$

where φ is any sentence and $\ulcorner\varphi\urcorner$ is a term denoting φ.

schema T. This leads us to a discussion of schema T, the problems it gives rise to, and how to circumvent these problems.

The first part of the essay does not require any training in mathematical logic.

Part I: Self-Reference

We start out by taking a closer look at paradoxes related to self-reference.

1 Paradoxes

A paradox is a "seemingly sound piece of reasoning based on seemingly true assumptions, that leads to a contradiction (or other obviously false conclusion)" (Audi, 1995). A classical example is **Zeno's Paradox** of Achilles and the Tortoise in which we seem to be able to prove that the tortoise can win any race against the much faster Achilles, if only the tortoise is given an arbitrarily small head-start (cf. (Erickson and Fossa, 1998) for a detailed description of this paradox). Another classical paradox is the **Liar Paradox**, which is the contradiction derived from the Liar sentence. Among the paradoxes we can distinguish those which are related to self-reference. These are called the **paradoxes of self-reference**. The Liar Paradox is one of these, and below we consider a few of the others.

1.1 Grelling's Paradox

A predicate is called **heterological** if it is not true of itself, that is, if it does not itself have the property that it expresses. Thus the predicate "long" is heterological, since it is not itself long (it consists only of four letters), but the predicate "short" is not heterological. The question that leads to the paradox is now:

> Is "heterological" heterological?

It is easy to see that we run into a contradiction independently of whether we answer 'yes' or 'no' to this question.

Grelling's paradox is self-referential, since the definition of the predicate "heterological" refers to *all* predicates, including the predicate "heterological" itself.

1.2 Richard's Paradox

Some phrases of the English language denote real numbers. For example, "the ratio between the circumference and diameter of a circle" denotes the number π. Assume that we have given an enumeration of all such phrases (e.g. by putting them into lexicographical order). Now consider the phrase

> “the real number whose nth decimal place is 1 if the nth decimal place of the nth phrase is 2, otherwise 1”.

This phrase defines a real number, so it must be among the enumerated phrases, say number k in this enumeration. But, at the same time, by definition, it differs from the number denoted by the kth phrase in the kth decimal place.

Richard’s paradox is self-referential, since the defined phrase refers to *all* phrases that define real numbers, including itself.

1.3 Berry’s Paradox

Berry’s Paradox is obtained by considering the phrase

> “the least natural number not specifiable by a phrase containing fewer than 100 symbols”.

The contradiction is that that natural number has just been specified using only 87 symbols!

The paradoxes may seem simply like amusing quibbles. We may think of them as nothing more than this when they are part of our imprecise natural language and not part of theories. When the reasoning and assumptions involved in the paradoxes are not attempted to be made completely explicit and precise, we might expect contradictions to be derivable because of this lack of precision. But having a theory—mathematical, philosophical or otherwise—containing a contradiction is of course devastating for the theory. It shows the entire theory to be inconsistent (unsound). The problem is that it turns out that in many of the intuitively correct theories in which some kind of self-reference is taking place, we can actually reconstruct the above paradoxes, and thereby show these theories to be inconsistent. This applies to the naive theories of truth, sets, and introspection as we will later see.

Before we turn to a more thorough study of the situations in which self-reference is to be expected to occur, we put a bit more structure on our notion of self-reference by introducing *reference relations.*

2 Reference Relations

Reference can be thought of as a relation R between a class of referring objects and a class of objects being referred to. R is called a **reference relation**, and it is characterized by the property that

$$(a, b) \in R \quad \text{iff} \quad b \text{ is referred to by } a.$$

The **domain** of R, that is, the set of a’s for which there is a b with $(a, b) \in R$, is denoted $\mathrm{dom}(R)$. The **range** of R, that is, the set of b’s for which there is an a with $(a, b) \in R$, is denoted $\mathrm{ran}(R)$. The relation R can be depicted as a graph

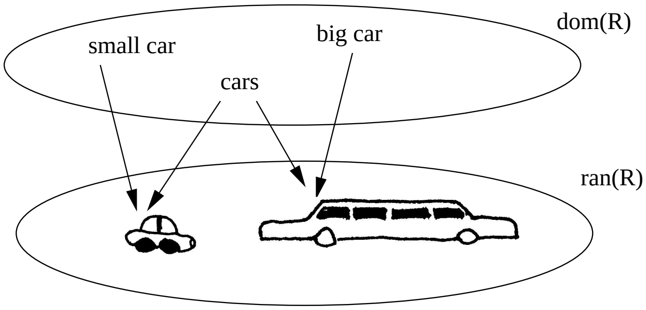


Figure 1: A reference relation.

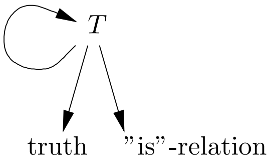
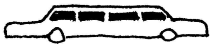


Figure 2: Reference relation for T.

on $\mathrm{dom}(R) \cup \mathrm{ran}(R)$, in which there is an edge from $a \in \mathrm{dom}(R)$ to $b \in \mathrm{ran}(R)$ iff $(a, b) \in R$. If e.g.

$$A = \{\text{“small car”}, \text{“big car”}, \text{“cars”}\}$$

and

$B = \Bigg\{$,  $\Bigg\}$

we could have the reference relation depicted on Figure 1. If $\mathrm{dom}(R) \cap \mathrm{ran}(R) = \emptyset$, as above, a referring object will always be isolated from the object it refers to, since these two objects will be members of two distinct and disjoint classes. Self-reference is thus only possible when $\mathrm{dom}(R) \cap \mathrm{ran}(R) \neq \emptyset$.

Let T be the self-referential sentence

> “This sentence is true”

(T for **truth teller**). The sentence refers to

(i) the sentence itself

(ii) the “is”-relation

(iii) the concept of truth.

Graphically, this could be represented by the reference relation in Figure 2.

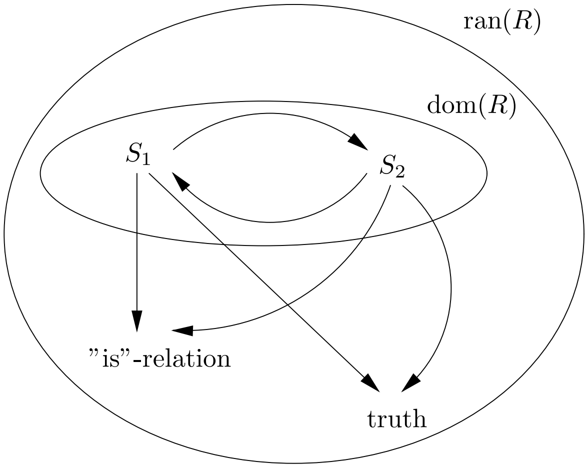


Figure 3: Reference relation for S_1 and S_2.

Notice the loop at T. The loop means that

$$(T, T) \in R,$$

that is, T is referred to by T, which is exactly the condition for T being self-referential. This leads us to the following definition:

> An object $a \in \text{dom}(R)$ is called **directly self-referential** if there is a loop at a in (the graph of) the reference relation.

Now consider the following two sentences, S_1 and S_2,

$$S_1 : \text{The sentence } S_2 \text{ is true.}$$
$$S_2 : \text{The sentence } S_1 \text{ is true.}$$

The reference relation for these two sentences become as depicted in Figure 3. Here the set of referring objects is $\text{dom}(R) = \{S_1, S_2\}$ and the set of objects referred to is

$$\text{ran}(R) = \{S_1, S_2, \text{“is”-relation}, \text{truth}\}\,.$$

Notice that $\text{dom}(R) \cap \text{ran}(R) \neq \emptyset$. None of these sentences are directly self-referential, but S_1 refers to S_2 which in turn refers back to S_1, and vice versa. This gives a *cycle* in the graph consisting of the nodes S_1 and S_2, and the two edges connecting them. We consider both of S_1 and S_2 to be *indirectly self-referential* since each of them refers to itself through the other sentence. Thus we define:

> An object $a \in \text{dom}(R)$ is called **indirectly self-referential** if a is contained in a cycle in (the graph of) the reference relation.

Kripke gives a very nice example of indirect self-reference in (Kripke, 1975). S_1 is the following statement, made by Jones,

$$S_1 : \text{Most of Nixon's assertions about Watergate are false.}$$

and S_2 is the following statement, made by Nixon,

$$S_2 : \text{Everything Jones says about Watergate is true.}$$

The reference relation for this pair of sentences will contain that of Figure 3, i.e. we have again a cycle between S_1 and S_2.

Let us consider a few additional examples of indirect self-reference. In the following, when an object is either directly or indirectly self-referential, we often simply call it **self-referential**.

2.1 Naive Set Theory

In naive set theory (as conceived in the early works of Georg Cantor. See e.g. (Cantor, 1932)) the concept of a set can be defined in the following way:

> By a set we understand any collection of mathematical objects (including sets).

We see that the concept of a set is defined in terms of mathematical objects which can themselves be sets. This means that what we have is a self-referential definition of the concept of a set. This self-reference makes the defined concept inconsistent, as we will see from Cantor's Paradox, introduced in Section 4.

2.2 Dictionary Reference

In a dictionary, the referring objects are the *definienda*, that is, the expressions or words being defined, and the objects referred to are the *definientia*, that is, the expressions or words that define the definienda. In Webster's 1828 dictionary the word "regain" is defined as:

regain : to *recover*, as what has *escaped* or been *lost*.

At the same time, the word "recover" is defined as:

recover : to *regain*; to *get* or *obtain* that which was *lost*.

Using only the words in italic, the reference relation for the above two dictionary definitions become as depicted in Figure 4. Since the definition of "regain" refers to the word "recover" and the definition of "recover" refers to the word "regain", there is a cycle between these two words in the graph. Each is defined through the other in an indirectly self-referential way. This means that unless we know the meaning of one of these words in advance, the dictionary definition will not be able to give us the full meaning of the other word.

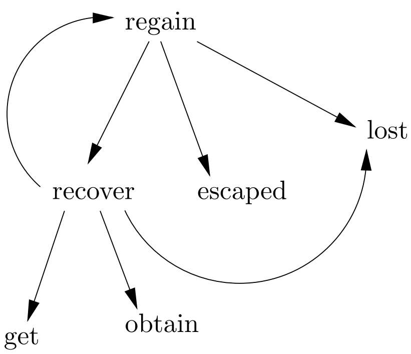


Figure 4: A dictionary reference relation.

This becomes even worse if we consider an English dictionary of the *entire* English language. Since every word is simply defined in terms of other words, we will not from the dictionary be able to learn the meaning of *any* of the words, unless we know the meaning of some of them in advance. This makes a dictionary insufficient as a definition of meaning for a language, as noted by Wittgenstein in the so-called *Blue Book* ((Wittgenstein, 1958)). Wittgenstein's way out was to think of a dictionary as supplied with a set of *ostensive definitions*. An **ostensive definition** of a word is a definition "by pointing out" the referent of the word—e.g. to say the word "banjo" while pointing to a banjo.

Wittgenstein's ideas are related to ideas of groundedness of ungroundedness of reference relations, as we will see in Section 5. But before that we will relate self-reference to reflection and universality.

3 Reflection and Self-Reference

Self-reference is often an epiphenomenon of *reflection* of some kind. The word reflection actually means "bending back". We use reflection to denote situations such as: viewing yourself in a mirror; exercising *introspection* (that is, reflecting on yourself and your own thoughts and feelings); having a theory which is contained in its own subject matter; having a picture which contains a picture of itself (Figure 6). Reflection can also be considered as a name for all the situations in which someone or something views itself "from the outside". In the framework of reference relations, we can choose to define:

> A reference relation R is said to have **reflection** if $\mathrm{dom}(R) \subseteq \mathrm{ran}(R)$.

By this definition, a reference relation has reflection iff every referring object is also an object that is referred to. That is, if R is the reference relation of some

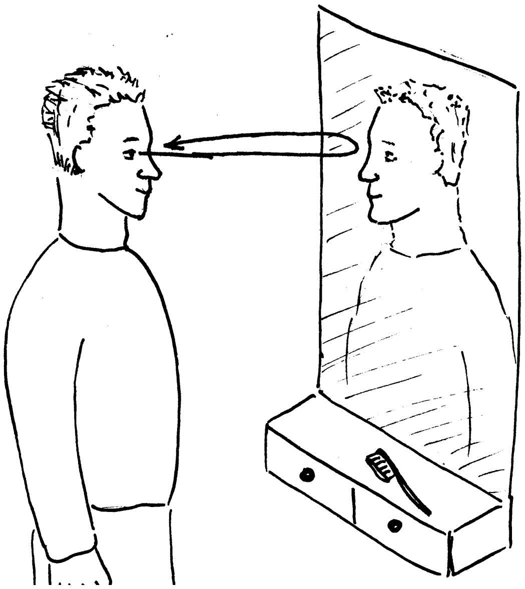


Figure 5: Reflection means “bending back”.

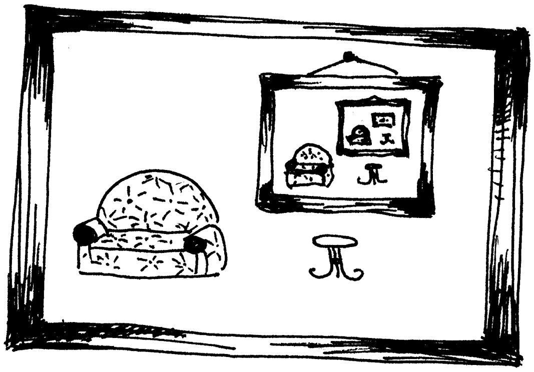


Figure 6: A picture containing itself.

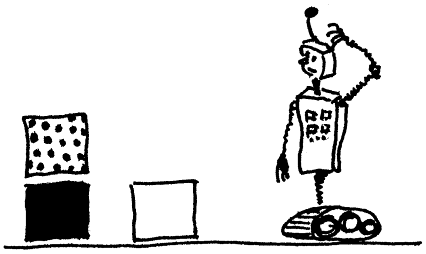


Figure 7: An agent in Blocks World.

theory, then that theory can refer (represent, describe) not only objects of the "external world" but also all the objects of the theory itself.

Reflection does not in itself necessarily lead to self-reference, though self-reference often comes together with reflection. We do only have self-reference if we among the elements of dom(R) can point out an element r which refers to r. Reflection means that every element r of dom(R) is referred to by another element q of dom(R), but for all such pairs (q, r) we might have $q \neq r$. In Section 4 we will show, though, that if reflection is combined with universality, then self-reference cannot be avoided.

Below we will consider some important examples of reflection.

3.1 Artificial Intelligence

A very explicit form of reflection is involved in the construction of artificial intelligence systems such as for instance robots. Such systems are called **agents**. Reflection enters the picture when we want to allow agents to reflect upon themselves and their own thoughts, beliefs, and plans. Agents that have this ability we call **introspective agents**.

An artificial intelligence agent is most often equipped with some formal language which it uses for representing its experiences and beliefs, and which it uses for reasoning about its environment. That is, such an agent has a model of the world it inhabits which is represented by a set of formal sentences.

Consider an agent situated in a *blocks world*[4] as depicted in Figure 7. The agent's task in this world is to move blocks to obtain some goal configuration (e.g. building a tower consisting of all blocks placed in a specific order). The agent's beliefs about this world could be represented in the agent by formal

[4] "Blocks worlds" are the classical example domains used in artificial intelligence.

sentences such as

$$\begin{aligned}&\mathit{on}(\mathit{black\ box}, \mathit{floor})\\ &\mathit{on}(\mathit{dotted\ box}, \mathit{black\ box})\\ &\mathit{on}(\mathit{white\ box}, \mathit{floor})\\ &\mathit{on}(\mathit{agent}, \mathit{floor}).\end{aligned}$$

For the agent to be introspective, though, it should also contain sentences concerning the agent's own beliefs. If the agent believes the sentence

$$\mathit{on}(\mathit{black\ box}, \mathit{floor})$$

to be part of its own model of the world, that could e.g. be represented by the sentence

$$\mathit{agent}(\ulcorner \mathit{on}(\mathit{black\ box}, \mathit{floor})\urcorner).$$

Now, the referring objects in this situation are obviously the sentences that make up the agent's model of the world. So if R denotes the reference relation of the agent, then $\mathrm{dom}(R)$ consists of all these sentences. The object referred to in the case of a sentence like $\mathit{on}(\mathit{black\ box}, \mathit{floor})$ is the black box on the floor, while the object referred to in the case of a sentence like $\mathit{agent}(\ulcorner \mathit{on}(\mathit{black\ box}, \mathit{floor})\urcorner)$ is the sentence $\mathit{on}(\mathit{black\ box}, \mathit{floor})$. If φ is any sentence then $\mathit{agent}(\ulcorner\varphi\urcorner)$ is a sentence referring to φ. This means that the set of objects referred to, $\mathrm{ran}(R)$, contains every sentence, i.e. we have $\mathrm{dom}(R) \subseteq \mathrm{ran}(R)$. By our definition, this means that R has *reflection.* This reflection—that the agent can refer to any of its own referring objects—turns out to provide a major theoretical obstacle to the construction of introspective agents, as we will see in Section 11.

3.2 Philosophy of Language

One of the major problems in the philosophy of language is to give a definition of truth for natural languages. Tarski suggests that every adequate theory of truth should give a predicate "true" satisfying

$$\varphi \text{ is true} \quad \text{iff} \quad \varphi$$

where φ is any sentence. In such a theory of truth we would also have reflection, since the referring objects are sentences, and any sentence φ can be referred to by the sentence

"φ is true".

Reflection is in itself not enough to give self-reference. In both examples above we had reflection but no self-reference, since there were no cycles in the reference relations. The problem is, though, that reflection often comes together with *universality,* and when we have both reflection and universality then self-reference cannot be avoided. Universality is the subject of the following section.

4 Universality and Self-Reference

When we make a statement about all entities in the world, this will necessarily also cover the statement itself. Thus such statements will necessarily be self-referential. We call such statements **universal** (as we call formulas of the form $\forall x \varphi(x)$ in predicate logic). Actually, we will use the term "universal" to denote any statement concerning all entities in the relevant domain of discourse. Correspondingly, in the framework of a reference relation R, we can define:

> An object $a \in \operatorname{dom}(R)$ is called **universal** if $(a, b) \in R$ for all $b \in \operatorname{ran}(R)$.

If R is the reference relation of our natural language then the sentence

$$\text{“All sentences are false”} \tag{1}$$

will be universal. The problem about universality is that reflection and universality together necessarily lead to self-reference, and thereby is likely to give rise to paradoxes. To see that reflection and universality together lead to self-reference, assume R has reflection and that $a \in \operatorname{dom}(R)$ is a universal object. Then we have $(a, b) \in R$ for all $b \in \operatorname{ran}(R)$, and since $\operatorname{dom}(R) \subseteq \operatorname{ran}(R)$ we especially get $(a, a) \in R$. That is, we have the following result:

> Assume R has reflection and that $a \in \operatorname{dom}(R)$ is universal. Then a is self-referential.

Universality enters the picture in the two examples of reflection previously given if we want the agent to be able to express universal statements about its environment or if we want to be able to apply the truth predicate to sentences that concern all sentences of the language (like e.g. the sentence (1)). In such cases self-reference cannot be avoided, and as we will see in the second part of the essay this will allow the paradoxes to surface and produce contradictions in the involved theories.

The problem sketched is not in any way only related to theories of agent introspection and truth. Any theory that is part of its own subject matter has reflection. Thus, if these theories make use of universal statements as well, then these theories contain self-referential statements, and then the paradoxes of self-reference will not be far away. Thus, self-reference is a problem to be taken seriously by any theory that is part of its own subject matter. This applies to theories of cognitive science, psychology, semiotics, mathematics, sociology, system science, cybernetics, computer science.

Note, that each of the paradoxes of self-reference considered in Section 1 involves both reflection and universality, since they all refer to the totality of objects of their own type: the predicate "heterological" refers to all predicates; the phrase defining a real number in Richard's paradox refers to all phrases defining real numbers; the phrase specifying a natural number in Berry's paradox refers to all phrases specifying natural numbers.

Let us conclude this section by considering another example of a universal object in a reflective setting. In the naive theory of sets (cf. Section 2.1) we can consider the set U of all sets. U is certainly a universal object, since it refers to all other sets.[5] At the same time, the theory of sets is reflective since for the reference relation R of sets, $\mathrm{dom}(R)$ and $\mathrm{ran}(R)$ are both the class of all sets. Thus U is a self-referential object, and this leads to trouble. Cantor have proved that the cardinality[6] of any set is smaller than the cardinality of the set of subsets of this set. This result is called **Cantor's Theorem**.[7] Let us see what happens if we apply Cantor's Theorem to the set U. First of all, we note that the set of all subsets of U is U itself, since U contains all sets. But then, by Cantor's Theorem, the cardinality of U is smaller than the cardinality of U, which is a contradiction. This contradiction is known as **Cantor's Paradox**. Cantor's Paradox proves that the naive theory of sets is inconsistent.

5 Ungroundedness and Self-Reference

Self-reference often occurs in situations that have an *ungrounded* nature. Given a reference relation R, we can define ungroundedness in the following way:

> An object $a \in \mathrm{dom}(R)$ is called **ungrounded** if there is an infinite path starting at a in the graph of the reference relation R. Otherwise a is called **grounded**.

Note, that if $\mathrm{dom}(R) \cap \mathrm{ran}(R) = \emptyset$, that is, if referring objects and objects being referred to are completely separated, then all elements are grounded.

If we take the dictionary example of Section 2.2, we can give a simple example of ungroundedness. Let R be the reference relation of Webster's 1828 dictionary, that is, let R contain all pairs (a, b) for which b is a word occurring in the definition of a. Since every word of the dictionary refers to at least one other word, every word will be the starting word of an infinite path of R. Here is a finite segment of one of these paths, taken from the 1828 dictionary:

$$\text{regain} \rightarrow \text{recover} \rightarrow \text{lost} \rightarrow \text{mislaid} \rightarrow \text{laid} \rightarrow \\ \text{position} \rightarrow \text{placed} \rightarrow \text{fixed} \rightarrow \dots$$

Now the problem that Wittgenstein considered can be stated in the following simple manner: in a dictionary all words are ungrounded. Since there are only finitely many words in the English language, any infinite path of words will contain repetitions. If a word occurs at least twice on the same path, it will

[5] It is natural to think of the objects being referred to by a set as the elements of the set.

[6] The cardinality of a set is a measure of its size.

[7] It is interesting at this point to note that the argument leading to Cantor's Theorem—a so-called *diagonal argument* (which he was the first to use)—has basically the same structure as Richard's Paradox.

be contained in a cycle. Thus, in *any* dictionary of the entire English language there will necessarily be words defined indirectly in terms of themselves. That is, any such dictionary will contain (indirect) self-reference.

Ungroundedness does not always lead to self-reference, but self-reference is very often a byproduct of ungroundedness. So whenever one encounters ungroundedness, one should be very careful to ensure that this ungroundedness does not lead to self-reference and paradoxes.

Actually, as showed by Steven Yablo in (Yablo, 1993), ungroundedness can lead to paradoxes even in cases where we do not have self-reference. **Yablo's Paradox** is obtained by considering an infinite sequence of sentences $S_1, S_2, \ldots$ defined by:

$$\begin{aligned} &S_1 : \text{All sentences } S_i \text{ with } i > 1 \text{ are false.} \\ &S_2 : \text{All sentences } S_i \text{ with } i > 2 \text{ are false.} \\ &S_3 : \text{All sentences } S_i \text{ with } i > 3 \text{ are false.} \\ &\vdots \end{aligned}$$

The reference relation for these sentences looks like this:

$$\begin{tikzcd}
S_1 \arrow[r] \arrow[rr, bend left] \arrow[rrr, bend left] \arrow[rrrr, bend left] & S_2 \arrow[r] \arrow[rr, bend right] \arrow[rrr, bend right] & S_3 \arrow[r] \arrow[rr, bend right] & S_4 \arrow[r] & S_5 \arrow[r] & \cdots
\end{tikzcd}$$

As one sees, there is no self-reference involved, but we still get a paradox: Assume S_i is true for some i. Then all S_j for $j > i$ must be false. In particular, S_{i+1} must be false. But since S_j is false for all $j > i+1$, S_{i+1} must also be true. This is a contradiction. Therefore all S_i must be false. But then S_1 should be true, which is again a contradiction.

It should be noted that even though ungroundedness does not always lead to self-reference, self-reference always leads to ungroundedness: any self-referential object a is contained in a cycle, and we get an infinite path from a by passing through this cycle repeatedly.

6 Vicious and Innocuous Self-Reference

Not all self-reference leads to paradoxes. There is no paradox involved in a self-referential sentence like

$$\text{“This sentence is true”.} \tag{2}$$

We can assume either that the sentence is true or that it is false, and neither of the cases will lead into contradiction. But as soon as we introduce a “not” in the sentence, that is, consider the following sentence instead

$$\text{“This sentence is \textit{not} true”} \tag{3}$$

we get a paradox (the Liar Paradox). Self-reference that leads to paradoxes we call **vicious self-reference** and self-reference that does not we call **innocuous self-reference**. It can be shown that self-reference can only be vicious if it involves negation or something equivalent (as the "not" in (3)). This means, for instance, that none of the paradoxes of self-reference considered above could be carried through if the occurrence of negation in their central definitions where removed (e.g. if we removed the "not" in the definition of heterological of Grelling's Paradox).

Part II: Logic

We now turn to a more formal treatment of self-reference, by formalizing some of the situations considered in the first part of the essay as theories of first-order predicate logic (henceforth simply called *first-order theories*). We will assume that all considered first-order theories contain the standard **numerals**:[8]

$$\bar{0}, \bar{1}, \bar{2}, \bar{3}, \ldots$$

We use $\ulcorner \cdot \urcorner$ to range over coding schemes. By a **coding scheme** we understand any injective mapping from sentences into numerals. That is, if φ is a sentence then $\ulcorner \varphi \urcorner$ is the numeral $\bar{n}$ for some natural number n. $\ulcorner \varphi \urcorner$ is a *name* for φ; we call it the **code number** of φ. If $\psi(x)$ is a formula containing x as its only free variable then $\psi(\ulcorner \varphi \urcorner)$ is a sentence expressing that "φ has the property expressed by ψ". In this sense, $\psi(\ulcorner \varphi \urcorner)$ *refers* to φ.

Schema T is, as before, defined as the theory containing each of the equivalences

$$T(\ulcorner \varphi \urcorner) \leftrightarrow \varphi$$

where T is a fixed one-place predicate symbol and φ is any sentence.

The aim of this part of the essay is to show that schema T is taking a central position in almost all situations in which we have self-reference. Indeed, schema T can be thought of as a *unifying principle* of all the different occurrences of self-reference.

7 Formalizing Paradoxes

We now try to formalize some of the most famous paradoxes of self-reference to show how these involve schema T. As mentioned, a paradox is a "seemingly sound *piece of reasoning* based on seemingly true *assumptions* that lead to a *contradiction*". Formalizing a paradox means to reconstruct it inside a formal theory (in our case, a first-order theory). This involves finding formal counterparts to each of the elements involved in the informal paradox. The formal counterpart of a "piece of reasoning" is a formal proof and the formal counterpart of an "assumption" is an axiom. Thus the formal counterpart of a piece of reasoning leading to a contradiction will be a formal proof of the inconsistency of the theory in question. Thus:

[8] Actually, any infinite collection of closed terms would do.

A **formalization** of a paradox is a formal proof of the inconsistency of the theory in which the axioms are the formal counterparts of the assumptions of the paradox.

7.1 The Liar Paradox

As already mentioned, the Liar Paradox is the contradiction that emerges from trying to determine whether the Liar sentence

> "This sentence is false"

is true or false. We will now try to formalize this paradox.

In general, sentences that are directly self-referential can be put in the following form:

$$\text{“This sentence has property } P\text{”.} \tag{4}$$

The assumption that characterizes such a sentence is that the term "this sentence" refers to the sentence itself. Another way of stating this assumption is to say that (4) should satisfy the following equivalence

$$\begin{aligned} &\text{This sentence has property } P \\ \Leftrightarrow\ &\text{“This sentence has property } P\text{” has property } P, \end{aligned} \tag{5}$$

that is, replacing the term "this sentence" by the sentence itself will not change the meaning of the sentence. Formally, this assumption can be expressed as the axiom

$$P(t) \leftrightarrow P(\ulcorner P(t) \urcorner) \tag{6}$$

where t is a term having the intended interpretation: "this sentence". This equivalence corresponds to the equivalence (5), in that "this sentence" have been replaced by t and the quotes "$\cdot$" have been replaced by $\ulcorner \cdot \urcorner$.

The Liar Paradox also rests on the assumption that our language has a truth predicate. The formal counterpart of this assumption is that our theory includes schema T. In the Liar sentence, P is the property "not true". Let therefore P in (6) denote the formula $\neg T(x)$. Then, in the theory consisting of schema T and (6), we get the following proof:

1.	$\neg T(t) \leftrightarrow \neg T\left(\ulcorner \neg T(t) \urcorner\right)$	(6) with P being $\neg T$
2.	$T\left(\ulcorner \neg T(t) \urcorner\right) \leftrightarrow \neg T(t)$	instance of schema T
3.	$T\left(\ulcorner \neg T(t) \urcorner\right) \leftrightarrow \neg T\left(\ulcorner \neg T(t) \urcorner\right)$	by 1. and 2.

This proves the theory consisting of (6) and schema T to be inconsistent, which is our formalization of the Liar Paradox.

7.2 Grelling's Paradox

We will now formalize Grelling's paradox. Recall that Grelling's Paradox is the paradox that emerges when trying to answer whether "heterological" is heterological. The formal counterpart of a predicate is a formula. A formula $\varphi(x)$ is then *heterological* if it is "not true of itself", that is, if

$$\neg T\,(\ulcorner \varphi(\ulcorner \varphi \urcorner) \urcorner)$$

holds, where T is a truth predicate. So to formalize Grelling's paradox we again need to have schema T among our axioms. We also need axioms that allow us to apply a formula to itself (that is, the code of itself). To obtain this, we introduce a function symbol *app* and axioms

$$\mathit{app}\,(\ulcorner \varphi(x_1) \urcorner, \tau) = \ulcorner \varphi(\tau) \urcorner \tag{7}$$

for all formulas φ and all terms τ. These axioms ensure us that $\mathit{app}(\ulcorner \varphi(x_1) \urcorner, \tau)$ denotes the result of "applying" $\varphi(x_1)$ to τ (that is, instantiating $\varphi(x_1)$ with τ). Now we can formalize the predicate "heterological" as the formula $\mathit{het}(x_1)$ given by

$$\mathit{het}(x_1) =_{\mathit{df}} \neg T\,(\mathit{app}(x_1, x_1))\,.$$

To obtain the contradiction we should ask whether $\mathit{het}(\ulcorner \mathit{het}(x_1) \urcorner)$ holds or not. We get the following proof:

$$\begin{array}{lll}
1. & \mathit{het}(\ulcorner \mathit{het}(x_1) \urcorner) \leftrightarrow \neg T\,(\mathit{app}\,(\ulcorner \mathit{het}(x_1) \urcorner, \ulcorner \mathit{het}(x_1) \urcorner)) & \text{by def. of } \mathit{het}(x_1) \\
2. & \mathit{het}(\ulcorner \mathit{het}(x_1) \urcorner) \leftrightarrow \neg T\,(\ulcorner \mathit{het}(\ulcorner \mathit{het}(x_1) \urcorner) \urcorner) & \text{by 1. and (7)} \\
3. & \mathit{het}(\ulcorner \mathit{het}(x_1) \urcorner) \leftrightarrow T\,(\ulcorner \mathit{het}(\ulcorner \mathit{het}(x_1) \urcorner) \urcorner) & \text{instance of schema T} \\
4. & \neg T\,(\ulcorner \mathit{het}(\ulcorner \mathit{het}(x_1) \urcorner) \urcorner) \leftrightarrow T\,(\ulcorner \mathit{het}(\ulcorner \mathit{het}(x_1) \urcorner) \urcorner) & \text{by 2. and 3.}
\end{array}$$

This proves the theory consisting of (7) and schema T to be inconsistent, which is our formalization of Grelling's paradox.

Richard's Paradox is formalized in much the same way as Grelling's Paradox, though the formalization becomes slightly more technical. For these reasons we choose to leave out a formalization of Richard's Paradox in this essay.

7.3 Berry's Paradox

Obviously, to formalize Berry's Paradox, we need axioms formalizing a reasonable part of arithmetic. Apart from this we only need a formal counterpart of the notion of specifiability (the formal counterpart of a "phrase" naturally being a formula). We can use the same trick as we did in the previous examples. A formula $\varphi(x)$ specifies the number n iff $\varphi(m)$ holds exactly when $m = n$. If we want to define a formula $\mathit{spec}(x, y)$ such that $\mathit{spec}(\ulcorner \varphi(x) \urcorner, n)$ holds precisely when $\varphi(x)$ specifies n, then it should look like

$$\mathit{spec}(x, y) =_{\mathit{df}} \forall z\,(z = y \leftrightarrow T(\mathit{app}(x, z)))$$

where T and *app* are defined as before. We will not go further into the details of formalizing this paradox, but refer to (Boolos, 1989) in which this is carried out. We just note that again schema T is central to the formalization. The notion of specifiability could not have been formalized without schema T or something equivalent.

We have now shown how to formalize several of the most famous paradoxes of self-reference, and, as we have seen, these formalized paradoxes all turn out to be reducible to schema T. That is, all these paradoxes have a common core which is schema T. What we can conclude is that:

(i) That schema T can be extracted from all these paradoxes helps us see the close formal relationship between the paradoxes of self-reference.

(ii) That all these paradoxes can be extracted from schema T helps us to see the importance of schema T in understanding the paradoxes of self-reference, and in understanding self-reference in general.

Below we consider some examples of occurrences of schema T in the philosophy of language, mathematics, and artificial intelligence.

8 The Naive Theory of Truth

As mentioned, Tarski thought of his schema T as describing the principle that any theory of truth should satisfy. The first-order theory consisting only of schema T is consistent. But for schema T to be a sensible principle of truth we must expect it to be consistent also when added to any consistent, "realistic" first-order theory. It should be a principle of truth working no matter which domain of discourse we would like to apply truth to. But, unfortunately, because of self-reference it is not so. In the formalizations of the paradoxes above we have seen several examples showing that schema T becomes inconsistent when added to even quite weak and harmless axioms (at least harmless when these axioms are taken by themselves or together with standard theories for arithmetic, set theory, or the like). In fact, it can easily be shown that all of the axioms assumed above in addition to schema T are *interpretable* in Peano Arithmetic, that is, they can all be translated into equivalent axioms of Peano Arithmetic.[9] This gives us the famous **Tarski's Theorem**:

Peano Arithmetic extended with schema T is inconsistent.

Note the interesting fact that *any* of the above paradoxes can be used to prove Tarski's Theorem—one just needs to show that the axioms of the formalized paradox are interpretable in PA (Peano Arithmetic). This shows that the contradiction derivable from the formalized paradox can be carried through in PA + schema T.

[9] For a precise definition of "interpretable in" we refer to (Mendelson, 1997) or a similar introduction to mathematical logic. At this point it is enough to note that when an axiom A is interpretable in a theory K it means that any proof in $K + A$ can be translated into a corresponding proof in K. It should be noted that to prove the interpretability we need to choose our coding scheme $\ulcorner \cdot \urcorner$ with care.

That schema T becomes inconsistent when standard arithmetic is added is a very serious drawback for the theory of truth expressed through schema T. It gives rise to an important problem of how we can restrict schema T to regain the essential consistency. This is the question that we take up in Section 12.

But let us first consider some more examples of situations in which schema T turns up, which makes the reasons to find consistent ways to restrict schema T even more urgent.

9 Gödel's Incompleteness Theorem

We now show how schema T is related to Gödel's famous First Incompleteness Theorem.

A version of the Incompleteness Theorem states that

> If PA is ω-consistent[10] then it is incomplete.[11]

To prove this, we can show that the assumption that PA is both ω-consistent and complete leads to a contradiction. On the basis of the formalizations of paradoxes that we have been considering, we see that this could be proved by showing that if PA were both ω-consistent and complete then some paradox would be formalizable in PA. This was, roughly, Gödel's idea.[12] He constructed a formula Bew (for "Beweis") in his theory satisfying, for all φ and all n,

$$\vdash Bew(\bar{n}, \ulcorner\varphi\urcorner) \quad \Leftrightarrow \quad n \text{ denotes a proof of } \varphi. \tag{8}$$

Assuming the theory to be ω-consistent and complete we can prove that

$$\vdash \exists x Bew(x, \ulcorner\varphi\urcorner) \quad \Leftrightarrow \quad \vdash \varphi$$

for every sentence φ. The proof runs like this: First we prove the implication from left to right. If $\vdash \exists x Bew(x, \ulcorner\varphi\urcorner)$ then there is some n such that $\nvdash \neg Bew(\bar{n}, \ulcorner\varphi\urcorner)$, by ω-consistency. By completeness we get $\vdash Bew(\bar{n}, \ulcorner\varphi\urcorner)$ for this n. By (8) above we get that n denotes a proof of φ. That is, φ is provable, so we have $\vdash \varphi$. To prove the implication from right to left, note that if $\vdash \varphi$ then there must be an n such that $\vdash Bew(\bar{n}, \ulcorner\varphi\urcorner)$, by (8). From this we get $\vdash \exists x Bew(x, \ulcorner\varphi\urcorner)$, as required. This concludes the proof.

Now, when we have

$$\vdash \exists x Bew(x, \ulcorner\varphi\urcorner) \quad \Leftrightarrow \quad \vdash \varphi$$

in a complete theory, we must also have

$$\vdash \exists x Bew(x, \ulcorner\varphi\urcorner) \leftrightarrow \varphi.$$

[10] A theory is called **ω-consistent** if, for every formula $\varphi(x)$ containing x as its only free variable, if $\vdash \neg\varphi(\bar{n})$ for every natural number number n, then it is not the case that $\vdash \exists x \varphi(x)$.

[11] A theory is incomplete if it contains a formula which can neither be proved nor disproved.

[12] Though he considered a different formal theory, P.

If we let the formula $\exists x Bew(x, \ulcorner\varphi\urcorner)$ be abbreviated by $T(\ulcorner\varphi\urcorner)$ then these equivalences read

$$\vdash T(\ulcorner\varphi\urcorner) \leftrightarrow \varphi$$

which is schema T!

That is, if we assume PA (or a related theory) to be ω-consistent and complete then schema T turns out to be interpretable in it. Now, Tarski's Theorem shows that there exists no such consistent theory. This gives us a proof of Gödel's Incompleteness Theorem. Furthermore, in the same way that one could use any of the paradoxes of self-reference to prove Tarski's Theorem, one can use ones favorite paradox of self-reference to prove Gödel's Theorem.

To summarize the process: first you assume your theory to be both ω-consistent and complete. Then you show that this makes schema T interpretable in the theory. Having schema T means that you can choose any paradox of self-reference and formalize it in the theory. The formalized paradox produces a contradiction in the theory, and thus shows that the theory cannot be both ω-consistent and complete.

Gödel himself actually had a footnote in his 1931 article, in which he proved the Incompleteness Theorem ((Gödel, 1931)), saying that any paradox of self-reference[13] could be used to prove the Incompleteness Theorem.

The reason that we have a result such as Gödel's Incompleteness Theorem is closely related to *reflection*. What Gödel ingeniously discovered was that formal theories can be reflected inside themselves, since numerals can be used to refer to formulas through the use of a coding scheme, $\ulcorner\cdot\urcorner$, and by means of these codes provability can be restated inside the theories as arithmetical properties.

10 Axiomatic Set Theory

Schema T also plays a central role in axiomatic set theory. By the **full abstraction principle** we understand the set of formulas on the form

$$\forall x \, (x \in \{y \mid \varphi(y)\} \leftrightarrow \varphi(x))\,^{14}$$

where φ is any formula. When Gottlob Frege tried to give a foundation for mathematics (set theory) through his works "Die Grundlagen der Arithmetik (1884)" and "Grundgesetze der Arithmetik (1893,1903)", the full abstraction principle were among his axioms. But in 1902 his system was shown to be inconsistent by Bertrand Russell. Russell constructed a paradox of self-reference which was formalizable within Frege's system. **Russell's Paradox** runs like this:

> Let M be the set of all sets that are not members of themselves. Is M a member of itself or not?

[13] He used the term "epistemic" about these paradoxes.

[14] The formula can be read: "for all sets x, x is in the set of y's for which $\varphi(y)$ holds if and and only if $\varphi(x)$ holds".

From each answer to this question the opposite follows. Notice the similarity between this paradox and Grelling's paradox considered in Section 1.1. **Russell's Paradox** can be formalized in any system containing the full abstraction principle. We let $M = \{y \mid y \notin y\}$, that is, $M = \{y \mid \varphi(y)\}$ where $\varphi(y) = y \notin y$. The abstraction principle instantiated by the formula φ now becomes

$$\forall x \, (x \in \{y \mid y \notin y\} \leftrightarrow x \notin x) \, .$$

Letting $x = \{y \mid y \notin y\}$, we get

$$\{y \mid y \notin y\} \in \{y \mid y \notin y\} \leftrightarrow \{y \mid y \notin y\} \notin \{y \mid y \notin y\}$$

which is a contradiction. Thus Frege's system, or indeed any system containing the full abstraction principle, is inconsistent.

The discovery of this inconsistency lead to extensive research in how the full abstraction principle could be restricted to regain consistency.

Actually, as we will now show, every instance of schema T can be interpreted in the corresponding instance of the abstraction principle. This means that if we can prove that a set of instances of schema T is inconsistent, then we have also proven that the corresponding set of instances of the abstraction principle is inconsistent. In other words, proving consistency results about restricted versions of schema T will also give corresponding consistency results about restricted versions of the abstraction principle.

The result is the following:

Every instance of schema T:

$$T(\ulcorner \varphi \urcorner) \leftrightarrow \varphi$$

can be interpreted in the corresponding instance of the abstraction principle:

$$\forall x \, (x \in \{y \mid \varphi\} \leftrightarrow \varphi) \, .$$

The proof is quite simple. If we have got a theory containing

$$\forall x \, (x \in \{y \mid \varphi\} \leftrightarrow \varphi)$$

then $T(\ulcorner \varphi \urcorner)$ can be interpreted in it by the following *extension by definitions*:[15]

$$\begin{gathered} \ulcorner \varphi \urcorner = \{y \mid \varphi\} \\ T(x) \leftrightarrow \bar{0} \in x. \end{gathered}$$

Since we have

$$\forall x \, (x \in \{y \mid \varphi\} \leftrightarrow \varphi)$$

[15] We refer again to (Mendelson, 1997) for a definition of the concept of "extension by definitions".

we get in particular

$$(\bar{0} \in \{y \mid \varphi\} \leftrightarrow \varphi)$$

which is the same as

$$T(\ulcorner \varphi \urcorner) \leftrightarrow \varphi,$$

using the definitions of $\ulcorner \cdot \urcorner$ and T. This proves $T(\ulcorner \varphi \urcorner) \leftrightarrow \varphi$ to be interpretable in $\forall x\, (x \in \{y \mid \varphi\} \leftrightarrow \varphi)$.

11 Agent Introspection

We now turn to our last example of an occurrence of schema T in a situation dealing with self-reference. We consider again the problem of constructing introspective agents, as introduced in Section 3.1. Since the agent's model of the world is supposed to consist of a set of sentences, we can think of this model as being a formal theory K. This could be a theory in any kind of formal language, but at this point we will assume that it is a theory in a first-order language. Then, for the agent to believe that e.g. the black box is on the floor would correspond to having

$$K \vdash \mathit{on}(\mathit{black\ box}, \mathit{floor}). \tag{9}$$

If the agent has introspection, it also has beliefs about its own model of the world. If it believes the sentence in (9) to be contained in its own model of the world we would have

$$K \vdash \mathit{agent}\left(\ulcorner \mathit{on}(\mathit{black\ box}, \mathit{floor}) \urcorner\right).$$

Now, if we assume that all of the agent's beliefs about itself to be correct, we should have

$$K \vdash \mathit{agent}(\ulcorner \varphi \urcorner) \quad \Leftrightarrow \quad K \vdash \varphi$$

for all sentences φ. Of course, not all of an agent's beliefs about itself will necessarily always be correct. But even so, the agent might believe this to be the case; and that would correspond to having

$$K \vdash \mathit{agent}(\ulcorner \varphi \urcorner) \leftrightarrow \varphi$$

for all sentences φ. Using "T" instead of "*agent*" this gives us, once again, schema T!

That is, if an agent has introspection and believes this introspection to be correct, then it will necessarily contain schema T in its model of the world. As we know from Tarski's Theorem and our formalized paradoxes this is very difficult to obtain without running into contradictions. At least, it is extremely sensitive to what other axioms we have in K. This is a major drawback in the design of introspective agents.

We have to expect that any kind of axioms could be in K, depending on the environment of the agent and its beliefs about it. The set of axioms of K could even change over time due to changes in the environment. If K includes schema

T it means that the agent could suddenly become inconsistent as a consequence of changes in the external world. This seems to prove that it is not possible for an introspective agent consistently to obtain and retain the belief that its introspection is correct.

This conclusion appears very counterintuitive, but again it has to do with the paradoxes of self-reference. If the agent has introspection, and believes this introspection to be correct, it can construct paradoxes of self-reference concerning its own beliefs, and these paradoxes make the agent inconsistent.

The problem is now to find ways to treat agent introspection such that this introspection will not lead into inconsistency. It seems that we have two possibilities: either to ensure that the agent will not be able to make self-referential statements (which would be a restriction on its introspective abilities) or to restrict its logical abilities such that self-referential statements could be assumed consistently. Such restrictions are the subject of the following section.

12 Taming Self-Reference

We have now seen that schema T occurs as the natural principle in a large number of situations of very different kinds. Schema T is the underlying principle in the naive theories of truth, sets, and agent introspection. But unfortunately, schema T is also the underlying principle in the paradoxes of self-reference, which means that most of the theories we are interested in become inconsistent when schema T is added. Since the inconsistency of schema T is a consequence of the presence of self-referential sentences, there seems to be two possible ways to get rid of the problem: ban self-referential sentences in our language or weaken the underlying logic so that these sentences will do no harm. That is, the different ways to restrict schema T in order to ensure consistency seems to divide into the following two major categories:

(i) Cutting away the problematic part (i.e. getting rid of the viciously self-referential sentences).

(ii) Making the problematic part unproblematic (i.e. ensure that self-reference does not lead to disaster).

12.1 Cutting Away the Problematic Part

Cutting away the problematic part means to restrict the set of instances of schema T such that the viciously self-referential sentences are excluded from entering the schema. By the **T-scheme** over M, where M is a set of sentences, we understand the following set of equivalences:

$$T(\ulcorner \varphi \urcorner) \leftrightarrow \varphi, \text{ for all } \varphi \in M.$$

If M does not contain sentences that are viciously self-referential, it can be proven that the T-scheme over M can consistently be added to *any* consistent theory.[16] This is because banning the viciously self-referential sentences from schema T makes it impossible to reconstruct the paradoxes of self-reference within the theory.

One very coarse way of disallowing self-reference was proposed by Tarski himself ((Tarski, 1956)): M should not be allowed to contain any sentence in which the predicate symbol T occurs. Note that this will ensure that none of the proofs of the formalized paradoxes considered in Section 7 can be carried through. This restriction is sufficient to reestablish consistency, but it is at the expense of a substantial loss of the expressive power of schema T. It means that iterated truth like in

> "It is true that it is not true that n is a prime number"

that formally looks like this

$$T\left(\ulcorner \neg T(prime(\bar{n})) \urcorner\right)$$

will not be treated correctly by the restricted T-scheme.

Several less coarse solutions have been proposed in the literature since Tarski. First of all, one notes that not all self-reference is vicious, so we can allow self-referential sentences in M as long as they are not vicious. As mentioned, for self-reference to be vicious, it needs to involve negation. A sentence in which the predicate symbol T is not in the scope of negation ($\neg$) is called a **positive sentence**. Positive sentences can be self-referential, but only of the innocuous kind. Donald Perlis and Solomon Fefermann showed independently ((Perlis, 1985), (Feferman, 1984)) that the T-schema over a set of positive sentences can consistently be added to any consistent theory.

Another way to exclude viciously self-referential sentences is to make restrictions on universality. As we saw in Section 4, reflection only necessarily leads to self-reference when it is combined with universality. Refraining from having universal sentences about truth like e.g.

> "All sentences are true"

in M we can again obtain a consistent, restricted T-scheme. More precisely, in (Bolander, 2002) it is shown that if none of the sentences of M contain $T(x)$ as a sub-formula with x quantified, then the T-scheme over M can consistently be added to any consistent theory.

Finally, the method of restricting negation and the method of restricting universality can be combined to get an even stronger T-scheme. M can consistently be allowed to contain any sentence in which $T(x)$ does not occur in the scope of negation (see (Bolander, 2002)).

[16] That is, can consistently be added to any consistent theory that does not in advance contain axioms for the T predicate.

12.2 Making the Problematic Part Unproblematic

Another way of ensuring consistency is to stick with self-reference (i.e. all instances of schema T) but to make sure that self-reference does not get the chance to become paradoxical. Such solutions seem again to divide into two categories:

(i) Restricting the form of schema T.

(ii) Restricting the underlying logic.

Below we consider each of these methods.

Restricting the Form of Schema T

Instead of having bi-implications

$$T(\ulcorner \varphi \urcorner) \leftrightarrow \varphi \tag{10}$$

in some cases it is sufficient to have e.g. the following implications

$$T(\ulcorner \varphi \urcorner) \rightarrow \varphi \quad \text{and} \quad T(\ulcorner \varphi \urcorner) \rightarrow T(\ulcorner T(\ulcorner \varphi \urcorner) \urcorner).$$

Some of these *restrictions on the form of schema T* will form consistent extensions to any consistent theory, even if we do not restrict the set of sentences that these schemas are instantiated with. Results of this type can be found in e.g. (Montague, 1963), (Thomason, 1980), and (McGee, 1985). Another possibility is to use a weak equivalence operator in (10) instead of the classical bi-implication operator $\leftrightarrow$. A result concerning such a weak equivalence operator can be found in (Feferman, 1984).

Restricting the underlying logic

Theories containing schema T become inconsistent because in them we can construct self-referential sentences that turn out to be true iff they are false. If we change the underlying logic such that sentences are allowed either to be nor true nor false, or both true and false, the self-referential sentences will no longer be able to prove the theories to be inconsistent. Kripke considers in (Kripke, 1975) the possibility of allowing sentences to have no truth-value, that is, to be neither true nor false. His trick is then to only assign truth-values to the *grounded sentences* of the language (cf. Section 5). By this, he ensures that no self-referential sentence will be given a truth-value (since every self-referential sentence is ungrounded). This corresponds to the fact that in a dictionary, as considered in Section 5, we can only, from the dictionary alone, assign meaning to the grounded words. The ungrounded words, among these the self-referentially defined ones, will be "undecided" (not be assigned any meaning). Kripke's theory can be used to construct formal systems in which we consistently have schema T, but in which the underlying logic is restricted (we cannot have classical negation, for instance, since this requires every sentence to be either true or false). Graham Priest (in (Priest, 1989) and others) proposes that we should allow sentences to be *both* true and false, because this is, in a sense, what paradoxical self-referential sentences are.

13 Conclusion

The paradoxes of self-reference still have no final solution that is generally agreed upon. This makes them, in a sense, *genuine paradoxes.* The presence of a paradox is always a symptom that some part of our fundamental understanding of a subject is crucially flawed. In Zeno's Paradox it was the understanding of infinity that was deficient. In the paradoxes of self-reference it seems that what we do not yet have a proper understanding of is the fundamental relation between something that *refers* (or *represents*) and something that *is referred to* (or *represented*) when these two can not be completely separated. As long as this relationship is not entirely grasped we will probably not get to a full understanding of the paradoxes of self-reference and their consequences for the theories of truth, sets, agent introspection, etc.

In Zeno's Paradox it was not an explicitly stated assumption that later proved to be defective. In the paradox it was implicitly assumed that "infinitely many things can not happen in finite time", but it was not until the development of the mathematical calculus that this assumption could be made explicit and rejected. In the case of Zeno's Paradox it was thus not simply a question of finding the failing assumption involved in the paradox. It was rather a question of discovering a new dimension of the world that had hitherto been hidden to the human eye. A similar thing might very well be the case for the paradoxes of self-reference. The right solution (assuming there is one) to the paradoxes is not to remove or restrict any of our explicit assumptions (that is, restrict schema T, underlying logic, or similar), but to discover a new dimension of the problem that will in the end give more, not fewer, axioms in some kind of extended logic. This new dimension is then expected to make explicit some assumptions about the general relations between referring objects and objects referred to; assumptions that a now invisible to us.[17]

Finally: What would be a suitable concluding remark in an essay like this?[18]

References

Audi, R., editor (1995). *The Cambridge Dictionary of Philosophy.* Cambridge University Press.

Bartlett, S. J., editor (1992). *Reflexivity—A Source-Book in Self-Reference.* North-Holland, Amsterdam.

Bolander, T. (2002). Restricted truth predicates in first-order logic. *(submitted for publication).* To be represented at LOGICA 2002.

[17] It has been proposed to solve the paradoxes of self-reference by extending logic to include *contexts.* Then a paradox such as the Liar will be resolved by the fact that the context before the Liar sentence is uttered is different from the context after it has been uttered. Such a solution seems to be of the kind I propose here, since the logic is extended (with contexts) rather than restricted. But, it should be noted that if we are allowed freely to refer to contexts then the Liar sentence can be strengthened to give a paradox even in this theory.

[18] Answer: A self-referential question which is its own answer.

Boolos, G. (1989). A new proof of the Gödel Incompleteness Theorem. *Notices of the American Mathematical Society*, 36:388–390. Reprinted in (Boolos, 1998).

Boolos, G. (1998). *Logic, Logic, and Logic.* Harvard University Press.

Cantor, G. (1932). *Gesammelte Abhandlungen.* Springer Verlag.

Erickson, G. W. and Fossa, J. A. (1998). *Dictionary of paradox.* University Press of America.

Feferman, S. (1984). Toward useful type-free theories I. *The Journal of Symbolic Logic*, 49(1):75–111.

Gaifman, H. (1992). Pointers to truth. *Journal of Philosophy*, 89(5):223–261.

Gilmore, P. C. (1974). The consistency of partial set theory without extensionality. In *Axiomatic set theory (Proc. Sympos. Pure Math., Vol. XIII, Part II, Univ. California, Los Angeles, Calif., 1967)*, pages 147–153. Amer. Math. Soc.

Gödel, K. (1931). Über formal unentscheidbare Satze der Principia Mathematica und verwandter Systeme I. *Monatshefte für Mathematik und Physik*, 38:173–198. Reprinted in (Gödel, 1986).

Gödel, K. (1986). *Collected works. Vol. I.* Oxford University Press. Publications 1929–1936, Edited and with a preface by Solomon Feferman.

Hatcher, W. S. (1982). *The Logical Foundations of Mathematics.* Pergamon Press.

Kripke, S. (1975). Outline of a theory of truth. *The Journal of Philosophy*, 72:690–716. Reprinted in (Martin, 1984).

Martin, R. L., editor (1984). *Recent Essays on the Liar Paradox.* Oxford University Press.

McGee, V. (1985). How truthlike can a predicate be? A negative result. *Journal of Philosophical Logic*, 14(4):399–410.

McGee, V. (1992). Maximal consistent sets of instances of Tarski's schema (T). *Journal of Philosophical Logic*, 21(3):235–241.

Mendelson, E. (1997). *Introduction to Mathematical Logic.* Chapman & Hall, 4 edition.

Montague, R. (1963). Syntactical treatments of modality, with corollaries on reflection principles and finite axiomatizability. *Acta Philosophica Fennica*, 16:153–166.

Perlis, D. (1985). Languages with self-reference I. *Artificial Intelligence*, 25:301–322.

Perlis, D. and Subrahmanian, V. S. (1994). Meta-languages, reflection principles and self-reference. In *Handbook of Logic in Artificial Intelligence and Logic Programming*, volume 2, pages 323–358. Oxford University Press.

Priest, G. (1989). Reasoning about truth. *Artificial Intelligence*, 39(2):231–244.

Priest, G. (1994). The structure of the paradoxes of self-reference. *Mind*, 103(409):25–34.

Tarski, A. (1956). The concept of truth in formalized languages. In *Logic, semantics, metamathematics*. Hackett Publishing Co., Indianapolis, IN. Papers from 1923 to 1938.

Thomason, R. H. (1980). A note on syntactical treatments of modality. *Synthese*, 44(3):391–395.

Wittgenstein, L. (1958). *Blue and Brown Books*. Blackwell.

Yablo, S. (1993). Paradox without self-reference. *Analysis*, 53(4):251–252.