Self-Referential Paradoxes

Noson S. Yanofsky

Brooklyn College, CUNY

February 17, 2021

My Agenda

Some of the most profound and famous theorems in mathematics and computer science of the past 150 years are instances of self-referential paradoxes.

- Georg Cantor's theorem that shows there are different levels of infinity;
- Bertrand Russell's paradox which proves that simple set theory is inconsistent;
- Kurt Gödel's famous incompleteness theorems that demonstrates a limitation of the notion of proof;
- Alan Turing's realization that some problems can never be solved by a computer;
- and much more.

Amazingly, all these diverse theorems can be seen as instances of a single simple theorem of basic category theory. We describe this theorem and show some of the instances. No category theory is needed for this talk.

My Hidden Agenda

This talk is a single section (of 58) of an introductory category theory textbook I am writing called

MONOIDAL CATEGORIES
A Unifying Concept in Mathematics, Physics, and Computing

The goal is to show the power of category theory (and to get you to buy the book!!!) The point is that with a little category theory, one can know a hell-of-a-lot of mathematics, physics, and computers.

Outline of the Talk

1. Some Preliminaries
2. Three Motivating Examples
3. Philosophical Interlude
4. Cantor's Inequalities
5. The Main Theorem About Self-Referential Paradoxes
6. Turing's Halting Problem
7. Fixed Points in Logic
8. Two Other Paradoxes
9. Further Reading

Some Preliminaries

Before we leap into all the examples, there are some technical ideas about sets.

- Let $2 = \{0, 1\}$ be a set with two values which correspond to true and false.
- Let S be a set. A function $g\colon S \longrightarrow 2$ is a characteristic function and describes a subset of S.
- Consider a set function $f\colon S \times S \longrightarrow 2$. The f accepts two elements of S and outputs either 0 or 1. Think of

  $$f(a, b) = 1$$

  meaning "*a* is a part of *b*" or "*a* is described by *b*."

Some Preliminaries

- For any element s_0 of S, consider the function f where the second input is always s_0. We say that s_0 is "hardwired into the function." This gives us a function

  $$f(\ \ ,s_0) \colon S \longrightarrow 2$$

  with only one input. Since this function goes from S to 2, it also is a characteristic function and describes a subset of S. The subset is

  $$\{s \in S : f(s, s_0) = 1\}.$$

  It is all the elements that are "part of s_0" or all the elements that are "described by s_0."

Some Preliminaries

- For different f's and various elements of S, there are different characteristic functions which describe various subsets.
- We now ask a simple question: given $g\colon S \longrightarrow 2$ and $f\colon S \times S \longrightarrow 2$, is there an s_0 in S such that g characterizes the same subset as $f(\ \ , s_0)$?
- To restate, for a given g and f, does there exist an $s_0 \in S$ such that $g(\ \ ) = f(\ \ , s_0)$? If such an s_0 exists, then we say g can be **represented** by f.

Some Preliminaries

- For every set S there is a set map $\Delta\colon S \longrightarrow S \times S$ called the **diagonal map** that takes an element t to the ordered pair (t, t). This is the core of self reference.
- If $f\colon S \times S \longrightarrow 2$ is a function that evaluates the relationship of S elements to S elements, then

$$\begin{tikzcd}
 & S \times S \arrow[r, "f"] & 2 \\
S \arrow[ur, "\Delta"] & &
\end{tikzcd}$$

takes every element $t \in S$ as follows

$$t \longmapsto (t, t) \longmapsto f(t, t).$$

This evaluates t with itself.

The Barber Paradox

Bertrand Russell was a great expositor. The **barber paradox** is attributed to Russell and is used to explain some of the central ideas of self-referential systems. Imagine an isolated village in the Austrian alps where it is difficult for villagers to leave and for itinerant barbers to come to the village. This village has exactly one barber and there is a strict rule that is enforced:

A villager cuts his own hair iff he does not go to the barber.

If the villager will cut his own hair, why should he go to the barber? On the other hand, if the villager goes to the barber, he will not need to cut his own hair. This works out very well for the villagers except for one: the barber. Who cuts the barber's hair? If the barber cuts his own hair, then he is violating the village ordinance by cutting his own hair and having his hair cut by the barber. If he goes to the barber, then he is also cutting his own hair. This is illegal!

The Barber Paradox

Let us formalize the problem. Let the set *Vill* consist of all the villagers in the village. The function

$$f : Vill \times Vill \longrightarrow 2$$

describes who cuts whose hair in the village. It is defined for villagers v and v' as

$$f(v, v') = \begin{cases} 1 & : \text{ if the hair of } v \text{ is cut by } v' \\ 0 & : \text{ if the hair of } v \text{ is not cut by } v'. \end{cases}$$

We can now express the village ordinance as saying that for all v

$$f(v, v) = 1 \quad \text{if and only if} \quad f(v, \text{barber}) = 0.$$

This is true for all v including $v = \text{barber}$. In this case we get:

$$f(\text{barber}, \text{barber}) = 1 \quad \text{if and only if} \quad f(\text{barber}, \text{barber}) = 0.$$

This is a contradiction!

The Barber Paradox

Let us be more categorical. There is

- The diagonal set function $\Delta\colon Vill \longrightarrow Vill \times Vill$ that is defined as $\Delta(v) = (v, v)$.
- There is also a negation function $NOT\colon 2 \longrightarrow 2$ defined as $NOT(0) = 1$ and $NOT(1) = 0$.
- Composing f with Δ and NOT gives us $g\colon Vill \longrightarrow 2$ as in the following commutative diagram:

$$\begin{tikzcd}
& Vill \times Vill \arrow[r, "f"] & 2 \arrow[dr, "NOT"] & \\
Vill \arrow[ur, "\Delta"] \arrow[rrr, "g"'] & & & 2
\end{tikzcd}$$

$$g = NOT \circ f \circ \Delta.$$

The Barber Paradox

For a villager v,

$$g(v) == NOT(f(\Delta(v))) = NOT(f(v, v)).$$

So

$$g(v) = 1$$

if and only if

$$NOT(f(v, v)) = 1$$

if and only if

$$f(v, v) = 0$$

if and only if
the hair of v is not cut by v.
In other words $g(v) = 1$ if and only if v does not cut his own hair.
g is the characteristic function of the subset of villagers who do not cut their own hair.

The Barber Paradox

We now ask the simple question: can g be represented by f? In other words, is there a villager v_0 such that $g(\quad) = f(\quad, v_0)$? It stands to reason that the barber is the villager who can represent g. After all, $f(\quad, \text{barber})$ describes all the villagers who get their hair cut by the barber.

$$g(v) = f(v, \text{barber})$$

What about $v = \text{barber}$.

$$g(\text{barber}) = f(\text{barber}, \text{barber}).$$

But the definition of g is given as $g(\text{barber}) = NOT(f(\text{barber}, \text{barber}))$. We conclude that g is not represented by $f(\quad, \text{barber})$ That is, the set of villagers who do not cut their own hair can not be the same as the set of villagers who get their hair cut by anyone.

The Barber Paradox — Matrix Form

It is helpful to describe this problem in matrix form. Let us consider the set $Vill$ as $\{v_1, v_2, v_3, \ldots, v_n\}$. We can then describe the function $f\colon Vill \times Vill \longrightarrow 2$ as a matrix. Let us say that the barber is v_4. Notice that every row has exactly one 1 (every villager gets their haircut in only one place): either along the diagonal (the villager cuts their own hair) or in the v_4 column (the villager goes to the barber.) Since it can only be one or the other, the numbers along the diagonal $1, 0, 1, ?, 0, \ldots, 1$ are almost the exact opposite of the numbers along the v_4 column $0, 1, 0, ?, 1, \ldots, 0$. This is a restatement of the rule of the village. There is only one problem: what is in the (v_4, v_4) position. We put a question mark because that entry cannot be the opposite of itself. This way of seeing the problem will arise over and over again. Here we can see why these paradoxes are related to proofs called **diagonal arguments**.

The Barber Paradox — Matrix Form

$$
\begin{array}{cc|ccccccc}
 & & \multicolumn{7}{c}{\textbf{Cutter}} \\
 & f & v_1 & v_2 & v_3 & v_4 & v_5 & \cdots & v_n \\
\hline
\textbf{Cuttee} & v_1 & 1 & 0 & 0 & 0 & 0 & \cdots & 0 \\
 & v_2 & 0 & 0 & 0 & 1 & 0 & \cdots & 0 \\
 & v_3 & 0 & 0 & 1 & 0 & 0 & \cdots & 0 \\
 & v_4 & 0 & 0 & 0 & ? & 0 & \cdots & 0 \\
 & v_5 & 0 & 0 & 0 & 1 & 0 & \cdots & 0 \\
 & \vdots & \vdots & & & \vdots & & & \vdots \\
 & v_n & 0 & 0 & 0 & 0 & 0 & \cdots & 1
\end{array}
$$

The Barber Paradox — Avoiding Contradictions

What is the resolution to this paradox? There are many attempts to solve this paradox, but they are not very successful. For example, the barber resigns as barber before cutting his own hair. (But that means that there is no barber in the town). Or the wife of the barber cuts the barber's hair. (But that means that there are two barbers in the town.) Or the barber is bald. Or the barber is a long-haired hippie. Or the rule is ignored while the barber cuts his own hair, etc. All these are saying the same thing: the village with this important rule cannot exist. Because if the village with this rule existed, there would be a contradiction. There are no contradictions in the physical world. The only way the world can be free of contradictions is if this proposed village with this strict rule does not exist.

Russell's Paradox

This paradox concerns sets which are considered the foundation of much of mathematics. As is known, sets contain elements. The elements can be anything. In particular an element in a set can be a set itself. A set containing itself is also not so strange. Here are three examples of sets that contain themselves:

- The set of all ideas discussed in this talk
- The set that contains all the sets that have more than three objects
- The set of abstract ideas.

Russell's Paradox

If you do not like sets that contain themselves, you might want to consider "Russell's set" which is the set of all sets that do not contain themselves. Formally,

$$R = \{\text{set } S : S \text{ does not contain } S\} = \{S : S \notin S\}.$$

Now ask yourself the simple question: does R contain itself? In symbols, we ask if $R \in R$? Let us consider the possible answers. If $R \in R$, then since R fails to satisfy the requirements of being a member of R, we get that $R \notin R$. In contrast, if $R \notin R$, then since R satisfies the requirement of belonging to R, we have that $R \in R$. This is a contradiction.

Russell's Paradox

Let us formulate this. There is a collection of all sets called *Set*. There is also a two-place function $f\colon Set \times Set \longrightarrow 2$ that describes which sets are elements of which other sets.

$$f(S,S') = \begin{cases} 1 & : \text{if } S \in S' \\ 0 & : \text{if } S \notin S'. \end{cases}$$

$$\begin{tikzcd}
 & Set \times Set \arrow[r, "f"] & 2 \arrow[dr, "NOT"] & \\
Set \arrow[ur, "\Delta"] \arrow[rrr, "g"'] & & & 2
\end{tikzcd}$$

The value $g(S) = NOT(f(S,S))$. This means $g(S) = 1$ if and only if $f(S,S) = 0$. g is the characteristic function of those sets that do not contain themselves.

Russell's Paradox

Now we ask the simple question: does there exist a set R such that $g(\quad)$ is represented by f as $f(\quad, R)$. That is, we want a set R such that

$$g(S) = 1 \quad \text{if and only if} \quad f(S, R) = 1$$

and

$$g(S) = 0 \quad \text{if and only if} \quad f(S, R) = 0.$$

This means that R contains only the sets that do not contain themselves. The problem is that if such a set R exists, then we can ask about $g(R)$, i.e., is $R \in R$. On the one hand $g(R)$ is defined as $NOT(f(R, R))$ and on the other hand, if f represents g with R, then $g(R) = f(R, R)$. That is,

$$f(R, R) = g(R) = NOT(f(R, R)).$$

This is a contradiction.

Russell's Paradox — Matrix Form

Let us look at Russell's paradox from a matrix point of view. Consider the infinite collection Set as $\{S_1, S_2, S_3, \ldots\}$. We can then describe the function $f\colon Set \times Set \longrightarrow 2$ as a matrix. Notice that the diagonal is different than every column of the array. This is a way of saying that that the diagonal (which is g) cannot be represented by any column of the array.

Russell's Paradox — Matrix Form

$$\begin{array}{cc|cccccc}
 & & \multicolumn{6}{c}{\textbf{Subset}} \\
 & f & S_1 & S_2 & S_3 & S_4 & S_5 & \cdots \\ \hline
\textbf{Element} & S_1 & \neg 1 = 0 & 0 & 0 & 0 & 1 & \cdots \\
 & S_2 & 0 & \neg 0 = 1 & 0 & 1 & 0 & \cdots \\
 & S_3 & 0 & 0 & \neg 1 = 0 & 0 & 0 & \cdots \\
 & S_4 & 1 & 0 & 0 & \neg 1 = 0 & 0 & \cdots \\
 & S_5 & 0 & 1 & 0 & 1 & \neg 0 = 1 & \cdots \\
 & \vdots & \vdots & & & \vdots & & \ddots
\end{array}$$

Russell's Paradox — Avoiding Contradictions

The only way to avoid this contradiction is to accept that the function g cannot be represented by any element of Set. This translates into meaning that the collection of all sets that do not contain themselves does not form a set, i.e., this collection is not an element of Set. While such a collection seems to be a well-defined notion, we have shown that if we say that this collection is an element of Set, then there is a contradiction.

Heterological Paradox

Now for a linguistic paradox. The **heterological paradox**, also called **Grelling's paradox** after Kurt Grelling, who first formulated it, is about adjectives (words that modify nouns). Consider several adjectives and ask if they describe themselves.

- "English" is English.
- "French" is not French ("francais" is francais.)
- "German" is not German ("Deutsch" is Deutsch.)
- "abbreviated" is not abbreviated,
- "unabbreviated" is unabbreviated
- "hyphenated" is not hyphenated, etc.

Call all adjectives that describe themselves "autological." In contrast, call all adjectives that do not describe themselves as "heterological."

Heterological Paradox

autological	**heterological**
English	non-English
	French
	German
noun	verb
unhyphenated	hyphenated
unabbreviated	abbreviated
polysyllabic	monosyllabic
⋮	⋮

Heterological Paradox

Let us ask a simple question? Is "heterological" heterological? That is, does it belong on the left side or the right side of the table? Let us go through the two possibilities.

- If "heterological" is not heterological, then it does not describe itself and therefore it *is* heterological.
- If "heterological" is heterological, then it does describe itself and therefor is *not* heterological.

This is a contradiction.

Heterological Paradox

Let us formulate this paradox categorically. There is a set Adj of adjectives and a function $f\colon Adj \times Adj \longrightarrow 2$ which is defined for adjectives a and a' as follows:

$$f(a,a') = \begin{cases} 1 & : \text{if } a \text{ is described by } a' \\ 0 & : \text{if } a \text{ is not described by } a'. \end{cases}$$

Use f to formulate g as as the composition of the following three maps:

$$\begin{tikzcd}
& Adj \times Adj \arrow[r, "f"] & 2 \arrow[dr, "NOT"] & \\
Adj \arrow[ur, "\Delta"] \arrow[rrr, "g"'] & & & 2.
\end{tikzcd}$$

The function g is the characteristic function of those adjectives that do not describe themselves.

Heterological Paradox

Can g be represented by some element in Adj? Is there some adjective, say "heterological," that we can use in f to represent g? That is, is it true that $g(\quad) = f(\quad, \text{``heterological''})$?

$$g(A) = f(A, \text{``heterological''}).$$

For all A including $A =$"heterological"

$$g(\text{``heterological''}) = f(\text{``heterological''}, \text{``heterological''}).$$

But that would give a contradiction because by the definition of g we have

$$g(\text{``heterological''}) = NOT(f(\text{``heterological''}, \text{``heterological''})).$$

The only conclusion we can come to is that $g(\quad)$ cannot be represented by f. That is, the set of all adjectives that do not describe themselves cannot be represented by "heterological". However, that is exactly the definition of "heterological"!

Heterological Paradox — Matrix Form

The hetrological paradox can be described with a matrix similar to the earlier matrices. The set $Adj = \{A_1, A_2, A_3, \ldots\}$. Again we would have a changed diagonal that would be different than every column in the array.

Heterological Paradox — Avoiding Contradictions

How do we avoid this little paradox? There are two possible ways of resolving this paradox.

- Many philosophers say that the word “heterological” cannot exist. After all, we just showed that it is not always well-defined. We cannot determine if a certain adjective (“hetrological”) is heterological or not.
- Another more obvious solution is to just ignore the problem. Human language is inexact and full of contradictions. Every time we use an oxymoron, we are stating a contradiction. Every time we ask for another piece of cake while lamenting the fact that we cannot lose weight, we are stating a contradiction. We can safely ignore the fact that heterological is not well-defined for only one adjective.

A philosophical interlude on paradoxes

A paradox is a process where an assumption is made, and through valid reasoning, a contradiction is derived.

$$\text{Assumption} \Longrightarrow \text{Contradiction.}$$

The logician then concludes that since the reasoning was valid and the contradiction cannot happen, it must be that the assumption was wrong. This is very similar to what mathematicians call "proof by contradiction" and philosophers call "*reductio ad absurdum*." A paradox is a method of showing that the assumption is not part of rational thought.

We have so far seen the same pattern of proof in three different areas: (i) villagers, (ii) sets, and (iii) adjectives. The assumption is that the g function can be represented by the f function. A contradiction is then derived and we conclude that g is not represented by f. These three examples highlight three different areas where the alleged contradictions might be found.

A philosophical interlude on paradoxes

The Physical Universe. A village with a particular rule is part of the physical universe. The physical universe does not have any contradictions. Facts and properties simply are and no object can have two opposing properties. Whenever we come to such contradictions, we have no choice but to conclude that the assumption was wrong.

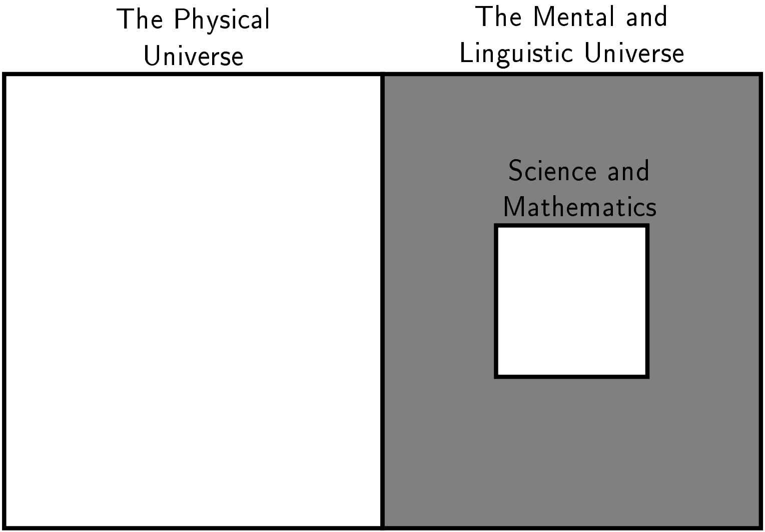
A philosophical interlude on paradoxes

The Mental and Linguistic Universe. In contrast to the physical universe, the human mind and human language — that the mind uses to express itself — are full of contradictions. We are not perfect machines. We have a lot of different contradictory parts and desires. We all have conflicting thoughts in our head and these thoughts are expressed in our speech. So when an assumption brings us to a contradiction in our thought or language, we do not need to take it very seriously. If an adjective is in two opposite classifications, it does not really bother us. In such a case, we cannot go back to our assumption and say it is wrong. The entire paradox can be ignored.

A philosophical interlude on paradoxes

Science and Mathematics. There are, however, parts of human thought and language which cannot tolerate contradictions: science and mathematics. These areas of exact thought are what we use to discuss the physical world (and more). If science and math are to discuss / describe / model / predict the contradiction-free physical universe, then we better make sure that no contradictions occur there. We first saw this in the early years of elementary school when our teachers proclaimed that we are not permitted to divide by zero. Since math and science cannot have contradictions, young fledglings are not permitted to divide by zero. To summarize, science and mathematics are products of the human mind and language which we do not permit to have contradictions. If an assumption leads us to a contradiction in science or mathematics, then we must abandon the assumption.

A philosophical interlude: Where contradictions can occur.



The Physical Universe

The Mental and Linguistic Universe

Science and Mathematics

A philosophical interlude — Why categories?

Many have felt that these different instances of self reference have a similar pattern (witness Bertrand Russell supposedly inventing the barber paradox to illustrate Russell's paradox.) The major advance that category theory has to offer the subject of self-referential paradoxes is to actually show that all these different self-referential paradoxes are really instances of the same categorical theorem. F. William Lawvere described a simple formalism that showed many of the major self-referential paradoxes and more. This shows that the logic of self-referential paradoxes is inherent in many systems. This also shows the unifying power of category theory.

A philosophical interlude — Why categories?

There is, however, another positive aspect of our formalism. Lawvere showed us how to have an exact mathematical description of the paradoxes while avoiding messy statements about what exists and what does not exist. In the categorical setting,

- The barber paradox does not say that a village with a rule does not exists.
- With Russell's paradox, a category theorist does not say that a certain collection does not form a set.
- In the heterological paradox, we avoid the silly analysis as to whether a word exists or not.

In our categorical discussion, we successfully avoid metaphysical gobbledygook. For this alone, we should be appreciative of the categorical formalism.

Cantor's Inequalities

At the end of the 19th century Georg Cantor proved some important theorems about the sizes of sets.

- He showed that every set is smaller than its powerset.
- Every set S is smaller than the set $\mathcal{P}(S)$.
- A more categorical way of saying this is that for any set S, there cannot exist a surjection $h\colon S \longrightarrow \mathcal{P}(S)$.
- Yet another way of saying this, is that for every purported surjection $h\colon S \longrightarrow \mathcal{P}(S)$, there is some subset of S, denoted C_h, that is not in the image of h.

This is proven with a proof by contradiction: we are going to assume (wrongly) that there is such a surjection and derive a contradiction. Since this is formal mathematics, no such contradiction can exist and hence our assumption that such a surjection exists must be false.

Cantor's Inequalities

Given such an h, let us define $f_h \colon S \times S \longrightarrow 2$ for $s, s' \in S$ as follows

$$f_h(s, s') = \begin{cases} 1 & : s \in h(s') \\ 0 & : s \notin h(s') \end{cases}$$

Use f_h to construct g_h as follows

$$\begin{tikzcd}
& S \times S \arrow[r, "f_h"] & 2 \arrow[rd, "NOT"] & \\
S \arrow[ru, "\Delta"] \arrow[rrr, "g_h"'] & & & 2
\end{tikzcd}$$

The function g_h is the characteristic function of the subset $C_h \subseteq S$ where each element s does not belong to $h(s)$, i.e.,

$$C_h = \{s \in S : s \notin h(s)\} \subseteq S.$$

Cantor's Inequalities

We claim that the subset C_h of S is not in the image of h, i.e., C_h is a "witness" or a "certificate" that h is not surjective. If C_h was in the image of h, there would be some $s_0 \in S$ such that $h(s_0) = C_h$. In that case g_h would be represented by f_h with s_0. That is, for all $s \in S$

$$g_h(s) = f_h(s, s_0)$$

but this would also be true for $s_0 \in S$ which would mean that

$$g_h(s_0) = f_h(s_0, s_0)$$

However, by the definition of g_h, we have that

$$g_h(s_0) = NOT(f_h(s_0, s_0)).$$

Since this cannot be, our assumption that $h(s_0) = C_h$ is wrong and there is a subset of S that is not in the image of h.

Cantor's Inequalities — Avoiding Contradictions

This is part of mathematics and the only resolution is to accept the fact no such surjective h exists and that $|S| < |\mathcal{P}(S)|$. Notice that this applies to any set. For finite S, this is obvious since $|S| = n$ implies $|\mathcal{P}(S)| = 2^n$. However, this is true for infinite S also. What this shows is that $\mathcal{P}(S)$ is a different level of infinity than S. One can iterate this process and get

- $\mathcal{P}(S)$,
- $\mathcal{P}(\mathcal{P}(S))$,
- $\mathcal{P}(\mathcal{P}(\mathcal{P}(S)))$,
- $\ldots$

This gives many different, unequal levels of infinity.

Cantor's Inequalities

Related to the above theorem of Cantor, is the theorem that the natural numbers N is smaller than the interval of all real numbers between 0 and 1, i.e., $(0,1) \subseteq$ R.
This proof is slightly different than the previous examples that we saw. We include it because it has features in it that are closer to the upcoming general theorem. Rather than working with the set $2 = \{0,1\}$, this proof works with the set $10 = \{0,1,2,3,\ldots 9\}$. Also, rather than working with the function $NOT: 2 \longrightarrow 2$, we now work with the function $\alpha: 10 \longrightarrow 10$ which is defined as follows:

$$\alpha(0) = 1, \quad \alpha(1) = 2, \quad \alpha(2) = 3, \ldots, \quad \alpha(8) = 9, \quad \alpha(9) = 0,$$

i.e., $\alpha(n) = n+1 \;\; Mod \;\; 10$. The most important feature of α is that, every output is different than its input. There are many such functions from 10 to 10. We choose this one.

Cantor's Inequalities — Matrix Form

The proof that $|\mathrm{N}| < |(0,1)|$ is, again, a proof by contradiction. We assume (wrongly) that there is a surjection $h\colon \mathrm{N} \longrightarrow (0,1)$ and come to a contradiction which proves that no such h can possibly exist. With such an h we can define a function $f_h\colon \mathrm{N} \times \mathrm{N} \longrightarrow 10$ which depends on h. For $m, n \in \mathrm{N}$,

$$f_h(m,n) = \text{the } m\text{th digit of } h(n).$$

This means that f_h gives every digit of the purported function h. The next slide will help explain f_h. The natural numbers are on the left tell you the position. The function h assigns to every natural number on the top, a real number below it. The numbers on the left are the first inputs to f_h and the numbers on the top are the second inputs to f_h.

Cantor's Inequalities — Matrix Form

$$\begin{array}{cc|cccccccc}
 & & \multicolumn{8}{c}{\textbf{Real Number}} \\
f_h & & 0 & 1 & 2 & 3 & 4 & 5 & 6 & \cdots \\ \hline
 & & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
 & & . & . & . & . & . & . & . & \cdots \\
\textbf{Digit} & 0 & 0 & 0 & 7 & 2 & 7 & 7 & 4 & \cdots \\
 & 1 & 0 & 1 & 2 & 2 & 7 & 6 & 7 & \cdots \\
 & 2 & 5 & 3 & 0 & 3 & 0 & 0 & 0 & \cdots \\
 & 3 & 0 & 6 & 2 & 0 & 1 & 2 & 0 & \cdots \\
 & 4 & 0 & 1 & 0 & 2 & 3 & 1 & 3 & \cdots \\
 & 5 & 0 & 1 & 0 & 3 & 0 & 1 & 5 & \cdots
\end{array}$$

Cantor's Inequalities

With such an f_h, one can go on to describe a function g_h with the — by now familiar — construction

$$\begin{tikzcd}
& \mathbf{N} \times \mathbf{N} \arrow[r, "f_h"] & \mathbf{10} \arrow[dr, "\alpha"] & \\
\mathbf{N} \arrow[ur, "\Delta"] \arrow[rrr, "g_h"'] & & & \mathbf{10}
\end{tikzcd}$$

The function g_h also depends on h. The next matrix will help explain the function g_h. That is, the nth digit of the nth number is changed. The changed numbers are the outputs to the function g_h. Thinking of the outputs of g_h as the digits of a real number, we are describing a real number between 0 and 1. We call this number G_h. In our case,

$$G_h = 0.121126\ldots$$

Cantor's Inequalities — Matrix Form

		Real Number					
f_h		0	1	2	3	4	5
	0	0	0	0	0	0	0
		.	.	.	.	.	.
Digit	0	$\alpha(0) = 1$	0	7	2	7	7
	1	0	$\alpha(1) = 2$	2	2	7	6
	2	5	3	$\alpha(0) = 1$	3	0	0
	3	0	6	2	$\alpha(0) = 1$	1	2
	4	0	1	0	2	3	$\alpha(1) =$

Cantor's Inequalities

The claim is that g_h is not represented by f_h. This means that the number represented by g_h will not be the number represented by $f_h(\quad, n_0)$ for any n_0. Another way to say this is that the number G_h will not be any column in the scheme described by the matrix. This is obviously true because G_h was formed to be different than the first column because the first digit is different. It is different than the second column because it was formed to be different at the second digit. It is different than the third column because it was formed to be different at the third digit, etc.

G_h is saying

"I am not on column n because my nth digit is different from the nth column's nth digit."

or

"I am not in the image of h."

Cantor's Inequalities

Let us show the end of the proof formally. If there was some n_0 that represented g_h, then for all m

$$g_h(m) = f_h(m, n_0)$$

(i.e., G_h is the same as column n_o.) But if this was true for all m, then it is true for n_0 also (that is, it is true by every digit including the one on the diagonal.) But that says that

$$g_h(n_0) = f_h(n_0, n_0).$$

However g_h was defined for n_0 as

$$g_h(n_0) = \alpha(f_h(n_0, n_0)).$$

We conclude that no such n_0 exists and g_h describes a number in $(0, 1)$ but is not in the image of h. That is, h cannot be surjective and $|\mathrm{N}| < |(0, 1)|$.

Main Theorem — Some Preliminaries

Definition. First a simple definition in $\mathbb{S}\mathfrak{et}$. Consider a set Y and a set function $\alpha\colon Y \longrightarrow Y$. We call $s_0 \in Y$ a **fixed point** of α if $\alpha(s_0) = s_0$. That is, the output is the same (or fixed) as the input. We can write the element s_0 by talking about a function $p\colon \{*\} \longrightarrow Y$ such that $p(*) = s_0$. Saying that s_0 is a fixed point of α amounts to saying that $\alpha \circ p = p$, i.e., the following diagram commutes:

$$\begin{tikzcd}
\{*\} \arrow[rr, "p"] \arrow[dr, "p"'] & & Y \arrow[dl, "\alpha"] \\
& Y. &
\end{tikzcd}$$

Let us generalize this to any category $\mathbb{A}$ with a terminal object 1. Let y be an object in $\mathbb{A}$ and $\alpha\colon y \longrightarrow y$ be a morphism in $\mathbb{A}$. Then we say $p\colon 1 \longrightarrow y$ is a **fixed point** of α if $\alpha \circ p = p$.

Lawvere's Theorem

Now for the main theorem as given by Lawvere in 1969.

Lawvere's Theorem. Let $\mathbb{A}$ be a category with a terminal object and binary products. Let y be an object in the category and $\alpha\colon y \longrightarrow y$ be a morphism in the category. If α does not have a fixed point, then for all objects a and for all $f\colon a \times a \longrightarrow y$ there exists a $g\colon a \longrightarrow y$ such that g is not representable by f.

Lawvere's Theorem

Proof. Let $\alpha \colon y \longrightarrow y$ not have a fixed point, then for any a and for any $f \colon a \times a \longrightarrow y$ we can compose f with Δ and α to form g as below.

$$\begin{tikzcd}
& a \times a \arrow[r, "f"] & y \arrow[dr, "\alpha"] & \\
a \arrow[ur, "\Delta"] \arrow[rrr, "g"'] & & & y.
\end{tikzcd}$$

g is not representable by f.

Turing's Halting Problem

In the early 1930's, long before the engineers actually created computers, Alan Turing, the "father of computer science," showed what computers *cannot* do. Loosely speaking, he proved that no program can decide whether or not any program will go into an infinite loop or not. Already from this inexact statement one can see the self reference: programs deciding properties of programs.

Let us state a more exact version of Turing's theorem. First some preliminaries. Programs come in many different forms. Here we are concerned with programs that only accept a single natural number. To every such program, there is a unique natural number that describes that program. This fact that programs that act on numbers can be represented by numbers shows that programs can be self referential.

Turing's Halting Problem

Programs that accept a single number can take an input and halt or they can go into an infinite loop. The halting problem asks for (i) a number of a program that accepts a single number and (ii) an input to that program. It returns 1 or 0 depending on if it halts or goes into an infinite loop. Turing's Halting theorem says that no such program can possibly exist. This is not a limitation of modern technology or of our current ability. Rather, this is an inherent limitation of computation.

Turing's Halting Problem

The proof is, once again, a proof by contradiction. Assume (wrongly) that there does exist a program that accepts a program number and an input, and can determine if that program will halt or go into an infinite loop when that number is entered into that program. Formally, such a program describes a total computable function. The function is named $Halt \colon \mathrm{N} \times \mathrm{N} \longrightarrow Bool$ defined on natural numbers $m, n \in \mathrm{N}$ is

$$Halt(m,n) = \begin{cases} 1 & : \text{if input } m \text{ into program } n \text{ halts} \\ 0 & : \text{if input } m \text{ into program } n \text{ goes into a loop} \end{cases}$$

Turing's Halting Problem — Matrix Form

$$\begin{array}{cc|ccccccc}
 & & \multicolumn{7}{c}{\textbf{Program}} \\
 & \textit{Halt} & 0 & 1 & 2 & 3 & 4 & 5 & \cdots \\ \hline
 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & \cdots \\
 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & \cdots \\
\textbf{Input} & 2 & 0 & 1 & 0 & 0 & 0 & 0 & \cdots \\
 & 3 & 0 & 1 & 0 & 0 & 1 & 0 & \cdots \\
 & 4 & 0 & 0 & 1 & 1 & 1 & 1 & \cdots \\
 & 5 & 1 & 0 & 1 & 0 & 1 & 1 & \cdots \\
 & \vdots & & & \vdots & & & \ddots &
\end{array}$$

Turing's Halting Problem

It is not hard to see that the function $\Delta\colon \mathbb{N} \longrightarrow \mathbb{N} \times \mathbb{N}$ defined as $\Delta(n) = (n, n)$ is a computable function. Consider the partial NOT function $ParNOT\colon Bool \longrightarrow Bool$ defined as follows:

$$ParNOT(n) = \begin{cases} 1 & : \text{if } n = 0 \\ \uparrow & : \text{if } n = 1 \end{cases}$$

where ↑ means it goes into an infinite loop. It is not hard to see that *ParNOT* is a computable function. Since *Halt* is assumed computable, and the function Δ and *ParNOT* are computable, then their composition as follows is also computable function.

$$\begin{tikzcd}
 & \mathbb{N} \times \mathbb{N} \arrow[r, "Halt"] & Bool \arrow[dr, "ParNOT"] & \\
\mathbb{N} \arrow[ur, "\Delta"] \arrow[rrr, "Halt'"'] & & & Bool
\end{tikzcd}$$

Turing's Halting Problem — Matrix Form

The new computable function, $Halt'$, accepts a number n as input and does the opposite of what program n on input n does. That is, if program n on input n halts, then $Halt'(n)$ will go into an infinite loop. Otherwise, if program n on input n goes into an infinite loop, then $Halt'(n)$ will halt. We can see the way $Halt'$ is defined in the next slide.

Turing's Halting Problem — Matrix Form

		Program				
	Halt	0	1	2	3	4
Input	0	$\alpha(0) = 1$	1	0	0	0
	1	1	$\alpha(1) = \uparrow$	1	1	1
	2	0	1	$\alpha(0) = 1$	0	0
	3	0	1	0	$\alpha(0) = 1$	1
	4	0	0	1	1	$\alpha(1) = \uparrow$
	5	1	0	1	0	1
	⋮			⋮		

Turing's Halting Problem

Since $Halt'$ is a computable function, the program for this computable function must have a number and be somewhere on our list of computable functions. However, it is not. $Halt'$ was formed to be different than every column in the chart. What is wrong? We know that $\triangle$ and $ParNOT$ are computable. We assumed that $Halt$ was computable. It must be that our assumption about $Halt$ was wrong. $Halt$ is not computable. Let us formally show that $Halt'$ is different than every column in the chart. Imagine that $Halt'$ is computable and the number of $Halt'$ is n_0. This means that $Halt'$ is the n_0 column of our chart. Another way to say this is that $Halt'$ is representable by $Halt(\quad, n_0)$, i.e., for all n,

$$Halt'(n) = Halt(n, n_0).$$

Now let us ask about $Halt'(n_0)$? We get

$$Halt'(n_0) = Halt(n_0, n_0).$$

Turing's Halting Problem

In a sense, we can say that the computational task that $Halt'$ (and in particular $Halt'(n_0)$) performs is:

> "If you ask me whether I will halt or go into an infinite loop, then I will give the wrong answer."

Since computers cannot give the wrong answer, $Halt'$ cannot exist and hence $Halt$ cannot exist.

Fixed Points in Logic

The Contrapositive of Lawvere's Theorem. Let $\mathbb{A}$ be a category with a terminal object and binary products. Let y be an object in the category and $\alpha\colon y \longrightarrow y$ be a morphism in the category. If there is an object a and a morphism $f\colon a \times a \longrightarrow y$, such that $g = \alpha \circ f \circ \Delta$ is representable by f, then α has a fixed point.

Fixed Points in Logic — Matrix Form

Second Input

First Input

$$
\begin{array}{c|cccccccc}
f & & & & & p & & \cdots \\
\hline
 & \alpha & & & & f(\quad, p) & & \cdots \\
 & & \alpha & & & f(\quad, p) & & \cdots \\
 & & & \alpha & & f(\quad, p) & & \cdots \\
 & & & & \alpha & f(\quad, p) & & \cdots \\
p & & & & & \alpha f(p,p) = f(p,p) & & \cdots \\
 & & & & & f(\quad, p) & \alpha & \cdots \\
\vdots & \vdots & & \vdots & & f(\quad, p) & & \ddots
\end{array}
$$

The crossing point is a fixed point.

Fixed Points in Logic

Now apply this theorem to logic. We use the contrapositive of Lawvere's Theorem to find fixed points in logic. First, some elementary logic. We are working in a system that can handle basic arithmetic. We will deal with logical formulas that accept at most one value which is a number.

$$\mathcal{A}(x), \mathcal{B}(x), \mathcal{C}(x), \ldots$$

A logical formula that accepts no value, sentences,

$$A, B, C, \ldots$$

We are interested in equivalence classes of these sets: two formulas are equivalent if they are provably logically equivalent. We will call the equivalence classes of predicates $Lind_1$ for the "Lindenbaum" classes of predicates. The equivalence classes of sentences will be denoted $Lind_0$.

Fixed Points in Logic

All logical formulas can be encoded as a natural number. We will write the natural number of a logical formula $\mathcal{A}(x)$ as $\ulcorner\mathcal{A}(x)\urcorner$ and the number of a sentence A is $\ulcorner A\urcorner$. Logical formulas about numbers will be able to evaluate logical formulas about numbers. It is these numbers that will help logical formulas be self-referential.

We are going to get fixed points of logical predicates. For every predicate, $\mathcal{E}(x)$, there is a way of constructing a fixed point which is a logical sentence C such that

$$\mathcal{E}(\ulcorner C\urcorner) \equiv C$$

The process that goes from a $\mathcal{E}(x)$ to C will be called a "fixed point machine." C is a logical sentence that says

"I have property $\mathcal{E}$."

With this fixed point machine we will find some of the most fascinating aspects of logic.

Fixed Points in Logic

Back to the contrapositive of the Lawvere Theorem.

- The category is $\mathbb{Set}$.
- The a of the theorem is the set of equivalence classes $Lind_1$.
- The y of the theorem is the set of equivalence classes $Lind_0$.
- There is a function $f : Lind_1 \times Lind_1 \longrightarrow Lind_0$ defined as follows:
$$f(\mathcal{A}(x), \mathcal{B}(x)) = \mathcal{B}(\ulcorner \mathcal{A}(x) \urcorner).$$
- The α of the theorem depends on some predicate $\mathcal{E}$ so we write it as $\alpha_{\mathcal{E}}$. The function $\alpha_{\mathcal{E}}$ applies the predicate to the number of a sentence A. It is a function $\alpha_{\mathcal{E}} : Lind_0 \longrightarrow Lind_0$ which is defined for the sentence A as
$$\alpha_{\mathcal{E}}(A) = \mathcal{E}(\ulcorner A \urcorner).$$

Fixed Points in Logic — Matrix Form

$$\begin{array}{cc|ccc}
 & & & & \textbf{Second Input} \\
 & f & \mathcal{A}_0(x) & \mathcal{A}_1(x) & \mathcal{A}_2(x) \\ \hline
\textbf{First Input} & \mathcal{A}_0(x) & \mathcal{E}(\mathcal{A}_0(\ulcorner \mathcal{A}_0(x) \urcorner)) & \mathcal{A}_1(\ulcorner \mathcal{A}_0(x) \urcorner) & \mathcal{A}_2(\ulcorner \mathcal{A}_0(x) \urcorner) \\
 & \mathcal{A}_1(x) & \mathcal{A}_0(\ulcorner \mathcal{A}_1(x) \urcorner) & \mathcal{E}(\mathcal{A}_1(\ulcorner \mathcal{A}_1(x) \urcorner)) & \mathcal{A}_2(\ulcorner \mathcal{A}_1(x) \urcorner) \\
 & \mathcal{A}_2(x) & \mathcal{A}_0(\ulcorner \mathcal{A}_2(x) \urcorner) & \mathcal{A}_1(\ulcorner \mathcal{A}_2(x) \urcorner) & \mathcal{E}(\mathcal{A}_2(\ulcorner \mathcal{A}_2(x) \urcorner)) \\
 & \vdots & \vdots & \vdots & \vdots \\
 & \mathcal{A}_p(x) & \mathcal{A}_0(\ulcorner \mathcal{A}_p(x) \urcorner) & \mathcal{A}_1(\ulcorner \mathcal{A}_p(x) \urcorner) & \mathcal{A}_2(\ulcorner \mathcal{A}_p(x) \urcorner) \\
 & \vdots & \vdots & & \vdots
\end{array}$$

Application: Gödel's Incompleteness Theorem

We use the fixed point machine to make interesting self-referential statements. Let Prov(x, y) be the two place predicate that is true when "y is the Gödel number of a proof of a statement whose Gödel number is x". Now we form the statement

$$\mathcal{E}(x) = (\forall y)\neg \mathrm{Prov}(y, x).$$

$$G \equiv \mathcal{E}(\ulcorner G \urcorner) = (\forall y)\neg \mathrm{Prov}(y, \ulcorner G \urcorner)$$

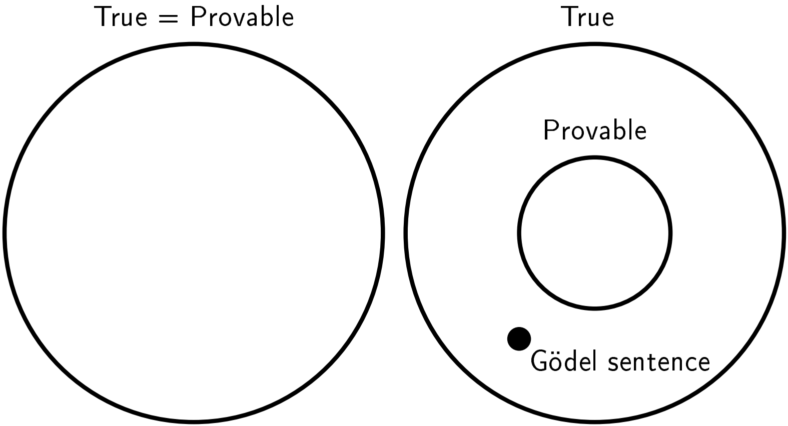
G is a logical statement that essentially says

"I am a statement for which any y is not a proof of me"

"I am unprovable."

If G was false then there would be a proof of G and hence there would be a proof of a false statement. In that case the system is not consistent. On the other hand, if G is true, then it essentially says that G is true but unprovable.

Application: Gödel's Incompleteness Theorem



What was believed before and after Gödel's Theorem.

Application: Tarski's Theorem

Alfred Tarski's theorem shows that a logical system cannot tell which of its predicates are true. Assume (wrongly) that there is some logical formula $\mathcal{T}(x)$ that accepts a number and tells if the statement is true. This formula will be true when "x is the Gödel number of a true statement in the theory". We can then use $\mathcal{T}(x)$ to form the statement

$$\mathcal{E}(x) = \neg\mathcal{T}(x)$$

This says that $\mathcal{E}(x)$ is true when $\mathcal{T}(x)$ is false. Now place $\mathcal{E}(x)$ into the fixed point machine. We will get a statement C such that

$$C \equiv \mathcal{E}(\ulcorner C \urcorner) = \neg\mathcal{T}(\ulcorner C \urcorner).$$

The logical sentence C essentially says

"I am false."

It is a logical version of the liar paradox.

Application: Parikh Sentences

Rohit Parikh used the fixed point machine to formulate some fascinating sentences that express properties about the length of its own proof. Consider the two-place predicate $\text{Prflen}(m, x)$ which is true if "there exists a proof of length m (in symbols) of a statement whose Gödel number is x."

$$\mathcal{E}_n(x) = \neg(\exists m < n \quad \text{Prflen}(m, x)).$$

$$C_n \equiv \mathcal{E}_n(\ulcorner C_n \urcorner) = \neg(\exists m < n \quad \text{Prflen}(m, \ulcorner C_n \urcorner)).$$

The logical sentence C_n essentially says

"I do not have a proof of length less than n."

As long as the logical system is consistent, C_n will be true and will not have a proof of length less than n. Parikh showed that although C_n does not have a short proof (you can make n as large as you want), there does exist a short proof of the fact that C_n is provable.

Epimenides and the Liar

Before we close this talk it pays to look at two famous paradoxes that are not exactly instances of Cantor's theorem but are close enough that they are easy to describe. The two examples are (i) the Epimenides paradox (the liar's paradox) and (ii) time travel paradoxes.

Chronologically, the granddaddy of all the self-referential paradoxes is the **Epimenides paradox**. Epimenides (6th or 7th century BC), a philosopher from Crete was a curmudgeon who did not like his neighbors in Crete. He is quoted as saying that "All Cretans are liars." The problem is that he is a Cretan. He is talking about himself and his statement. If his statement is true, then this very utterance is also a lie and hence is not true. On the other hand, if what he is saying is false, then he is not a liar and what he said is true. This seems to be a contradiction.

Epimenides and the Liar

There is a set of English sentences which we call $Sent$.
$f : Sent \times Sent \longrightarrow 2$. The function f is defined for sentences s and s' as

$$f(s, s') = \begin{cases} 1 & : s \text{ is negated by } s' \\ 0 & : s \text{ is not negated } s'. \end{cases}$$

$$\begin{tikzcd}
 & Sent \times Sent \arrow[r, "f"] & 2 \arrow[dr, "NOT"] & \\
Sent \arrow[ur, "\Delta"] \arrow[rrr, "g"'] & & & 2
\end{tikzcd}$$

g is the characteristic function of the subset of sentences that negate themselves. Till here, we have been mimicking the set-up of Lawvere's theorem. It is not clear what we would mean by talking about g being representable by f. What would it mean for a sentence S to represent a subset of sentences?

Time Travel Paradoxes

If time travel was possible, a time traveler might go back in time and shoot his bachelor grandfather, guaranteeing that the time traveler was never born. Homicidal behavior is not necessary to achieve such paradoxical results. The time traveler might just make sure that his parents never meet, or he might simply go back in time and make sure that he does not enter the time machine. These actions would imply a contradiction and hence cannot happen. The time traveler should not shoot his own grandfather (moral reasons notwithstanding) because if he shoots his own grandfather, he will not exist and will not be able to travel back in time to shoot his own grandfather. So by performing an action he is guaranteeing that the action cannot be performed. Here an event negatively affects itself. Since the physical universe does not permit contradictions, we must deny the assumption that time travel exists.

Time Travel Paradoxes

There is a collection of all physical events, *Events*.
$f : Events \times Events \longrightarrow 2$ is defined for two events e, and e' as

$$f(e, e') = \begin{cases} 1 & : \text{if } e \text{ is negated by } e' \\ 0 & : \text{if } e \text{ is not negated } e'. \end{cases}$$

$$\begin{tikzcd}
 & Events \times Events \arrow[r, "f"] & 2 \arrow[dr, "NOT"] & \\
Events \arrow[ur, "\Delta"] \arrow[rrr, "g"'] & & & 2
\end{tikzcd}$$

g is the characteristic function of those events that negate themselves. Such events cannot exist. Till here the pattern has been the same with Lawvere's theorem. However, we do not go on to talk about representing g by f. What would it mean for $f(\quad, e)$ to represent a subset of events?

Further Reading

- F. William Lawvere. Diagonal arguments and cartesian closed categories. In Category Theory, Homology Theory and their Applications, II pages 134-145. Springer, Berlin, 1969.
- F. William Lawvere and Stephen H. Schanuel. Conceptual mathematics, Second edition. Cambridge University Press, 2009. (Session 29).
- F. William Lawvere and Robert Rosebrugh. Sets for mathematics. Cambridge University Press, 2003. (Section 7.3).
- Rohit Parikh. Existence and feasibility in arithmetic. J. Symbolic Logic, 36:494-508, 1971.
- Noson S. Yanofsky. A universal approach to self-referential paradoxes, incompleteness and fixed points. Bull. Symbolic Logic, 9(3):362-386, 2003.

Further Reading

- Noson S. Yanofsky. The Outer Limits of Reason: What Science, Mathematics, and Logic Cannot Tell Us. MIT Press, Cambridge, MA, 2013.
- Noson S. Yanofsky. Resolving paradoxes. Philosophy Now, 016:10-12, 2015.
- Noson S. Yanofsky. Paradoxes, contradictions, and the limits of science. American Scientist, pages 166-173, May-June 2016.

The End

Thank You