

## **Virtual Logic — The Smullyan Machine**

*by Louis H. Kauffman<sup>1</sup>*

This is the seventh column in this series on “Virtual Logic”. In this column I will discuss an imaginary machine devised by the logician Raymond Smullyan (R. Smullyan, *Gödel’s Incompleteness Theorems*, Oxford Univ. Press 1992). Smullyan managed to compress the essence of Gödel’s theorem on the incompleteness of formal systems into the properties of a devilish machine.

This column consists in two parts. In the first part we find a story/satire about such a machine, with the Smullyan structure at its core. In this story, the protagonist is bent on detecting a flaw in the machine and he operates with strict two-valued logic. In such logic a statement is either true or false. Thus we call the statement “If unicorns can fly then all numbers are less than pi.” true because it is not definitely false. In general “A implies B” is taken to be false only if A is true and B is false. This is the one significant case where “A implies B” must be false. All other cases, such as A false and B true are taken to be true. This is the classical logical convention. It works quite well in its own domain, but it has its limits. One of these limits occurs when there is a gradation of qualities. For example in statements about tall and short the truth is relative to your idea of this discrimination. Another limit is in the realm of self-referential statements. Certainly the Liar Paradox — “This statement is false.” is neither true nor false in any timeless sense.

The satire goes on for four sections. Section 5 answers a problem from the last column and poses a new problem. Section 6 is a discussion of the issues raised by our satire.

Part of the protagonist’s dialogue with the machine is related to limitations of logic in the face of self-reference. As the reader will see, there are situations where the existence of a self-reference can limit actual behaviour (e.g. of a machine or a person) in the face of rules and classical logic. This is the part that is an encapsulation of the structure of Gödel’s Theorem. Near the end of the satire the protagonist runs into limits of logic in the face of self-reference. He discovers a useful sentence for politicians whose constituencies believe classical logic at all costs: “If this sentence is true, then I did not lie.”

---

[1] October 26 and November 15, 1998 Mittag-Leffler Institute Djursholm, Sweden

**I. The Machine**

No sooner had I taken Smullyan's book from the library than there came a knock on my door and a ring of the doorbell. I opened the door and found on the stoop, a neatly printed leaflet advertising just such a Machine with many promises, a money back guarantee and the usual admonition to act fast to insure the big savings in a once-only sale of this harbinger of truth.

Ever since I read that leaflet, I was unable to wait. I had to satisfy my curiosity and order the Smullyan Machine. The ad promised

"The Truth in a Box",  
 "A Machine that Never Lies",  
 "New Intelligence for the Millennium",  
 "No longer can you afford to be without the truth. Obtain the truth at an affordable price!"

A few days later I received a carton in the mail. It was labelled

"The Smullyan Machine — Nothing But The Truth!"

I opened the box and found the Machine. It is a beautifully configured wooden box with a button marked "PRINT", a button marked "TEST", a standard keyboard, a liquid crystal display, a slot for paper tape to emerge and a cord to attach to the wall socket. The instructions are embossed on a bronze plaque on the front of the machine (see panel).

I had a vague feeling of unease as I typed in my first sentence. I typed:

"Every even number greater than four is the sum of two odd prime numbers."

(I had hoped to finally learn the truth of this proposition, known as Goldbach's Conjecture.) Of course I did not type the quotation marks. They are just a device in collaboration with you, the reader, to distinguish the text input and output for the Machine.

Immediately, the Machine's printer hummed and a strip of tape came out on which read

"Every even number greater than four is the sum of two odd prime numbers."

Well, that's settled I said and ran to show my wife. Look dear, says I, Goldbach's conjecture is true after all!

How nice she says, and look dear, there is a message on the liquid crystal display. So I looked, and it said "So you think that Goldbach's conjecture is true do you? Press the PRINT button." I pressed the button. The machine printed

"There is a counterexample to Goldbach's conjecture, but this tape is too constricted for me to print it and my memory is too small to hold it."

## Smullyan Enterprises Inc.

**IMPORTANT INSTRUCTIONS — READ CAREFULLY**

1. Each time you press the button marked PRINT, the Machine will print a statement on the tape. The empty statement is a legally a statement and it will occupy a length of tape that is 7 centimeters long.
2. Printings produced by the machine are formed in English upper and lower case letters  $\{A,B, \dots, W,X,Y,Z. a,b,c,\dots,w,x,y,z\}$  plus Arabic numerals  $\{0,1,2,3,4,5,6,7,8,9\}$  plus the standard punctuation marks: the period  $\{.\}$ , the comma  $\{,\}$ , the round left and right parentheses  $\{,\}$ , the tilde  $\sim$ , the equals sign  $=$  and the empty character (1 centimeter in length)  $\cdot$ . These and only these characters will be used by the machine in printing its statements.
3. A printing by the machine consists in a finite character string. If  $X$  is a character string, then the following types of strings will be called “M-sentences” (Machine sentences).

(a)  $P(X)$  (b)  $\sim P(X)$  (c)  $PR(X)$  (d)  $\sim PR(X)$

The M-sentences each have interpretations in terms of the machine action. Interpretations are given in the form “A means B” where B is the interpretation of A. When we say “M can print Y” for a character string Y, we mean that it is possible for Y (exactly in that form) to be printed on the tape after the operator presses PRINT. Note that the fact that the machine prints XYZ where X, Y and Z are character strings, does NOT mean that it can print just Y alone (although this may be the case).

- (a)  $P(X)$  means that M can print X.  
 (b)  $\sim P(X)$  means that M cannot print X.  
 (c)  $PR(X)$  means that M can print  $X(X)$ .  
 (d)  $\sim PR(X)$  means that M cannot print  $X(X)$ .

An M-sentence is said to be true if and only if its interpretation is true.

4. **Any M-sentence Printed By The Smullyan Machine Is True.** Smullyan Enterprises GUARANTEES the truth of any M-sentence printed by a Smullyan Machine.. We will return your money in exchange for an invalid machine. *Caveat Emptor:* Smullyan enterprises makes no statement about the interpretation or validity of non M-sentence character strings that the machine can print. Enjoy your Machine!
5. **MACHINE OPERATION** (a) Plug the machine into a standard source of ac current — 100 to 240 V, 50-60 Hz. (b) If, without typing on the keyboard, the button PRINT is pressed, the Machine will print a sentence on the tape. (c) If the operator first presses the button TEST and then types a character string on the keyboard, then the typed string will appear on the liquid crystal display of the machine. Subsequent pressing of PRINT will either result in the printing of a character string, or the printing of a blank stretch of tape. (d) Sometimes the machine will put characters on the liquid crystal display after it prints on the tape (or produces blank tape). Smullyan enterprises takes no responsibility for interpretations of these liquid crystal displays. The liquid crystal displays are NOT instances of printing. When we say the machine prints a character string, this refers to the tape and only the tape. (e) Unplug the machine for convenient storage

Oh no! Says I. This Machine has already shown itself to be inconsistent. I want my money back.

Wait dear, she says. Why don't you read the instructions again. Maybe you are pushing the wrong buttons.

Very well says I, lets look again. So I looked and saw the clause about M-sentences. The machine only guarantees to print true M-sentences. So I wrote

"P(The moon is made of green cheese.)"

and pressed the print button.

Immediately the Machine printed

"The moon is made of green cheese."

and on the liquid crystal display it read

"And so are you."

## **II. Logical Testing**

The only way to find out about the machine was to work with M-Sentences. I was determined to trap the Machine in a lie and get my money back. So I tried the input

"P(This sentence is not printable by the Smullyan Machine.)".

I pressed PRINT and the Machine printed

"This sentence is not printable by the Smullyan Machine."

and the liquid crystal display read

"Press PRINT again."

I did and found

"P(This sentence is not printable by the Smullyan Machine.)"

This last was an M-sentence of the form P(X), and indeed true, since the machine had just printed X.

I was about to despair when my wife heard my groans and said: But dear, why don't you use symbolic logic. You are always explaining its virtues to me at 11:00 at night when we should be asleep!

### III. In Symbols

I followed her advice and thought: I want to catch this Machine by giving it an M-sentence that asserts its own unprintability. If I can make such an M-sentence then the Machine will be stuck, for if it prints it then the sentence will be false, and if it can not print it then the sentence will be true. My monkey-wrench M-sentence will have to be of the form

$$\sim P(X)$$

since this M-sentence asserts the unprintability of X.

But I want X to be identical with  $\sim P(X)$  as character strings. This is impossible since  $\sim P(X)$  is four characters longer than X!

It was after 11:00PM by now and I showed this impossibility to my wife. She said: Well dear I am very sleepy, but I think that you should read the Instructions again. There may be no way to make your monkey-wrench, but at least you will find out! Now please don't bother me again. I am going to bed!

So I looked again at the instructions. The M-sentences were stipulated to be of the form

$$\begin{aligned} &P(X), \\ &\sim P(X), \\ &PR(X), \\ &\sim PR(X). \end{aligned}$$

I had forgotten the R! R(X) refers to X(X) and so R(X) can have fewer characters than its referent, like any good name.

What about  $\sim PR(X)$ ? This sentence means that X(X) is unprintable by the Machine. Could X(X) be identical with  $\sim PR(X)$ ?

Of course! I will just let  $X = \sim PR$ . My monkey wrench would read

$$\sim PR(\sim PR).$$

This M-sentence asserts its own unprintability. I was ready.

### IV. Submission of the Monkey Wrench

I carefully typed  $\sim PR(\sim PR)$  and pressed the PRINT button. The machine whirred and printed blank tape. The liquid crystal display flickered and read

“I am a Gödelian Machine! The M-sentence that you have submitted to me asserts its own unprintability. Therefore, I cannot print it, for once printed, the M-sentence will be false. I never print a false M-Sentence. Since I cannot print it, your M-sentence is true! I am an Incomplete Machine, unable to print all true M-sentences. But you will not find me lying. It all depends on what you mean by lying. Ha! As a reward, please press the print button once more.”

I pressed the print button and the following text appeared on the tape.

**HOW TO PROVE ANYTHING!**

Do you wish to prove a proposition P?

No problem. Mathematicians have made too much of this tedious enterprise. Here is the all purpose proof.

THEOREM. P is true.

PROOF. Let S be the sentence:

"If this sentence is true then P is true."

Note that S is of the form:

"S implies P."

The truth value of a sentence in the form "A implies B." is true whenever A is false. Therefore S cannot be false, since this would imply that S is true. Hence S is true. Since S is true and S says that the truth of S implies the truth of P, it follows that P is true. This completes the proof of P.//

With the help of this proof-schema you can effortlessly prove any theorem in Mathematics and win any argument in Science or Economics. Note that the Theorem has immense Political Applications (PAPP) since it can be used at any time to prove

"I did not lie."

Just take S to be the sentence

"If this sentence is true, then I did not lie."

and proceed as above. Copies of this paragraph will be produced by the Smullyan Machine upon demand by entering ~ PR(~ PR) and pressing the PRINT button twice. Smullyan Enterprises suggests the distribution of this paragraph to members of all houses of government, particularly the Executive Branch. It is suggested that Mathematics professors handle this information with caution, as it may be hazardous to their health.

The Machine had me. I could not fault it, and now I had to ponder this paradox of provability. I went to bed.

**V. From Number Six**

At the end of the last column, I asked the reader to transcribe

"((a Implies b) and (b Implies c)) Implies (a Implies c)"

into the calculus of indications and verify that it always has the marked value in that calculus.

Recall that  $X \text{ OR } Y = XY$

$\text{NOT } X = \langle X \rangle$

$X \text{ AND } Y = \langle \langle X \rangle \langle Y \rangle \rangle$

$A \text{ IMPLIES } B = \langle A \rangle B$

$\text{TRUE} = \langle \rangle$

$\text{FALSE} = \langle \langle \rangle \rangle =$

in the calculus of indications.

Thus the sentence transcribes as

$$\langle \langle a \text{ IMPLIES } b \rangle \langle b \text{ IMPLIES } c \rangle \rangle \text{ IMPLIES } \langle a \rangle c.$$

and this fully transcribes to

$$Q = \langle \langle \langle \langle a \rangle b \rangle \langle \langle b \rangle c \rangle \rangle \rangle \langle a \rangle c.$$

Since  $\langle \langle X \rangle \rangle = X$  for any  $X$ , we see that

$$Q = \langle \langle a \rangle b \rangle \langle \langle b \rangle c \rangle \langle a \rangle c.$$

To see that  $Q$  is always true, use  $\langle XY \rangle Y = \langle X \rangle Y$ . then

$$Q = \langle b \rangle \langle \langle b \rangle c \rangle \langle a \rangle c = \langle b \rangle \langle c \rangle \langle a \rangle c = \langle b \rangle \langle a \rangle \langle c \rangle c.$$

Now use  $\langle c \rangle c = \langle \rangle$  and  $X \langle \rangle = \langle \rangle$  to conclude that

$$Q = \langle \rangle.$$

This completes the solution to the exercise.

**EXERCISE FOR NEXT TIME:** The Smullyan Machine's last printout was based on a self-referential sentence  $S$ . Transcribe this sentence into the calculus of indications and discuss its structure.

Until next time, enjoy yourself and remember,

$$\sim \text{PR}(\sim \text{PR}).$$

## VI. Notes

Here are some comments on our satire about the Smullyan Machine.

First of all, the Machine and its instructions embody a version of the distinction between language and meta-language that is used in studying mathematical formal systems. The formal language of the machine is the language consisting of strings

of symbols with the special notion of the “M-sentences” of the form  $P(X)$ ,  $\sim P(X)$ ,  $PR(X)$  and  $\sim PR(X)$ . The smallest possible “Smullyan Machine” would be one that only used the symbols  $(,)\sim, P$  and  $R$ .

In our satire the instructions that come with the Machine are in a meta-language that is explicitly intended to be descriptive of the machine’s operation, just like any manual for a piece of technology that one might use in the modern world. Machines have design and description that informs their structure. However, the Smullyan machine seems also to be self-describing in various ways. Even though the instructions say that there is no guarantee about the interpretation of the symbol strings that the machine makes (except for the M-sentences that it prints) the operator is led to interpret many of these as sensible English and/or mathematical communications. We have depicted him as beginning to sort out the distinction as the story goes on, but at the end the machine prints a whole paragraph of logical paradox (“how to prove anything”) that is under no guarantee by Smullyan Enterprises.

That paradox, by the way is an important variant of the Liar Paradox and we will come back to it in a later column.

Eventually the protagonist in our story sorts out the language/meta-language distinction that is created here, and he tries to make an M-sentence that is true but that the Machine cannot print. He tries to show that the Machine is incomplete! Note that there is given a notion of truth of M-sentences and behind this there is a notion of truth in a more general framework. This notion is concretely based on the way we determine truth in everyday life about definite actions in the world. Either the Machine does or does not print a given pattern of symbols. Either the operator does or does not read these symbols. It is implicitly assumed that “printing” is a very definite and clear operation. The reader of the printing can always determine just what has been printed.

Without these assumptions about precision, the game with the machine would not work in the very logical way that is intended. Again, this is just what we do in the world of technology, the world of mathematics and the world of games. We create a clearing where it is intended that the rules of logic should apply. In this intention statements are supposed to either be true or false, not both and not neither.

In our satire it is pointed out by the Instructions that this precision of logical structure is only guaranteed for the interpretations of the special M-sentences and for the reliability of the Machine’s rule-following. The rest may appear logical or illogical. There is no guarantee. Here I have placed the fellow who bought the machine in the same situation as any person confronted with the dialogue of another. There is no guarantee that the other’s behaviour will be logical, no guarantee that he will make sense at all.

What I take as your meaning is how I have interpreted it.

As scientists, cyberneticians and mathematicians we sort out communications by applying contexts and feelings. Some of these contexts are rule-governed like the M-sentences of the machine. But just as with the Machine’s incompleteness,

there is a built-in limitation to any formal language that we use to attempt to obtain precision. So we resort to dialogue, meta-dialogue, and our feelings to continue the communication and to reach newness and agreement.

Leibniz had a dream that all human disputes would someday be settled by a special language. Today we have many such languages and the machines to run them. It becomes more and more apparent that something beyond a special rule-driven formality is needed to reach real insight. As the formal systems grow in complexity, so does the need for understanding and sophistication in their operation. There is a difficulty and a real paradox here, because what is needed is a seamless continuity between language and meta-language. The rule-driven formal systems are based on making a distinction between language and meta-language. It is here that we (finally!) reach the concerns of second order cybernetics. It is through second order cybernetics that we understand that it is possible for the distinction between language and meta-language to flexibly come and go. We do this. We live in a language that can talk about itself, that emanates from its own self-description. We live beyond the formal systems that we create. We can even ask whether all that we do could be part of an all-embracing formal system (whose rules might even become known to us!). It is circular to be self-describing, but it is the way we are.

To return to the incompleteness of the Machine, our protagonist tries at first to make a machine sentence that says “I am unprintable.” by using the sentence-form “ $\sim P(X)$ ” and looking for an  $X$  that will solve the “equation”

$$\sim P(X) = X.$$

This would be a statement that is identical to the statement of its own unprintability. However, everything about this Machine is grounded in the concreteness of symbol strings that it prints. Identity is the character—by—character identity of strings of symbols. On this account the first attempt at solving such an equation is impossible since  $\sim P(X)$  has four more characters in its string than  $X$ , and so  $\sim P(X)$  can never be identical with  $X$ . Self-reference cannot be achieved in this way.

Concrete identity of symbol strings has in back of itself the idea of definite yes-no similarity. It is assumed that each string of symbols is recognisable and that the discrimination is possible. Again this is a notion of identity from one of the worlds that we create. We construct identity of this kind by demanding that it be so. And in the actual world (Think of the information of the witnesses in a court of law.) there is hard work needed to obtain even a fraction of this clarity. We ourselves do not seem to be just symbol strings and our identity is so fluid that we can hardly be said to be identical with ourselves (except, in the Western mode, by definition!).

Our protagonist solved the problem of self-reference by solving a different equation. He solved

$$\sim PR(X) = X(X).$$

He turned to this equation because, in the language of the Machine,  $R(X)$  refers to  $X(X)$ . Reference was built into the language of the machine in this very simple form. The “name”  $R(X)$  can have fewer characters (less complexity) than the  $X(X)$  to which it refers. As a result we can solve this equation uniquely with  $X \sim PR$  so that the M-sentence

$\sim PR(\sim PR)$

is interpreted as asserting its own unprintability.

Self-reference and incompleteness arise from this basic property of reference. The entity that does the referring can have smaller complexity than the entity to which it makes the reference.

I am more complex than my name. That is the game of reference. In the second order, we still play this game but realise that it is a game. The MAKING of the name IS the name. The making of the name requires the full complexity (and simplicity) of the whole. I am identical with my full (and partly unspoken) name.

We become the language that we are.



