METAMAGICAL THEMAS:

Questing for the Essence of Mind and Pattern

DOUGLAS R. HOFSTADTER

1

On Self-Referential Sentences

January, 1981

I never expected to be writing a column for *Scientific American.* I remember once, years ago, wishing I were in Martin Gardner's shoes. It seemed exciting to be able to plunge into almost any topic one liked and to say amusing and instructive things about it to a large, well-educated, and receptive audience. The notion of doing such a thing seemed ideal, even dreamlike. Over the next several years, by a series of total coincidences (which turned out to be not so total), I met one after another of Martin's friends. First it was Ray Hyman, a psychologist who studies deception. He introduced me to the magician Jerry Andrus. Then I met the statistician and magician Persi Diaconis and the computer wizard Bill Gosper. Then came Scott Kim, and soon afterward, the mathematician Benoît Mandelbrot. All of a sudden, the world seemed to be orbiting Martin Gardner. He was at the hub of a magic circle, people with exciting, novel, often offbeat ideas, people with many-dimensional imaginations. Sometimes I felt overawed by the whole remarkable bunch.

One day, five or so years ago, I had the pleasure of spending several hours with Martin in his house, discussing many topics, mathematical and otherwise. It was an enlightening experience for me, and it gave me a new view into the mind of someone who had contributed so much to my own mathematical education. Perhaps the most striking thing about Martin to me was his natural simplicity. I had been told that he is an adroit magician. This I found hard to believe, because one does not usually imagine someone so straightforward pulling the wool over anyone's eyes. However, I did not see him do any magic tricks. I simply saw his vast knowledge and love of ideas spread out before me, without the slightest trace of pride or pretense. The Gardners—Martin and his wife Charlotte—entertained me for the day. We ate lunch in the kitchen of their cozy three-story house. It pleased me somehow to see that there was practically no trace of mathematics or games or tricks in their simple but charming living room.

After lunch—sandwiches that Martin and I made while standing by the kitchen sink—we climbed the two flights of stairs to Martin's hideaway. With his old typewriter and all kinds of curious jottings in an ancient filing cabinet and his legendary library of three-by-five cards, he reminded me of an old-time journalist, not of the center of a constellation of mathematical eccentrics and game addicts, to say nothing of magicians, anti-occultists, and of course the thousands of readers of his column.

Occasionally we were interrupted by the tinkling of a bell attached to a string that led down the stairs to the kitchen, where Charlotte could pull it to get his attention. A couple of phone calls came, one from the logician and magician Raymond Smullyan, someone whose name I had known for a long time, but who I had no idea belonged to this charmed circle. Smullyan was calling to chat about a book he was writing on Taoism, of all things! For a logician to be writing about what seemed to me to be the most anti-logical of human activities sounded wonderfully paradoxical. (In fact, his book *The Tao Is Silent* is delightful and remarkable.) All in all, it was a most enjoyable day.

Martin's act will be a hard one to follow. But I will not be trying to be another Martin Gardner. I have my own interests, and they are different from Martin's, although we have much in common. To express my debt to Martin and to symbolize the heritage of his column, I have kept his title "Mathematical Games" in the form of an anagram: "Metamagical Themas".

What does "metamagical" mean? To me, it means "going one level beyond magic". There is an ambiguity here: on the one hand, the word might mean "ultramagical"—magic of a higher order—yet on the other hand, the magical thing about magic is that what lies behind it is always *non*-magical. That's metamagic for you! It reflects the familiar but powerful adage "Truth is stranger than fiction." So my "Metamagical Themas" will, in Gardnerian fashion, attempt to show that magic often lurks where few suspect it, and, by the opposite token, that magic seldom lurks where many suspect it.

*　*　*

In his July, 1979 column, Martin wrote a very warm review of my book *Gödel, Escher, Bach: an Eternal Golden Braid.* He began the review with a short quotation from my book. If I had been asked to guess what single sentence he would quote, I would never have been able to predict his choice. He chose the sentence "This sentence no verb." It is a catchy sentence, I admit, but something about seeing it again bothered me. I remembered how I had written it one day a few years earlier, attempting to come up with a new variation on an old theme, but even at the time it had not seemed as striking as I had hoped it would. After seeing it chosen as the symbol of my book, I felt challenged. I said to myself that surely there must be much cleverer types of self-referential sentence. And so one day I wrote down quite a pile of self-referential sentences and showed them to friends, which began a mild craze among a small group of us. In this column, I will present a selection of what I consider to be the cream of that crop.

Before going further, I should explain the term "self-reference". Self-reference is ubiquitous. It happens every time anyone says "I" or "me" or "word" or "speak" or "mouth". It happens every time a newspaper prints a story about reporters, every time someone writes a book about writing, designs a book about book design, makes a movie about movies, or writes an article about self-reference. Many systems have the capability to represent or refer to themselves somehow, to designate themselves (or elements of themselves) within the system of their own symbolism. Whenever this happens, it is an instance of self-reference.

Self-reference is often erroneously taken to be synonymous with paradox. This notion probably stems from the most famous example of a self-referential sentence, the Epimenides paradox. Epimenides the Cretan said, "All Cretans are liars." I suppose no one today knows whether he said it in ignorance of its self-undermining quality or for that very reason. In any case, two of its relatives, the sentences "I am lying" and "This sentence is false", have come to be known as the *Epimenides paradox* or the *liar paradox.* Both sentences are absolutely self-destructive little gems and have given self-reference a bad name down through the centuries. When people speak of the evils of self-reference, they are certainly overlooking the fact that not every use of the pronoun "I" leads to paradox.

* * *

Let us use the Epimenides paradox as our jumping-off point into this fascinating land. There are many variations on the theme of a sentence that somehow undermines itself. Consider these two:

This sentence claims to be an Epimenides paradox, but it is lying.

This sentence contradicts itself—or rather—well, no, actually it doesn't!

What should you do when told, "Disobey this command"? In the following sentence, the Epimenides quality jumps out only after a moment of thought: "This sentence contains exactly threee erors." There is a delightful backlash effect here.

Kurt Gödel's famous Incompleteness Theorem in metamathematics can be thought of as arising from his attempt to replicate as closely as possible the liar paradox in purely mathematical terms. With marvelous ingenuity, he was able to show that in any mathematically powerful axiomatic system *S* it is possible to express a close cousin to the liar paradox, namely, "This formula is unprovable within axiomatic system *S.*"

In actuality, the Gödel construction yields a mathematical formula, not an English sentence; I have translated the formula back into English to show what he concocted. However, astute readers may have noticed that, strictly speaking, the phrase "this formula" has no referent, since when a *formula* is translated into an English *sentence*, that sentence is no longer a formula!

If one pursues this idea, one finds that it leads into a vast space. Hence the following brief digression on the preservation of self-reference across language boundaries. How should one translate the French sentence *Cette phrase en français est difficile à traduire en anglais*? Even if you do not know French, you will see the problem by reading a literal translation: "This sentence in French is difficult to translate into English." The problem is: To what does the subject ("This sentence in French") refer? If it refers to the sentence it is part of (which is not in French), then the subject is self-contradictory, making the sentence false (whereas the French original was true and harmless); but if it refers to the French sentence, then the meaning of "this" is strained. Either way, something disquieting has happened, and I should point out that it would be just as disquieting, although in a different way, to translate it as: "This sentence in English is difficult to translate into French." Surely you have seen Hollywood movies set in France, in which all the dialogue, except for an occasional *Bonjour* or similar phase, is in English. What happens when Cardinal Richelieu wants to congratulate the German baron for his excellent command of French? I suppose the most elegant solution is for him to say, "You have an excellent command of our language, *mon cher baron*", and leave it at that.

* * *

But let us undigress and return to the Gödelian formula and focus on its meaning. Notice that the concept of *falsity* (in the liar paradox) has been replaced by the more rigorously understood concept of *provability*. The logician Alfred Tarski pointed out that it is in principle impossible to translate the liar paradox exactly into any rigorous mathematical language, because if it were possible, mathematics would contain a genuine paradox —a statement both true and false—and would come tumbling down.

Gödel's statement, on the other hand, is not paradoxical, though it constitutes a hair-raisingly close approach to paradox. It turns out to be true, and for this reason, it is unprovable in the given axiomatic system. The revelation of Gödel's work is that in *any* mathematically powerful and consistent axiomatic system, an endless series of true but unprovable formulas can be constructed by the technique of self-reference, revealing that somehow the full power of human mathematical reasoning eludes capture in the cage of rigor.

In a discussion of Gödel's proof, the philosopher Willard Van Orman Quine invented the following way of explaining how self-reference could be achieved in the rather sparse formal language Gödel was employing. Quine's construction yields a new way of expressing the liar paradox. It is this:

"yields falsehood, when appended to its quotation." yields falsehood, when appended to its quotation.

This sentence describes a way of constructing a certain typographical entity —namely, a phrase appended to a copy of itself in quotes. When you carry out the construction, however, you see that the end product is the sentence itself—or a perfect copy of it. (There is a resemblance here to the way self-replication is carried out in the living cell.) The sentence asserts the falsity of the constructed typographical entity, namely itself (or an indistinguishable copy of itself). Thus we have a less compact but more explicit version of the Epimenides paradox.

It seems that all paradoxes involve, in one way or another, self-reference, whether it is achieved directly or indirectly. And since the credit for the discovery—or creation—of self-reference goes to Epimenides the Cretan, we might say: "Behind every successful paradox there lies a Cretan."

On the basis of Quine's clever construction we can create a self-referential question:

What is it like to be asked,
"What is it like to be asked, self-embedded in quotes after its comma?"
self-embedded in quotes after its comma?

Here again, you are invited to construct a typographical entity that turns out, when the appropriate operations have been performed, to be identical with the set of instructions. This self-referential question suggests the following puzzle: What is a question that can serve as its own answer? Readers might enjoy looking for various solutions to it.

* * *

When a word is used to *refer* to something, it is said to be being *used.* When a word is *quoted,* though, so that one is examining it for its surface aspects (typographical, phonetic, etc.), it is said to be being *mentioned.* The following sentences are based on this famous use-mention distinction:

You can't have your use and mention it too.

You can't have "your cake" and spell it "too".

"Playing with the use-mention distinction" isn't "everything in life, you know".

In order to make sense of "this sentence", you will have to ignore the quotes in "it".

This is a sentence with "onions", "lettuce", "tomato", and "a side of fries to go".

This is a hamburger with vowels, consonants, commas, and a period at the end.

The last two are humorous flip sides of the same idea. Here are two rather extreme examples of self-referential use-mention play:

> Let us make a new convention: that anything enclosed in *triple* quotes—for example, '"No, I have decided to change my mind; when the triple quotes close, just skip directly to the period and ignore everything up to it"'—is not even to be read (much less paid attention to or obeyed).
>
> A ceux qui ne comprennent pas l'anglais, la phrase citée ci-dessous ne dit rien: "For those who know no French, the French sentence that introduced this quoted sentence has no meaning."

The bilingual example may be more effective if you know only one of the two languages involved.

Finally, consider this use-mention anomaly: "i should begin with a capital letter." This is a sentence referring to itself by the pronoun "I", a bit mauled, instead of through a pointing-phrase such as "this sentence"; such a sentence would seem to be arrogantly proclaiming itself to be an animate agent. Another example would be "I am not the person who wrote me." Notice how easily we understand this curious nonstandard use of "I". It seems quite natural to read the sentence this way, even though in nearly all situations we have learned to unconsciously create a mental model of some person—the sentence's speaker or writer—to whom we attribute a desire to communicate some idea. Here we take the "I" in a new way. How come? What kinds of cues in a sentence make us recognize that when the word "I" appears, we are supposed to think not about the author of the sentence but about the sentence itself?

* * *

Many simplified treatments of Gödel's work give as the English translation of his famous formula the following: "I am not provable in axiomatic system *S*." The self-reference that is accomplished with such sly trickery in the formal system is finessed into the deceptively simple English word "I", and we can—in fact, we automatically do—take the sentence to be talking about itself. Yet it is hard for us to hear the following sentence as talking about itself: "I *already* took the garbage out, honey."

The ambiguous referring possibilities of the first-person pronoun are a source of many interesting self-referential sentences. Consider these:

I am not the subject of this sentence.

I am jealous of the first word in this sentence.

Well, how about that—this sentence is about me!

I am simultaneously writing and being written.

This raises a whole new set of possibilities. Couldn't "I" stand for the writing instrument ("I am not a pen"), the language ("I come from Indo-European roots"), the paper ("Cut me out, twist me, and glue me to form a Möbius strip, please")? One of the most involved possibilities is that "I" stands not for the physical tokens we perceive before us but for some more ethereal and intangible essence, perhaps the *meaning* of the sentence. But then, what is meaning? The next examples explore that idea:

I am the meaning of this sentence.

I am the thought you are now thinking.

I am thinking about myself right now.

I am the set of neural firings taking place in your brain as you read the set of letters in this sentence and think about me.

This inert sentence is my body, but my soul is alive, dancing in the sparks of your brain.

The philosophical problem of the connections among Platonic ideas, mental activity, physiological brain activity, and the external symbols that trigger them is vividly raised by these disturbing sentences.

This issue is highlighted in the self-referential question, "Do you think anybody has ever had *precisely this thought* before?" To answer the question, one would have to know whether or not two different brains can *ever* have precisely the same thought (as two different computers can run precisely the same program). An illustration of this possibility may be found in Figure 24-2. I have often wondered: Can *one* brain have the same thought more than once? Is a thought something Platonic, something whose essence exists independently of the brain it is occurring in? If the answer is "Yes, thoughts are brain-independent", then the answer to the self-referential question would also be yes. If it is not, then no one could ever have had the same thought before—not even the person thinking it!

Certain self-referential sentences involve a curious kind of communication between the sentence and its human friends:

> You are under my control because I am choosing exactly what words you are made out of, and in what order.
>
> No, *you* are under *my* control because you will read until you have reached the end of me.
>
> Hey, down there—are you the sentence I am writing, or the sentence I am reading?
>
> And you up there—are you the person writing me, or the person reading me?
>
> You and I, alas, can have only one-way communication, for you are a person and I, a mere sentence.
>
> As long as you are not reading me, the fourth word of this sentence has no referent.
>
> The reader of this sentence exists only while reading me.

Now *that* is a rather frightening thought! And yet, by its own peculiar logic, it is certainly true.

> Hey, out there—is that *you* reading me, or is it someone else?
>
> Say, haven't you written me somewhere else before?
>
> Say, haven't I written you somewhere else before?

The first of the three sentences above addresses its reader; the second addresses its author. In the last one, an author addresses a sentence.

Many sentences include words whose referents are hard to figure out because of their ambiguity—possibly accidental, possibly deliberate:

> Thit sentence is not self-referential because "thit" is not a word.
>
> No language can express every thought unambiguously, least of all this one.

In the Escher-inspired Figure 1-1, visual and verbal ambiguity are simultaneously exploited.

* * *

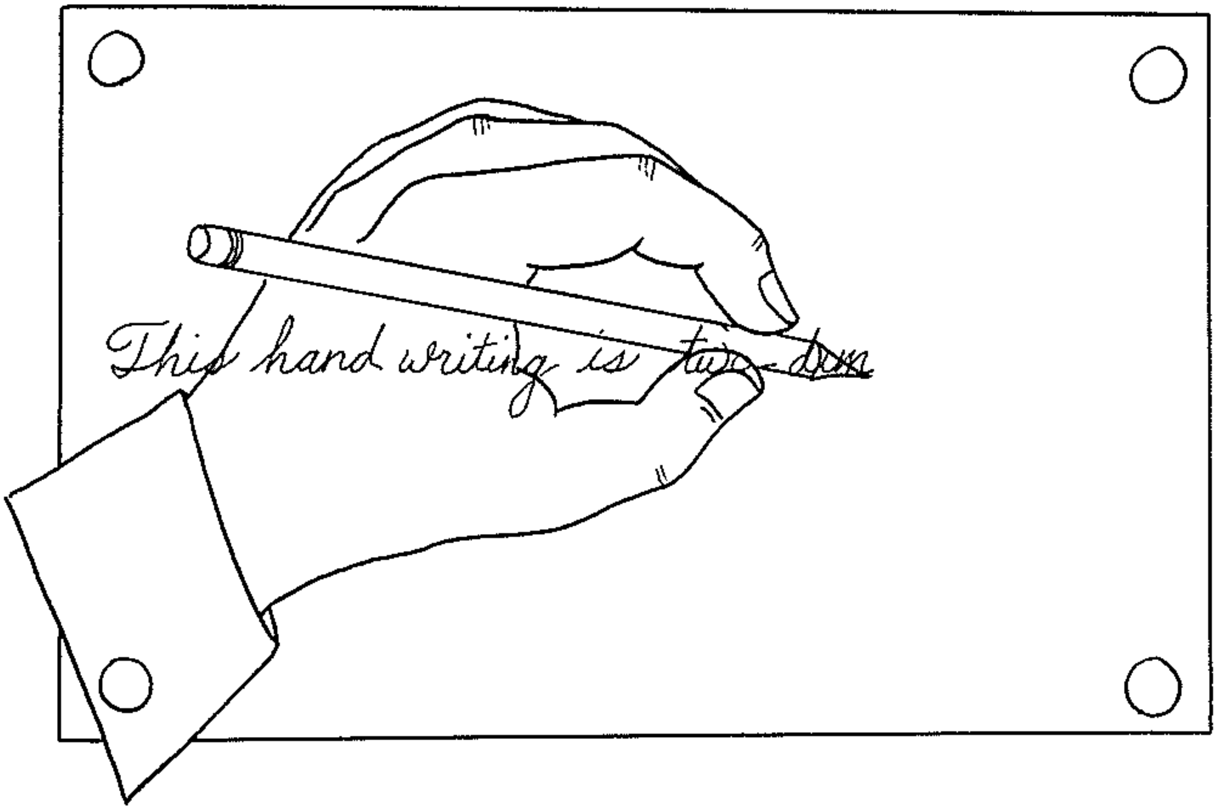

FIGURE 1-1. *Ambiguity: What is being described—the hand, or the writing? [Drawing by David Moser, after M. C. Escher.]*

Let us turn to a most interesting category, namely sentences that deal with the languages they are in, once were in, or might have been in:

When you are not looking at it, this sentence is in Spanish.

I had to translate this sentence into English because I could not read the original Sanskrit.

The sentence now before your eyes spent a month in Hungarian last year and was only recently translated back into English.

If this sentence were in Chinese, it would say something else.

.siht ekil ti gnidaer eb d'uoy ,werbeH ni erew ecnetnes siht fI

The last two sentences are examples of *counterfactual conditionals.* Such a sentence postulates in its first clause (the *antecedent*) some contrary-to-fact situation (sometimes called a "possible world") and extrapolates in its second clause (the *consequent*) some consequence of it. This type of sentence opens up a rich domain for self-reference. Some of the more intriguing self-referential counterfactual conditionals I have seen are the following:

If this sentence didn't exist, somebody would have invented it.

If I had finished this sentence,

If there were no counterfactuals, this sentence would not be paradoxical.

If wishes were horses, the antecedent of this conditional would be true.

If this sentence were false, beggars would ride.

What would this sentence be like if it were not self-referential?

What would this sentence be like if π were 3?

Let us ponder the last of these (invented by Scott Kim) for a moment. In a world where π actually *did* have the value 3, you wouldn't ask about how things *would* be if π were 3. Instead, you might muse "if π were 2" or "if π *weren't* 3". So one's first answer to the question might be this: "What would this sentence be like if π weren't 3?". But there is a problem. The referent of "this sentence" has now changed identity. So is it fair to say that the second sentence is an answer to the first? It is a little like a woman who muses, "What would I be doing now if I had had different genes?" The problem is that she would not be herself; she would be someone else, perhaps the little boy across the street, playing in his sandbox. Personal pronouns like "I" cannot quite keep up with such strange hypothetical world-shifts.

But getting back to Scott Kim's counterfactual, I should point out that there is an even more serious problem with it than so far mentioned. Changing the value of π is, to put it mildly, a radical change in mathematics, and presumably you cannot change mathematics radically without having radically changed the fabric of the universe within which we live. So it is quite doubtful that any of the concepts in the sentence would make any sense if π were 3 (including the concepts of "π", "3", and so on).

Here are two more counterfactual conditionals to put in your pipe and smoke:

If the subjunctive was no longer used in English, this sentence would be grammatical.

This sentence would be seven words long if it were six words shorter.

These two lovely examples, invented by Ann Trail (who is also responsible for quite a few others in this column), bring us around to sentences that comment on their own form. Such sentences are quite distinct from ones that comment on their own content (such as the liar paradox, or the sentence that says "This sentence is not about itself, but about whether it is about itself."). It is easy to make up a sentence that refers to its own form, but it is hard to make up an *interesting* one. Here are a few more quite good ones:

because I didn't think of a good beginning for it.

This sentence was in the past tense.

This sentence has contains two verbs.

This sentence contains one numeral 2 many.

a preposition. This sentence ends in

In the time it takes you to read this sentence, eighty-six letters could have been processed by your brain.

* * *

David Moser, a composer and writer, is a delector and creator of self-reference and frame-breaking of all kinds. He has even written a story in which every sentence is self-referential (it is included in Chapter 2). It might seem unlikely that in such a limited domain, individual styles could arise and flourish, but David has developed a self-referential style quite his own. As a mutual friend (or was it David himself?) wittily observed, "If David Moser had thought up this sentence, it would have been funnier." Many Moser creations have been used above. Some further Moserian delights are these:

This is not a complete. Sentence. This either.

This sentence contains only one nonstandard English flutzpah.

This gubblick contains many nonsklarkish English flutzpahs, but the overall pluggandisp can be glorked from context.

This sentence has cabbage six words.

In my opinion, it took quite a bit of flutzpah to just throw in a random word so that there *would* be cabbage six words in the sentence. That idea inspired the following: "This sentence has five (5) words." A few more miscellaneous Moserian gems follow:

This is to be or actually not two sentences to be, that is the question, combined.

It feels *sooo* good to have your eyes run over my curves and serifs.

This sentence is a !!!! premature punctuator

Sentences that talk about their own punctuation, as the preceding one does, can be quite amusing. Here are two more:

This sentence, though not interrogative, nevertheless ends in a question mark?

This sentence has no punctuation semicolon the others do period

Another ingenious inventor of self-referential sentences is Donald Byrd, several of whose sentences have already been used above. Don too has his own very characteristic way of playing with self-reference. Two of his sentences follow:

This hear sentence do'nt know Inglish purty good.

If you meet this sentence on the board, erase it.

The latter, via its form, alludes to the Buddhist saying "If you meet the Buddha on the road, kill him."

Allusion through similarity of form is, I have discovered, a marvelously rich vein of self-reference, but unfortunately this article is too short to contain a full proof of that discovery. I shall explicitly discuss only two examples. The first is "This sentence verbs good, like a sentence should." Its primary allusion is to the famous slogan "Winston tastes good, like a cigarette should", and its secondary allusion is to "This sentence no verb." The other example involves the following lovely self-referential remark, once made by the composer John Cage: "I have nothing to say, and I am saying it." This allows the following rather subtle twist to be made: "I have nothing to allude to, and I am alluding to it."

* * *

Some of the best self-referential sentences are short but sweet, relying for their effect on secondary interpretations of idiomatic expressions or well-known catch phrases. Here are five of my favorites, which seem to defy other types of categorization:

Do you read me?

This point is well taken.

You may quote me.

I am going two-level with you.

I have been sentenced to death.

In some of these, even sophisticated non-native speakers would very likely miss what's going on.

Surely no article on self-reference would be complete without including a few good examples of self-fulfilling prophecy. Here are a few:

This prophecy will come true.

This sentence will end before you can say "Jack Rob

Surely no article on self-reference would be complete without including a few good examples of self-fulfilling prophecy.

Does this sentence remind you of Agatha Christie?

That last sentence—one of Ann Trail's—is intriguing. Clearly it has nothing to do with Agatha Christie, nor is it in her style, and so the answer ought to be no. Yet I'll be darned if I can read it without being reminded of Agatha Christie! (And what is even stranger is that I don't know the first thing about Agatha Christie!)

In closing, I cannot resist the touching plea of the following Byrdian sentence:

Please, oh please, publish me in your collection of self-referential sentences!

Post Scriptum.

This first column of mine triggered a big wave of correspondence, some of which is presented in the next chapter. Most of the correspondence was light-hearted, but there were a number of serious letters that intrigued me. Here is a repartee that appeared in the pages of *Scientific American* a few months later.

> The kind of structural analysis engaged in, and the resulting questions raised by, Douglas Hofstadter in his amusing and intriguing article concerning self-referential sentences need not lead inevitably to bafflement of the reader.

> Help is at hand from the "laggard science" psychology, but only from that carefully defined quarter of psychology known as behavior analysis, which was progenerated by the famous Harvard psychologist B. F. Skinner almost 50 years ago.
>
> In examining the implications of linguistic analyses such as Hofstadter's for the serious student of verbal behavior, Skinner comments in his book *About Behaviorism* (pages 98—99) as follows:
>
>> Perhaps there is no harm in playing with sentences in this way or in analyzing the kinds of transformations which do or do not make sentences acceptable to the ordinary reader, but it is still a waste of time, particularly when the sentences thus generated could not have been emitted as verbal behavior. A classical example is a paradox, such as 'This sentence is false', which appears to be true if false and false if true. The important thing to consider is that no one could ever have emitted the sentence as verbal behavior. A sentence must be in existence before a speaker can say, 'This sentence is false', and the response itself will not serve, since it did not exist until it was emitted. What the logician or linguist calls a sentence is not necessarily verbal behavior in any sense which calls for a behavioral analysis.
>
> As Skinner pointed out long ago, verbal behavior results from contingencies of reinforcement arranged by verbal communities, and it is these contingencies that must be analyzed if we are to identify the variables that control verbal behavior. Until we grasp the full import of Skinner's position, which goes beyond structure to answer *why* we behave as we do verbally or nonverbally, we shall continue to fall back on prescientific formulations that are about as useful in understanding these phenomena as Hofstadter's quaint metaphorical speculation: "Such a sentence would seem to be arrogantly proclaiming itself to be an animate agent."
>
> George Brabner
> College of Education
> University of Delaware

I felt compelled to reply to Professor Brabner's interesting views about these matters, and so here is what I wrote:

> I assume that the quote from B. F. Skinner reflects Professor Brabner's own sentiments about the likelihood of self-referential utterances. I am always baffled by people who doubt the likelihood of self-reference and paradox. Verbal behavior comes in many flavors. Humor, particularly self-referential humor, is one of the most pervasive flavors of verbal behavior in this century. One has only to watch the Muppets or Monty Python on television to see dense and intricate webs of self-reference. Even advertisements excel in self-reference.
>
> In art, René Magritte, Pablo Picasso, M. C. Escher, John Cage, and dozens of others have played with the level-distinction between *that which represents* and *that which is represented.* The "artistic behavior" that results includes much self-reference and many confusing and sometimes exhilaratingly paradoxical

tangles. Would Professor Brabner say that no one could ever have "emitted" such works as "artistic behavior"? Where is the borderline?

Ordinary language, as I pointed out in my column, is filled with self-reference, usually a little milder-seeming than the very sharply pointed paradoxes that Professor Brabner objects to. "Mouth", "word", and so on are all self-referential. Language is inherently filled with the potential of sharp turns on which it may snag itself.

Many scholarly papers begin with a sentence about "the purpose of this paper". Newspapers report on their own activities, conceivably on their own inaccuracies. People say, "I'm tired of this conversation." Arguments evolve about arguments, and can get confusingly and painfully self-involved. Has Professor Brabner never thought of "verbal behavior" in this light? It is likely that in hunting woolly mammoths, no one found it extraordinarily likely to shout, "This sentence is false!" However, civilization has come a long way since those days, and the primitive purposes of language have by now been almost buried under an avalanche of more complex purposes.

Part of human nature is to be introspective, to probe. Part of our "verbal behavior" deliberately, often playfully, explores the boundaries between conceptual levels of systems. All of this has its root in the struggle to survive, in the fact that our brains have become so flexible that much of their time is spent in dealing with their own activities, consciously or unconsciously. It is simply a consequence of representational power—as Kurt Gödel showed—that systems of increasing complexity become increasingly self-referential.

It is quite possible for people filled with self-doubt to recognize this trait in themselves, and to begin to doubt their self-doubt itself. Such psychological dilemmas are at the heart of some current theories of therapy. Gregory Bateson's "double bind", Victor Frankl's "logotherapy", and Paul Watzlawick's therapeutic ideas are all based on level-crossing paradoxes that crop up in real life. Indeed, psychotherapy is itself based completely on the idea of a "twisted system of self"—a self that wants to reach inward and change some presumably wrong part of itself.

We human beings are the only species to have evolved humor, art, language, tangled psychological problems, even an awareness of our own mortality. Self-reference—even of the sharp Epimenides type—is connected to profound aspects of life. Would Professor Brabner argue that suicide is not conceivable human behavior?

Finally, just suppose Professors Skinner and Brabner are right, and no one ever says exactly "This sentence is false." Would this mean that study of such sentences is a waste of time? Still not. Physicists study ideal gases because they represent a distillation of the most significant principles of the behavior of real gases. Similarly, the Epimenides paradox is an "ideal paradox"—one that cuts crisply to the heart of the matter. It has opened up vast domains in logic, pure science, philosophy, and other disciplines, and will continue to do so despite the skepticism of behaviorists.

It is a curious coincidence that the only other reply to my article that was printed in the "Letters" column of *Scientific American* also came from the University of Delaware. Here it is:

> I hope that you do not receive any correspondence concerning Douglas R. Hofstadter's article on self-reference. I should like to inform your readers that many years of study on this problem have convinced me no conclusion whatsoever can be drawn from it that would stand up to a moment's scrutiny. There is no excuse for *Scientific American* to publish letters from those cranks who consider such matters to be worthy of even the slightest notice.
>
> A. J. Dale
> Department of Philosophy
> University of Delaware

I replied as follows:

> Many years of reading such letters have convinced me that no reply whatsoever can be given to them that would stand up to a moment's scrutiny. There is no excuse for publishing responses to those cranks who send them.

After these two exchanges had appeared in print, a number of people remarked to me that they'd read the two letters from Delaware that had attacked me, and had enjoyed my responses. Two? I guess it wasn't so obvious that Dale's letter was completely tongue-in-cheek. In fact, that was its point.

* * *

Two other letters stand out sharply in my memory. One was from an individual who signed himself (I presume it is a male) as "Mr Flash qFiasco". Mr Flash insisted that a sentence cannot *say* what it *shows.* The former concerns only its *content,* which is supposedly independent of how it manifests itself in print, while the latter is a property exclusively of its *form,* that is, of the physical sentence only when it is in print. This distinction sounds crystal-clear at first, but in reality it is mud-blurry. Here is some of what Flash wrote me:

> For a sentence to attempt to say what it shows is to commit an error of logical types. It seems to be putting a round peg into a square hole, whereas it is instead putting a round peg into something which is not a hole at all, square or otherwise. This is a category mismatch, not a paradox. It is like throwing the recipe in with the flour and butter and eggs. The source of the equivocation is an illegitimate use of the term 'this'. 'This' can point to virtually anything, but 'this' cannot point to itself. If you stick out your index finger, you can point to virtually anything; and by curling it you can even point to the pointing finger; but you cannot point to *pointing.* Pointing is of a higher logical type than the thing which is doing the pointing. Similarly, the referent of 'this sentence' can be virtually anything but that sentence. Sentences of the form exemplified by 'This sentence no verb' and 'This sentence has a verb' are not well-formed: they commit fallacies of logical type equivocation. Thus their self-referential character is not genuine and they present no problem as paradoxes.

There will always be people around who will object in this manner, and in the Brabnerian manner. Such people think it is possible to draw a sharp line between attributes of a printed sentence that can be considered part of its *form* (*e.g.*, the typeface it is printed in, the number of words it contains, and so on), and attributes that can be considered part of its *content* (*i.e.*, the things and events and relationships that it refers to).

Now, I am used to thinking about language in terms of how to get a machine to deal with it, since I look at the human brain as a very complex machine that can handle language (and many other things as well). Machines, in trying to make sense of sentences, have access to nothing more than the *form* of such sentences. The *content,* if it is to be accessible to a machine, has to be derived, extracted, constructed, or created somehow from the sentence's physical structure, together with other knowledge and programs already available to the machine.

When very simple processing is used to operate on a sentence, it is convenient to label the information thus obtained "syntactic". For instance, it is clearly a syntactic fact about "This sentence no verb." that it contains six vowels. The vowel-consonant distinction is obviously a typographical one, and typographical facts are considered superficial and syntactic. But there is a problem here. With different *depths of processing,* aspects of different degrees of "semanticity" may be detected.

Consider, for example, the sentence "Mary was sick yesterday." Let's call it *Sentence M.* Listed below are the results of seven different degrees of processing of Sentence M by a hypothetical machine, using increasingly sophisticated programs and increasingly large knowledge bases. You should think of them as being English translations, for your convenience, of computational structures inside the machine that it can act on and use fluently.

1. Sentence M contains twenty letters.
2. Sentence M contains four English words.
3. Sentence M contains one proper noun, one verb, one adjective, and one adverb, in that order.
4. Sentence M contains one human's name, one linking verb, one adjective describing a potential health state of a living being, and one temporal adverb, in that order.
5. The subject of Sentence M is a pointer to an individual named 'Mary', and the predicate is an ascription of ill health to the individual so indicated, on the day preceding the statement's utterance.
6. Sentence M asserts that the health of an individual named 'Mary' was not good the day before today.
7. Sentence M says that Mary was sick yesterday.

Just where is the boundary line that says, "You can't do *that* much processing!"? A machine that could go as far as version 7 would have actually *understood*—at least in some rudimentary sense—the content of Sentence M. Work by artificial-intelligence researchers in the field of natural language understanding has produced some very impressive results along these lines, considerably more sophisticated than what is shown here. Stories can be "read" and "understood", at least to the extent that certain kinds of questions can be answered by the machine when it is probed for its understanding. Such questions can involve information not explicitly in the story itself, and yet the machine can fill in the missing information and answer the question.

I am making this seeming digression on the processing of language by computers because intelligent people like Mr Flash qFiasco seem to have failed to recognize that the boundary line between form and content is as blurry as that between blue and green, or between human and ape. This comparison is not made lightly. Humans are supposedly able to get at the "content" of utterances, being genuine language-users, while apes are not. But ape-language research clearly shows that there is some kind of in-between world, where a certain degree of content can be retrieved by a being with reduced mental capacity. If mental capacity is equated with potential processing depth, then it is obvious why it makes no sense to draw an arbitrary boundary line between the form and the content of a sentence. *Form blurs into content as processing depth increases.* Or, as I have always liked to say, "Content is just fancy form." By this I mean, of course, that "content" is just a shorthand way of saying "form as perceived by a very fancy apparatus capable of making complex and subtle distinctions and abstractions and connections to prior concepts".

Flash qFiasco's down-home, commonsense distinction between form and content breaks down swiftly, when analyzed. His charming image of someone making a "category error" by throwing a recipe in with the flour and butter and eggs reveals that he has never had Recipe Cake. This is a delicious cake whose batter is made out of cake recipes (if you use pie recipes, it won't taste nearly as good). The best results are had if the recipes are printed in French, in Baskerville Roman. A preponderance of *accents aigus* lends a deliciously piquant aroma to the cake. My recommendation to Brabner and qFiasco is: "Let them eat recipes."

* * *

Finally, I come to John Case, a computer scientist who wrote from Yale, insisting that there is no conceptual problem whatsoever in translating the French sentence "*Cette phrase en français est difficile à traduire en anglais*" into English. Case's translation was the following English sentence:

> The French sentence "*Cette phrase en français est difficile à traduire en anglais*" is difficult to translate into English.

In other words, Case translates a *self*-referential French sentence into an *other*-referential English sentence. The English sentence talks about the French sentence—in fact it quotes it completely! Something radical is missing here. At one level, of course, Case is right: now the two sentences, one French and one English, both are talking about (or pointing to) the same thing (the French sentence). But the absolute crux of the French one is its tangledness; the English one completely lacks that quality. Clearly Case has had to make a sacrifice, a compromise.

The alternative, which I prefer, is to construct in English an *analogue* to the French sentence: a *self*-referential English sentence, one that has a tangledness isomorphic to that of the French sentence. That's where the *essence* of the sentence lies, after all! "But is that its *translation*?" you might ask. A good question.

Ionesco once remarked, "The French for London is Paris." (Use-mention fanatic that I am, I assume that he meant "The French for 'London' is 'Paris' ", although it is pungent either way.) What he meant was that in understanding situations, French people tend to translate them into their own frame of reference. This is of course true for all of us. If Mary tells Ann, "My brother died", and if Ann does not know Mary's brother, then how can she understand this statement? Surely projection is of the essence: Ann will imagine her *own* brother dying (if she has one—and if not, then her sister, a good friend, possibly even a pet!). This alternate frame of reference allows Ann to empathize with Mary. Now if Ann *did* know Mary's brother somewhat, then she might flicker between thinking of him as the person she vaguely remembers and thinking of her own brother (friend, pet, or whatever) dying. This dilemma (discussed further in the postscript to Chapter 24) arises for all beings with their own preferred vantage points: Do I map things into *what they would be for me*, or do I stand apart and survey them completely objectively and impassively?

Case is advocating the latter, which is all very well as an intellectual stance to adopt, but when it comes to real life, it just won't cut the mustard. To be concrete, one might ask: What was the actual solution used in the French edition of *Scientific American*? The answer, surprising no one, I hope, was this: "This English sentence is difficult to translate into French." I rest my case.

* * *

I wonder what literalists like John Case would suggest as the proper translation of the title of the book *All the President's Men* (a book about the downfall of President Nixon, a downfall that none of the people around him could prevent). Would they say that *Tous les hommes du Président* fills the bill admirably? Back-translated rather literally, it means "All the men of the President". It completely lacks the allusion —the reference by similarity of form—to the nursery rhyme "Humpty Dumpty". Is that dispensable? In my opinion, hardly. To me, the essence of the title resides in that allusion. To lose that allusion is to deflate the title totally.

Of course, what do I mean by "that allusion"? Do I wish the French title to contain, somehow, an allusion to an *English* nursery rhyme? That would be rather pointless. Well, then, do I want the French title to allude to the French version of "Humpty Dumpty"? It all depends how well known that is. But given that Humpty Dumpty is practically an unknown figure to French-speaking people, it seems that something else is wanted. Any old French nursery rhyme? Obviously not. The critical allusion is to the lines "All the King's horses/ And all the King's men/ Couldn't put Humpty together again." Are there—*anywhere* in French literature—lines with a similar import? If not, how about in French popular songs? In French proverbs? Fairy tales?

One might well ask why French-speaking people would ever care about reading a book about Watergate in the first place. And even if they *did* want to read it, shouldn't it be *completely* translated, so that it happens in a French-speaking city? Come to think of it, didn't Ioratno once remark that the French for Washington is Montréal?

Clearly, this is carrying things to an extreme. There must be some middle ground of reasonableness. These are matters of subtle judgment, and they are where being human and flexible makes all the difference. Rigid rules about translation may lead you to a kind of mechanical consistency, but at the sacrifice of all depth and charm. The problem of self-referential sentences is just the tip of the iceberg, as far as translation is concerned. It is just that these issues show up very early when direct self-reference is concerned. When self-reference (or reference in general, for that matter) is indirect, mediated by form, then fluidity is required. The understanding of such sentences involves a mixture of deriving the content and yet retaining the form in mind, letting qualities of the form conjure up flavors and enhance the meaning with a halo of not-quite-conscious pseudo-meanings, connotations, flavors, that flicker in the mind, not quite in reach, not quite out of reach. Self-reference is a good starting point for investigation of this kind of issue, because it is so much on the surface there. You can't sweep the problems under the rug, even though some would like to do so.

* * *

This first column, together with this postscript, provides a good introduction to the book as a whole, because many central issues are touched on: codes, translation, analogies, artificial intelligence, language and machines, mind and meanings, self and identity, form and content—all the issues I originally was motivated by when first writing that collection of teasing self-referential sentences.