METAMAGICAL THEMAS

An anagrammatic title introduces a new contributor to this column

by Douglas R. Hofstadter

I never expected to be writing a column for *Scientific American*. Let me say at once that I am not replacing Martin Gardner; no one could do that. Martin is, however, retiring at the end of this year. Until then he and I shall be sharing this space.

I remember once, years ago, wishing I were in Martin'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 thereafter 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 an imagination of many dimensions. Sometimes I felt overawed by the whole remarkable bunch.

One day five 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 a man who had contributed much to my own mathematical education. Perhaps the most striking thing about Martin to me was his natural simplicity. I have been told that he is an adroit magician. This is hard to believe, because one cannot imagine someone so straightforward pulling the wool over anyone's eyes.

I did not, however, 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 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, antioccultists 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 his wife could pull it to get his attention. A couple of phone calls came, one from the logician and magician Raymond Smullyan, another man I knew well by reputation but did not know belonged to this charmed circle. It was a most enjoyable day.

Martin's act, as they say, will be a hard one to follow. I shall not, however, be trying to be another Martin Gardner. I have my own interests, and no one except Martin himself could have all his interests. Nevertheless, 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 higher order, yet on the other hand the magical thing about magic is that what lies behind it is always nonmagical. 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, magic seldom lurks where many suspect it. Herewith, dear reader, I take my own plunge.

In his column for July, 1979, Martin wrote a kindly review of my book *Gödel, Escher, Bach: An Eternal Golden Braid*. He began the review with a short quotation. If I had been asked to guess what single sentence of my book 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, but something about seeing it again bothered me. I remembered writing it one day, attempting to come up with a new variation on an old theme, but it did not seem 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 some more self-referential sentences and showed them to friends, which began a mild craze in a small group. Here I shall present a selection of what I consider the cream of that crop.

I should not go further without explaining 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 probably stems from the first 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 reason. In any case two of its descendants, the sentences "I am lying" and "This sentence is false," have come to be known as the "Epimenides paradox" or the "liar paradox." (The latter term is more descriptive, the former more orotund. I shall use both.) 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 seem to be 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 one do when told, "Disobey this command"? In the following sentence the Epimenides quality jumps out only after a moment of thought: "This sentance has three 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. Astute readers may have noticed that strictly speaking the phrase "this formula" has no referent, since the English version is no longer a formula!

If one pursues the 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 understand the problem by reading a translation: "This sentence in French is difficult to translate into English." To what does the subject in the latter sentence refer? If it refers to the English sentence, then the subject is self-contradictory, making the English sentence false (whereas the French original was true and harmless); if it refers to the French sentence, then the self-reference is gone. Either way something quite disquieting has happened, and I should point out that it would be just as disquieting, although in a different way, to translate it: "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 phrase, 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 has been slipped into 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, constitutes a hair-raisingly close approach to paradox, yet it is not paradoxical. It turns out to be true, and for this reason it is unprovable in the axiomatic system. The revelation of Gödel's work is that, in any mathematically powerful axiomatic system that is consistent, 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 W. V. 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."

The 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. (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.

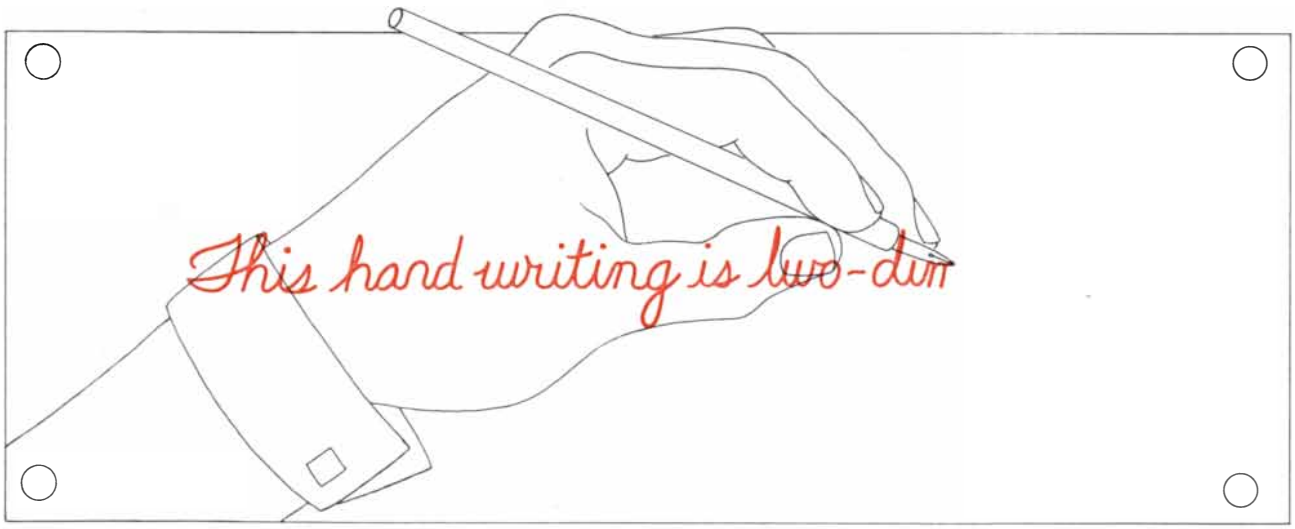
It seems that all paradoxes involve, in one way or another, self-reference, whether it is achieved directly or indirectly. And since the credit for the discovery—or creation—of self-reference goes to Epimenides the Cretan, we might say: "Behind every successful paradox there lies a Cretan."

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 the reader is asked 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, however, so that one is examining it for its linguistic aspects, it



A self-referential drawing of a self-referential sentence ("This hand writing is two-dimensional")

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)."

"À 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."

That is a sentence referring to itself by the (mauled) pronoun "I" 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 usage 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 cue 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 sentence 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 the 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 an entire 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 into 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."

These lead to:

"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 in your mind."

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 have precisely the same thought (as two different computers can run precisely the same program). 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.

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

"You are under my control because you will read until you have reached the end of me."

"Are you the person who is writing me, or the person who is perceiving me?"

"You and I can have only one-way communication, because you are a person and I am 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."

The last of the preceding group of sentences is a rather frightening thought!

"Hey, out there—is that *you* reading me, or is it someone else?"

"Say, haven't you written me somewhere else before?"

The first of the above two sentences addresses its reader; the second addresses its author. Here we have an author addressing a sentence:

"Say, haven't I written you somewhere else before?"

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."

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."

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

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." You might well say "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. Is it fair to say, then, that the second sentence is an answer to the first? It is a little like a person who muses, "What would I be doing now if I had had different genes?" The problem is that he would not be himself; he would be someone else, perhaps the woman across the street mowing her lawn. The pronoun "I" cannot quite keep up with such strange hypothetical world shifts.

There is an even worse problem with the counterfactual above. Changing the value of π is, to put it mildly, a radical change in mathematics, and presumably you cannot change mathematics radically without also radically changing 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 admirable examples, invented by Ann C. Trail (who is also responsible for others in this column), bring us around to sentences that comment on their own form. These 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 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."

"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 music student at Indiana University, is a specialist in self-reference of all kinds. (He has written a story in which every sentence is self-referential.) It might seem unlikely that in such a tiny domain individual styles could arise and flourish, but Moser has developed a self-referential style quite his own. As a friend (or was it Moser himself?) wittily observed, "If David Moser had thought of 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 non-standard English flutzpah."

"This gubblick contains many non-sklarkish English flutzpahs, but the overall pluggandisp can be glorked from context."

"This sentence has sofa six words."

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

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

"It feels so 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 last 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 been featured here. He too has a characteristic way of playing with self-reference. Two of his sentences follow:

"This here sentence don't know english to good."

"If you meet this sentence on the road, erase it."

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

Allusion through similarity of form is another rich vein of self-reference, but unfortunately I can give only two further 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 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. 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."

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

"Does this sentence remind you of Agatha Christie?"

This last sentence 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.)

As I have indicated, the sentences featured in this column were invented by many people, not all of whose names I know. In addition to those already mentioned the inventors include Robert Filman, Margaret Minsky and me. Readers are invited to submit their own self-referential concoctions. Warning: The habit can become addictive!

In closing I cannot resist the plea of the following sentence: "Please publish me in your collection of self-referential sentences."