SO WHAT *IS* this book, — usually known by its acronym, " "— really all about?

That question has hounded me ever since I was scribbling its first drafts in pen, way back in 1973. Friends would inquire, of course, what I was so gripped by, but I was hard pressed to explain it concisely. A few years later, in 1980, when found itself for a while on the bestseller list of , the obligatory one-sentence summary printed underneath the title said the following, for several weeks running: " ." After I protested vehemently about this utter hogwash, they finally substituted something a little better, just barely accurate enough to keep me from howling again.

Many people think the title tells it all: a book about a , an , and a . But the most casual look will show that these three individuals *per se,* august though they undeniably are, play but tiny roles in the book's content. There's no way the book is about those three people!

Well, then, how about describing as ' "? Again, this is a million miles off — and yet I've heard it over and over again, not only from nonreaders but also from readers, even very ardent readers, of the book.

And in bookstores, I have run across gracing the shelves of many diverse sections, including not only math, general science, philosophy, and cognitive science (which are all fine), but also religion, the occult, and God knows what else. Why is it so hard to figure out what this book is about? Certainly it's not just its length. No, it must be in part that delves, and not just superficially, into so many motley topics — fugues and canons, logic and truth, geometry, recursion, syntactic structures, the nature of meaning, Zen Buddhism, paradoxes, brain and mind, reductionism and holism, ant colonies, concepts and mental representations, translation, computers and their languages, DNA, proteins, the genetic code, artificial intelligence, creativity, consciousness and free will — sometimes even art and music, of all things! — that many people find it impossible to locate the core focus.

The Key Images and Ideas that Lie at the Core of

Needless to say, this widespread confusion has been quite frustrating to me over the years, since I felt sure I had spelled out my aims over and over in the text itself. Clearly, however, I didn't do it sufficiently often, or sufficiently clearly. But since now I've got the chance to do it once more — and in a prominent spot in the book, to boot — let me try one last time to say why I wrote this book, what it is about, and what its principal thesis is.

In a word, is a very personal attempt to say how it is that animate beings can come out of inanimate matter. What is a self, and how can a self come out of stuff that is as selfless as a stone or a puddle? What is an ' ', and why are such things found (at least so far) only in association with, as poet once wonderfully phrased it, "

" — that is, only in association with certain kinds of gooey lumps encased in hard protective shells mounted atop mobile pedestals that roam the world on pairs of slightly fuzzy, jointed stilts?

approaches these questions by slowly building up an analogy that likens inanimate molecules to meaningless symbols, and further likens selves (or ' "'s or " ", if you prefer — whatever it is that distinguishes animate from inanimate matter) to certain special swirly, twisty, vortex-like, and patterns that arise only in particular types of systems of meaningless symbols. It is these strange, twisty patterns that the book spends so much time on, because they are little known, little appreciated, counterintuitive, and quite filled with mystery. And for reasons that should not be too difficult to fathom, I call such strange, loopy patterns "

" throughout the book, although in later chapters, I also use the phrase " " to describe basically the same idea.

This is in many ways why — or more precisely, his art — is prominent in the " ", because , in his own special way, was just as fascinated as I am by strange loops, and in fact he them in a variety of contexts, all wonderfully disorienting and fascinating. When I was first working on my book, however, was totally out of the picture (or out of the loop, as we now say); my working title was the rather mundane phrase " ", and I gave no thought to inserting paradoxical pictures, let alone playful dialogues. It's just that time and again, while writing about my notion of strange loops, I would catch fleeting glimpses of this or that print flashing almost subliminally before my mind's eye, and finally one day I realized that these images were so connected in my own mind with the ideas that I was writing about that for me to deprive my readers of the connection that I myself felt so strongly would be nothing less than perverse. And so 's art was welcomed on board. As for , I'll come back to his entry into my "

" a little later.

Back to strange loops, right now. was inspired by my long-held conviction that the " " notion holds the key to unraveling the mystery that we conscious beings call " " or " ". I was first hit by this idea when, as a teen-ager, I found myself obsessedly pondering the quintessential strange loop that lies at the core of the proof of 's famous theorem in mathematical logic — a rather arcane place, one might well think, to stumble across the secret behind the nature of selves and ' "'s, and yet I practically heard it screaming up at me from the pages of and that this was what it was all about.

This preface is not the time and place to go into details — indeed, that's why the tome you're holding was written, so it would be a bit presumptuous of me to think I could outdo its author in just these few pages! — but one

thing has to be said straight off: the strange loop that arises in formal systems in mathematics (*i.e.*, collections of rules for churning out an endless series of mathematical truths solely by mechanical symbol-shunting without any regard to meanings or ideas hidden in the shapes being manipulated) is a loop that allows such a system to " ", to talk about itself, to become " ", and in a sense it would not be going too far to say that by virtue of having such a loop, a formal system :

Meaningless Symbols Acquire Meaning Despite Themselves

What is so weird in this is that the formal systems where these skeletal ' " come to exist are built out of nothing but meaningless symbols. The self, such as it is, arises solely because of a special type of swirly, tangled among the meaningless symbols. But now a confession: I am being a bit coy when I repeatedly type the phrase " " (as at the ends of both of the previous sentences), because a crucial part of my book's argument rests on the idea that meaning cannot be kept out of formal systems when sufficiently complex isomorphisms arise. Meaning comes in despite one's best efforts to keep symbols meaningless!

Let me rephrase these last couple of sentences without using the slightly technical term " ". When a system of " " symbols has patterns in it that accurately track, or mirror, various phenomena in the world, then that tracking or mirroring imbues the symbols with some degree of meaning — indeed, such tracking or mirroring is no less and no more than what meaning is. Depending on how complex and subtle and reliable the tracking is, different degrees of meaningfulness arise. I won't go further into this here, for it's a thesis that is taken up quite often in the text, most of all in Chapters 2, 4, 6, 9, and 11.

Compared to a typical formal system, human language is unbelievably fluid and subtle in its patterns of tracking reality, and for that reason the symbols in formal systems can seem quite arid; indeed, without too much trouble, one can look at them as totally devoid of meaning. But then again, one can look at a newspaper written in an unfamiliar writing system, and the strange shapes seem like nothing more than wondrously intricate but totally meaningless patterns. Thus even human language, rich though it is, can be drained of its seeming significance.

As a matter of fact, there are still quite a few philosophers, scientists, and so forth who believe that patterns of symbols (such as books or movies or libraries or CD-ROM's or computer programs, no matter how complex or dynamic) have meaning on their own, but that meaning instead, in some most mysterious manner, springs only from the organic chemistry, or perhaps the quantum mechanics, of processes that take place in carbon-based biological brains. Although I have no patience with this parochial, bio-chauvinistic view, I nonetheless have a pretty clear sense of its intuitive appeal. Trying to don the hat of a believer in the primacy, indeed the uniqueness, of brains, I can see where such people are coming from.

Such people feel that some kind of " " takes place only inside our " ", somewhere behind pairs of eyeballs, even though they can never quite put their finger on how or why this is so; moreover, they believe that this semantic magic is what is responsible for the existence of human selves, souls, consciousness, " "'s. And I, as a matter of fact, quite agree with such thinkers that selves and semantics — in other words, that me's and meanings — *do* spring from one and the same source; where I take issue with these people is over their contention that such phenomena are due entirely to some special, though as yet undiscovered, properties of the microscopic hardware of brains.

As I see it, the only way of overcoming this magical view of what ' " and consciousness are is to keep on reminding oneself, unpleasant though it may seem, that the " " that nestles safely inside one's own cranium is a purely physical object made up of completely sterile and inanimate components, all of which obey exactly the same laws as those that govern all the rest of the universe, such as pieces of text, or CD-ROM's, or computers. Only if one keeps on bashing up against this disturbing fact can one slowly begin to develop a feel for the way out of the mystery of consciousness: that the key is not the out of which brains are made, but the that can come to exist inside the stuff of a brain.

This is a liberating shift, because it allows one to move to a different level of considering what brains are: as that support complex patterns that mirror, albeit far from perfectly, the world, of which, needless to say, those brains are themselves denizens — and it is in the inevitable self-mirroring that arises, however impartial or imperfect it may be, that the strange loops of consciousness start to swirl.

Smashes through 's Maginot Line

I've just claimed that the shift of focus from material components to abstract patterns allows the quasi-magical leap from inanimate to animate, from nonsemantic to semantic, from meaningless to meaningful, to take place. But how does this happen? After all, not *all* jumps from matter to pattern give rise to consciousness or soul or self, quite obviously: in a word, not all patterns are conscious. What kind of pattern is it, then, that is the telltale mark of a ? 's answer is: .

The irony is that the first strange loop ever found — and my model for the concept in general — was found in a system . I speak of and 's famous treatise , a gigantic, forbidding work laced with dense, prickly symbolism filling up volume after volume, whose creation in the years 1910–1913 was sparked primarily by its first author's desperate quest for a way to circumvent paradoxes of self-reference in mathematics.

At the heart of lay 's so-called " ", which, much like the roughly contemporaneous Maginot Line, was designed to keep " " out in a most staunch and watertight manner.

For the French, the enemy was Germany; for , it was self-reference. believed that for a mathematical system to be able to talk about itself in any way whatsoever was the kiss of death, for self-reference would — so he thought — necessarily open the door to self-contradiction, and thereby send all of mathematics crashing to the ground. In order to forestall this dire fate, he invented an elaborate (and infinite) hierarchy of levels, all sealed off from each other in such a manner as to definitively — so he thought — block the dreaded virus of self-reference from infecting the fragile system.

It took a couple of decades, but eventually the young Austrian logician realized that and 's mathematical Maginot Line against self-reference could be most deftly circumvented (just as the Germans in World War II would soon wind up deftly sidestepping the real Maginot Line), and that self-reference not only had lurked from Day One in , but in fact plagued poor in a totally unremovable manner. Moreover, as made brutally clear, this thorough riddling of the system by self-reference was not due to some weakness in , but quite to the contrary, it was due to its . Any similar system would have exactly the same "defect". The reason it had taken so long for the world to realize this astonishing fact is that it depended on making a leap somewhat analogous to that from a brain to a self, that famous leap from inanimate constituents to animate patterns.

For , it all came into focus in 1930 or so, thanks to a simple but wonderfully rich discovery that came to be known as " " — a mapping whereby the long linear arrangements of strings of symbols in any formal system are mirrored precisely by mathematical relationships among certain (usually astronomically large) whole numbers. Using his mapping between elaborate patterns of meaningless symbols (to use that dubious term once again) and huge numbers, showed how a statement *about* any mathematical formal system (such as the assertion that is contradiction-free) can be translated into a mathematical statement number theory (the study of whole numbers). In other words, any metamathematical statement can be imported *into* mathematics, and in its new guise the statement simply asserts (as do all statements of number theory) that certain whole numbers have certain properties or relationships to each other. But on another level, it also has a vastly different meaning that, on its surface, seems as far removed from a statement of number theory as would be a sentence in a novel.

By means of 's mapping, any formal system designed to spew forth truths about "mere" numbers would also wind up spewing forth truths — inadvertently but inexorably — about its own properties, and would thereby become " ", in a manner of speaking. And of all the clandestine instances of self-referentiality plaguing and brought to light by , the most concentrated doses lurked in those sentences that talked about their *own* , and in particular said some very odd things about themselves, such as " ". And let me repeat: such twisting-back, such looping-around, such self-enfolding, far from being an eliminable defect, was an inevitable by-product of the system's vast power.

Not too surprisingly, revolutionary mathematical and philosophical consequences tumbled out of 's sudden revelation that self-reference abounded in the bosom of the bastion so carefully designed by to keep it out at all costs; the most famous such consequence was the so-called " " of formalized mathematics. That notion will be carefully covered in the chapters to come. and yet, fascinating though it is, incompleteness is not in itself central to 's thesis. For , the most crucial aspect of Gödel's work is its demonstration that a statement's can have deep consequences. even in a supposedly meaningless universe. Thus it is the of 's sentence G (the one that asserts " ") that guarantees that

(which is precisely what G itself claims). It is as if the sentence's hidden meaning had some kind of power over the vacuous symbol-shunting, meaning-impervious rules of the system, preventing them from ever putting together a demonstration of G, no matter what they do.

Upside-down Causality and the Emergence of an " "

This kind of effect gives one a sense of crazily twisted, or upside-down, causality. After all, shouldn't meanings that one chooses to read into strings of meaningless symbols be totally without consequence? Even stranger is that the sentence G is not provable inside is its self-referential meaning; indeed, it would seem that G, being a statement about whole numbers, *ought* to be provable, but — thanks to its extra level of meaning as a statement about itself, asserting its own nonprovability — it is not.

Something very strange thus emerges from the loop: the revelation of the causal power of meaning in a rule-bound but meaning-free universe. And this is where my analogy to brains and selves comes back in, suggesting that the twisted loop of trapped inside an inanimate bulb called a " " also has causal power — or, put another way, that a mere pattern called " " can shove around inanimate particles in the brain no less than inanimate particles in the brain can shove around patterns. In short, an " " comes about — in my view, at least — via a kind of vortex whereby patterns in a brain mirror the brain's mirroring of the world, and eventually mirror themselves, whereupon the vortex of " " becomes a real, causal entity. For an imperfect but vivid concrete analogue to this curious abstract phenomenon, think of what happens when a TV camera is pointed at a TV screen so as to display the screen on itself (and that screen on itself, etc.) — what in I called a " ", and in my later writings I sometimes call a " ".

When and only when such a loop arises in a brain or in any other substrate, is a — a unique new " " — brought into being. Moreover, the more self-referentially rich such a loop is, the more conscious is the self to which it gives rise. Yes, shocking though this might sound, consciousness is not an on/off phenomenon, but admits of degrees, grades, shades. Or, to put it more bluntly, there are bigger souls and smaller souls.