23

On the Seeming Paradox of Mechanizing Creativity

September, 1982

IT is a commonly heard statement that there is such a thing as the "creative spark", that an "unanalyzable leap of the imagination" takes place when a great mind comes up with a new idea or work of art. Great creators are sometimes said to be a "quantum leap" away from ordinary mortals. People like Mozart are held to be somehow divinely inspired, to have magical insights for which they could no more be expected to be able to account than spiders for the wondrous webs they weave. It is all felt to be somehow too deep down, too hidden, too occult a gift, to be mechanical in any sense. Creativity, in fact, is perhaps one of the last refuges of the soul. "You may mechanize your *logic,* " says the English professor to the computer scientist, "but you'll never lay a finger on *poetry.* " (You may substitute music or any other domain of artistic creation for poetry.)

Is this kind of statement irrational? Is it a reflection of a deep-seated fear that even this most sacred aspect of humanity is doomed to be taken over soon by metallic machines, or by silicon chips? Why make such a big deal out of an activity of the human mind which, like every other activity in life, has shades and degrees? After all, the creative blurs with the mundane so much that it would be hopeless, would it not, to try to cull what is truly creative from what is not? Or—is there some clean dividing line that distinguishes the run-of-the-mill workaday deviser of ditties from the Great Composer of Eternal Symphonic Masterpieces? And if so, is it possible that here lies the elusive difference between the living and the dead, the human and the machine, the mental and the mechanical?

With such a "magical" view of creativity, there is, of course, a problem. It would seem to imply that the poor composer of ditties is actually dead and mechanical inside; that only certified geniuses like Mozart are qualitatively different from machines—and that even old Mozart was nonmechanical only when he was composing (certainly not when he was merely sipping ale at a tavern!). Probably most people who believe in the magical view of creativity would dispute this way of portraying their position. They would maintain that Mozart was nonmechanical *all* the time; moreover that you and I, no less than Mozart, are also nonmechanical all the time. No matter that some, even many, human abilities have already been mechanized or will be mechanized someday.

About the touchy question of the mechanization of the mental, many educated people feel that, although a machine may now or someday be able to do a creditable job of acting like a person, any machine's performance will always remain lackluster and dull, and that after a while, this dullness will always shine through. You'll simply be able to tell that it is unoriginal, that its ideas and thoughts are all being drawn from some storehouse of formulas and *clichés*, that ultimately there is nothing alive and dynamic—no *élan vital*—behind its *façade.* If it comes up with a *bon mot* now and then, well, *tant mieux*—but even the best will just be an automaton *par excellence.* There may be nothing specific to point to other than the "vibes" you pick up of its dullness and unoriginality, but after a while they will inevitably start to come in loud and clear. (Incidentally, I would be delighted if some of the more vocal antimechanists felt that way, instead of insisting, as they more often do, that operational tests are of no use in deciding who or what possesses "genuine mental states".)

This sense that you will eventually be able to "just tell", from its inevitable lack of sparkle, that you're dealing with a machine and not a person, seems to depend upon a tacit assumption about human thought, one with which I fully agree: namely, that "creative spark" is not the exclusive property of just a few rare individuals down the centuries, but quite to the contrary, it is an intrinsic ingredient of the everyday mental activity of everyone, even the most run-of-the-mill people. In short, it seems that people who feel that machines—even intelligent ones—will always remain duller than minds are tacitly relying on the following thesis: Creativity is part of the very fabric of all human thought, rather than some esoteric, rare, exceptional, and fluky by-product of the ability to think, which every so often surfaces in places spread far and wide.

With this thesis I agree. Where I differ with the antimechanists is over the matter of whether creativity lies *beyond* intelligence. I see creativity and insight, for machines no less than for people, as intimately bound up with intelligence, so that I cannot imagine a noncreative yet intelligent machine —something that, in order to make a point about what is essentially human, they seem to be willing and able to do. To me, "noncreative intelligence" is a flat-out contradiction in terms.

* * *

In this column, I would like to describe some ideas I have about how creativity is founded on mechanisms, mechanisms that, to be sure, lie deeply hidden in the depths of the structure of our brains, but mechanisms that nonetheless exist and can perhaps be approximated using the hardware and software of the machines we have today, crude though they are in certain ways. The gist of my notion is that having creativity is an automatic consequence of having the proper representation of *concepts* in a mind. It is not something you add on afterward. It is built into the way concepts are. To spell this out more concretely: If you have succeeded in making an accurate model of *concepts,* you have thereby also succeeded in making a model of the creative process, and even of consciousness.

Another way of talking about concepts is to talk about memory, which is the "place" where concepts are stored. It is the organization of memory that defines what concepts are. Incidentally, when I first wrote the preceding sentence, it ended differently. It said, "It is the organization of memory that defines what concepts will be accessible under what conditions." But on rereading it, I felt it was too weak that way. It took for granted the notion that all readers have a clear concept of what a concept is. But that is hardly takable-for-granted! Granted, we all have *some* concept of what a concept is, but a *clear* one?

So I dropped the phrase beginning with "will be accessible" and replaced it with a stark "are". This way, the sentence does more than simply state that memory is a storehouse of some things called concepts. It emphasizes that what establishes the "concepthood" of something is the way it is integrated into memory. Or to put it the other way 'round, nothing is a concept except by virtue of the way it is connected up with other things that are also concepts. In other words, the property of being a concept is a property of connectivity, a quality that comes from being embedded in a certain kind of complicated network, and from nowhere else. Put this way, concepts sound like structural or even topological properties of vast tangly networks of sticky mental spaghetti.

That's more or less the image I feel it is important to convey: namely, that concepts derive all their power from their connectivity to one another. And now, having expressed that idea, I can return to the sentence as it was originally put: It is the organization of memory that defines what concepts will be accessible under what conditions—and surely, the happy choice of the right concept at the right time is the essence of the creative. Therefore it is imperative to study deeply the nature of that network—to ask the question "What is a concept?".

Some questions that come to mind are: What is the relationship between a general, or Platonic, concept, such as that of "tree", and the concept you form of some specific tree? That is, what is the distinction between semantic or perceptual *categories* and the representations of individual *instances* of them? How is a given situation filed away in memory so that one has access to it under an enormous variety of future situations—access that is often via analogy or other abstract pathways, rather than by simplistic superficial traits? Or, to flip that coin, how does a given situation cause the highly selective retrieval from memory of a small number of previous situations that seem relevant? Only through a deep understanding of the organization of memory—which is to say, only by answering the question "What is a concept?"—will we be able to make models of the creative process. This will be a long and arduous process, not one that will yield answers overnight, or even in a few decades. Nonetheless, we have the right beginnings, in the sciences of cognitive psychology and artificial intelligence. Philosophers of mind and neuroscientists will undoubtedly contribute as well. The union of all these disciplines is called "cognitive science".

* * *

A question that arises at the outset is: "What kinds of objects have concepts stored inside them, and what kinds do not?" One of my favorite passages that opens this question wide is in Dean Wooldridge's book *Mechanical Man: The Physical Basis of Intelligent Life,* and it runs this way:

> When the time comes for egg laying, the wasp *Sphex* builds a burrow for the purpose and seeks out a cricket which she stings in such a way as to paralyze but not kill it. She drags the cricket into the burrow, lays her eggs alongside, closes the burrow, then flies away, never to return. In due course, the eggs hatch and the wasp grubs feed off the paralyzed cricket, which has not decayed, having been kept in the wasp equivalent of a deepfreeze. To the human mind, such an elaborately organized and seemingly purposeful routine conveys a convincing flavor of logic and thoughtfulness—until more details are examined. For example, the wasp's routine is to bring the paralyzed cricket to the burrow, leave it on the threshold, go inside to see that all is well, emerge, and then drag the cricket in. If the cricket is moved a few inches away while the wasp is inside making her preliminary inspection, the wasp, on emerging from the burrow, will bring the cricket back to the threshold, but not inside, and will then repeat the preparatory procedure of entering the burrow to see that everything is all right. If again the cricket is removed a few inches while the wasp is inside, once again she will move the cricket up to the threshold and reenter the burrow for a final check. The wasp never thinks of pulling the cricket straight in. On one occasion this procedure was repeated forty times, with the same result.

One can make the obvious remark that perhaps not the wasp but the experimenter was the one in the rut—but humor aside, this is a rather shocking revelation of the mechanical underpinning, in a living creature, of what looks like quite reflective behavior.

There seems to be something supremely unconscious about the wasp's behavior here, something totally opposite to what we feel *we* are all about, particularly when we talk about our own consciousness. I propose to call the quality here portrayed *sphexishness,* and its opposite *antisphexishness* (a vexish word to pronounce!), and then I propose that consciousness is simply the possession of antisphexishness to the highest possible degree. The point is that sphexishness and antisphexishness are two extremes along a continuum. Let me give a few examples distributed along that continuum, starting at the most sphexish and finishing with the most antisphexish:

1. A stuck record. This can be especially ironic if it's a recording of something that has a vibrant, lifelike dynamism to it (such as the music of contemporary composer Steve Reich), and then the illusion is shattered by the mechanical repetition of the jumping needle.
2. The *Sphex* wasp herself, and other examples from the insect world. For instance, suppose you have a mosquito in your bedroom. You try to swat it, and miss. It takes off and flies around the room, losing you. But after a while, it settles down and you spot it somewhere on the wall. Again you try to swat it and miss. As this cycle progresses, is the mosquito aware of the repetition? Does it begin to sense that there is an organized conspiracy against it, or does each new swat attempt come as fresh and unexpected as the previous one? Does the mosquito formulate some such notion as "the animate agent trying to wipe me out"? Sadly for the mosquito (but fortunately for you), it seems highly doubtful.
3. A herd of cattle in a corral, waiting to get branded. There is general commotion and hubbub, caused by the noise each cow makes at the moment of branding, and propagated outward by the cows closest to it. But does each cow in the corral recognize the overall pattern? Is its increased state of agitation due to the fact that the cow sees what is coming, or is it rather just a kind of vague apprehension, perhaps merely a raised adrenaline level without any specific meaning or referential quality?
4. A dog who is fooled every time by a faking motion in which you pretend to throw a ball, but instead don't release it. Actually, I don't know any dog who would fall for such an elementary trick. However, I do know a dog (who shall remain nameless—although he does happen to be an Airedale) who did not catch on when I threw his toy to an upstairs landing instead of down the hall (where he expected it). I led him up the stairs and showed him where it was. I expected he would know to go upstairs the next time. But no such luck. He just ran down the hallway again. Even after I had thrown his toy upstairs fifteen times more, he *still* ran down the hallway, then came back looking confused. Poor doggie! True, some of those seventeen painful times he did start going up the stairs, but each time he got only partway up, then turned around, and hightailed it down the hallway. To me, it was a disappointingly sphexish kind of behavior for a dog.
5. Glassy-eyed gamblers in Las Vegas, glued to their slot machines. To this can be added glassy-eyed teen-agers and college students glued to video games and pinball machines. Is there not some kind of deadening rut here? And yet so many people do this over and over again with seeming pleasure.

6. A happy-go-lucky person who sings or whistles all the time—and if you listen closely, you notice that it's always the same little refrain, day in, day out, year in, year out: never any variety.
7. People who make what seems to be the same joke, only in slightly different guises, over and over and over again. Or inveterate punsters, who simply cannot stop making one pun after another.
8. Junior-high-school students who fill each other's yearbooks with those same pat phrases and corny poems as *your* junior-high class did.
9. A mathematician who exploits one single technique to advantage in paper after paper, making advances in many different branches in mathematics, yet always with a distinct, idiosyncratic touch, and always, in some deep sense, just doing "the same old trick" again and again.
10. People whose rut-stuck behavior leads them down harmful pathways in their lives, for instance in their romances or their jobs. We all know people who "blow it" in the same way each time when faced with a situation that matters.
11. Social trends that become completely stylized and predictable, such as the endless trashy sitcoms that television networks keep churning out, the movies one after another based on some gimmick exploited in slightly different ways. For instance, one could perceive the movies *Breaking Away, The Black Stallion,* and *Chariots of Fire* as simply three ways of plugging specific values for variables into one successful formula—an upcoming championship race, a lovable underdog, a rival, and, of course, ultimate victory. And these are sophisticated, compared to some books and movies that much more blatantly exploit famous predecessors.
12. Styles in art that become dated and routinized to the point of no longer being creative. This happens to every style, but at the moment of its happening, there are always some people who are breaking out of the rut and creating totally new styles. However, there are others who become technically proficient at an old style, and who continue to create in an old-fashioned vein.

How different are these last few examples from the stuck record, or from the *Sphex* wasp? What is the real difference we feel as we progress down this list?

I would summarize it by saying that it is a general *sensitivity to patterns,* an ability to spot patterns of unanticipated types in unanticipated places at unanticipated times in unanticipated media. For instance, *you* just spotted an unanticipated pattern—five repetitions of a word. And I'm sure you picked up on all the French phrases crowded together earlier on in this chapter. Neither in your schooling nor in your genes was there any explicit preparation for such acts of perception. All you had going for you is *an ability to see sameness.* All human beings have that readiness, that alertness, and that is what makes them so antisphexish. Whenever they get into some kind of "loop", they quickly sense it. Something happens inside their heads—a kind of "loop detector" fires. Or you can think of it as a "rut detector", a "sameness detector"—but no matter how you phrase it, the possession of this ability to *break out of loops of all sorts* seems the antithesis of the mechanical. Or, to put it the other way around, the essence of the mechanical seems to be in its lack of novelty and its repetitiveness, in its trappedness in some kind of precisely delimited space. This is why the wasp, the dog, even some humans seem so mechanical.

* * *

How many computers do you know that would react with outrage (or guffaws) to the simultaneous occurrence on a single mailing list of "Bernie Weinreb", "Bernie W. Weinreb". "Mr. Bernie Weinreb, R.M.", "Barnie Weinrab", and so forth? Computers do not have automatic sensitivity to patterns in the data that they deal with. And of course, how could they be expected to? As one old saw goes, they do only what they are programmed to do. Computers are not inherently bored by adding long columns of numbers, even when all the numbers are the same. But people are. What is the difference?

Clearly there is something lacking in the machine that allows it to have this unbounded tolerance for repetitive actions. This thing that is lacking can be described in a few words: It is the ability to watch oneself as one deals with the world, to perceive in one's own activities a pattern, and to be able to do so at many levels of abstraction. Thus, consider the case of a hypothetical self-watching computer. To be sensitive in this way, it should get bored whenever it is forced to add a long column of identical numbers together. Wouldn't you? It should get bored whenever it is forced to do just adding over and over again, even when the numbers are different. Wouldn't you? It should even get bored when asked to do many arithmetic operations in any sort of repetitive pattern! Wouldn't you? Any loop of any sort should become tedious! Wouldn't it?

But where does it stop? Surely if a computer could perceive that all it *ever* does is pull up one instruction after another from memory (a piece of hardware, not to be confused with human memory), execute those instructions, and change various registers, it would yawn very boredly and probably soon go to sleep. And by the same token, you or I, if we ever gained access to the firings of our neurons, would find watching the activity to be one of the most stultifying things imaginable.

But this is not the kind of self-watching I mean. Watching one's own internal microscopic patterns is bound to be boring, because any complex system is bound to be made up out of thousands, millions, or even more copies of small elements (such as gears, transistors, cells, and so on). What is critical is to be able to watch activities on a completely different level—the *collective* level, in which huge patterns of activity of these many components assume regular behaviors perceptible on their own. A hurricane is a huge pattern of activity of tiny atoms, but one that has such regularity and pattern that we can predict hurricanes without ever thinking of their constituent atoms. A *thought* is a huge pattern of activity of tiny cells, of which much the same can be said.

Antisphexishness has to do with self-perception at this kind of level. Rather than watching its neurons or transistors or registers, an antisphexish being watches its own high-level patterns, looking for similarities somewhat the way meteorologists might look for one hurricane following another in a regular way.

Thus we should not expect or even want a self-watching computer to be able to see down to the level of its circuitry; it would not watch itself doing machine-language operations such as *ADD, STORE,* and *JUMP* in loop-like patterns. The effects of such operations are to change larger things called "data structures" in memory. Self-watching involves monitoring those changes as they happen, filtering out the dull ones, and recording certain aspects of the interesting ones in *other* data structures. (The fact that such monitoring, filtering, and recording would, on a more microscopic level, involve the very same kinds of elementary machine-language operations would be invisible to the computer, since it should be shielded from that detailed a view of itself.) Thus patterns in the changes taking place in *one* set of data structures would get recorded in another set of data structures. Should we then not set up a third level of data structures, to watch the second level, should patterns occur in it? And a fourth, to watch the third? This seems prime territory for an infinite regress: an endless hierarchy of structures, each one monitoring changes in the level below it.

Now that is quite true, and it is because you are a self-watching human being that you caught onto this pattern, and probably before I had spelled it out. It is in the nature of human pattern perception to be able to detect such infinite regresses, and to stop them short before they ever get anywhere. But what about the hypothetical self-watching computer, with its infinitely many layers of watchers?

Well, surely one of the most salient features—no, definitely the *most* salient feature—of what I have just described is the pattern of the data structures themselves: the hierarchy stretching upwards repetitively towards infinity. Shouldn't this pattern be as blatant to a self-watcher as it is to us? Indeed yes, it should. If we were to label the bottom level '0' and the first watching level '1', then logically we should label the further levels '2', '3', and so on. Each level in this potentially infinite set can be identified with a natural number. Once the pattern is perceived by a watcher, that watcher can form the general concept of "all the levels seen at once", associated with the concept of "all the natural numbers conceived of at once". The conventional name for the set of all natural numbers is 'ω' (omega), which we can take as the name of a *new* watching level that looks out for patterns in this potentially infinite tower of watchers.

You need not worry, by the way, that in proposing such a self-watching computer I am presupposing an infinite machine. Precisely the opposite. The whole purpose of stopping infinite regress in its tracks is so that we will *not* need to actually build an infinite tower of data structures and watching processes, a feat that would clearly be impossible, aside from being monumentally sphexish. At any stage, only a finite amount of recording would have been done, so that only a finite number—in fact, a small number—of levels of structure would exist. The only requirement is that there should exist the *potential* to extend it further.

It would be the ω-watcher that would perceive (as you and I and any human being would) the infinite-regress pattern of attempts to build the ω-tower itself. The ω-watcher would catch any such infinite regress before it could start. If a change in level 0 caused a change in level 1 that caused yet another change in level 2, and if these changes seemed to be patterned in such a way that an inevitable infinite ripple upwards would ensue, the ω-watcher, ever alert for such patterns in the other watchers, would come to the rescue, shouting "Wait! Enough! Halt!" Thus in fact, no infinite regress would actually occur; it would be nipped in the bud by the same sorts of mechanisms that allow you to cut off a bore at a party. "Excuse me, I think I'll go get some more punch."

* * *

The problem is, there's nothing to prevent the ω-level itself from going into loops—so if we're going to obviate that, we have to have a higher watcher—conventionally called "$\omega+1$". Uh-oh! Before I even had a chance to begin spelling it out, you sniffed a new infinite regress! (You ruin all my fun!) Well, I'm going to spell it out, anyway. Level $\omega+1$ needs to be watched by level $\omega+2$, and that level by level $\omega+3$. Thus we have a *second* potentially infinite tower of watchers, all of whom will be watched over by the Grand Watcher: level 2ω. But if there can be *two* towers, then why not *three*? And so, of course, it goes. Wheels within wheels, patterns of patterns of patterns. We get watchers 2ω, 3ω, and now our tower of towers needs a new Great-Grand Watcher: ω^2. And then—

Excuse me; I think I'll go get some more punch. There is a problem once you start getting into infinite regresses composed of other infinite regresses—the whole thing just never stops, and it becomes a *bore.* Or not exactly a bore, but a very complex and confusing thing, whose reality and relevance become ever more questionable. And yet, when you bring it back to the domain of sphexishness, it becomes the very real and very relevant question of how to build a machine that can sense unanticipated patterns in its own behavior.

This is related to a classic problem in the theory of computability, called the *halting problem*: It is the question of whether there exists any computer program that can inspect other programs before they run, and reliably predict whether or not they will go into infinite loops ("going into an infinite loop" means, of course, never coming to a halt—and conversely, "halting" means avoiding any infinite loop). The answer turns out to be "Definitely not", and for elegant, deep reasons. (Recall Chapter 21.) Of course, the thing hinges on getting this halting inspector to try to predict its own behavior when looking at itself trying to predict its own behavior when looking at itself trying to predict its own behavior when . . . Excuse me; I think I'll go get some more punch.

This halting-problem idea is closely related to our question about self-watching programs, but it is not really the same thing. First of all, the halting problem is concerned with an inspection to be carried out on programs *before* they are running, like looking at blueprints of buildings before they are built to see if they are earthquake-proof. Here we are talking about a program that is observing some program *while* it is running—and what's more, it's not just "some program" that it is watching, but *itself.* Of course, not *all* of its attention is being devoted to seeing if it's gotten into a rut (for that would itself constitute ruttish behavior!), but while it's doing other things, it's keeping its eye peeled, so to speak, for signs of ruttishness inside itself.

In computability theory, when a program or system of any sort turns back on itself in this manner, the turning-back-on-itself is known as *diagonalization.* To some people, diagonalization seems a bizarre exercise in artificiality, a construction of a sort that would never arise in any realistic context. To others, its flirtation with paradox is tantalizing and provocative, suggesting links to many deep aspects of the universe. Now here we see a *dynamic* diagonalization—a self-watching program—that seems to be closely connected with what makes a human being so utterly different from a stuck record or a *Sphex* wasp. Surely that is not such a bizarrely artificial thing to ponder!

Probably the most significant difference between the halting problem and the idea of a self-watching program is that in trying to build an artificial intelligence, we are not really so concerned with the mathematical perfection of our self-watching system as with its likelihood of survival in a complex world; after all, that's what intelligence is about. So if there is a mathematical theorem telling us that no program whatsoever will be a *perfect* self-watcher, able to catch itself in any conceivable kind of infinite regress, well, that is simply a statement that *perfect* intelligence is unreachable—something that ought to please us rather than dismay us, since it would be rather horrible and disappointing if someone came up with some finite program after a while, and could legitimately announce, "Well, folks, here it is at last: the end-all of intelligence, a *perfectly* intelligent program."

But don't worry about that. The metamathematical work of Kurt Gödel, Alan Turing, Stephen Kleene, and others, on such things as the halting problem and the theory of infinite ordinals (such as the towers of numbers and ω's), tells us that this scenario will not come to pass, for neither is there a perfect halting inspector, nor is there any ultimate scheme for naming ordinals. What this latter result means is that there is no finite mechanism that can possibly detect all patterns, patterns of patterns, patterns of patterns of patterns of patterns (aha!—fooled you that time, didn't I?), and so on.

* * *

In his famous paper "Minds, Machines, and Gödel", the English philosopher J. R. Lucas attempted to capitalize on these sorts of "negative" results of metamathematics by claiming that they provided the key element in a proof that no machine could ever be conscious in the way that humans are. Let Lucas speak for himself:

> At one's first and simplest attempts to philosophize, one becomes entangled in questions of whether when one knows something one knows that one knows it, and what, when one is thinking of oneself, is being thought about, and what is doing the thinking. After one has been been puzzled and bruised by this problem for a long time, one learns not to press these questions: the concept of a conscious being is, implicitly, realized to be different from that of an unconscious object. In saying that a conscious being knows something, we are saying not only that he knows it, but that he knows that he knows it, and that he knows that he knows that he knows it, and so on, as long as we care to pose the question: there is, we recognize, an infinity here, but it is not an infinite regress in the bad sense, for it is the questions that peter out, as being pointless, rather than the answers. The questions are felt to be pointless because the concept contains within itself the idea of being able to go on answering such questions indefinitely. Although conscious beings have the power of going on, we do not wish to exhibit this simply as a succession of tasks they are able to perform, nor do we see the mind as an infinite sequence of selves and super-selves and super-super-selves. Rather, we insist that a conscious being is a unity, and though we talk about parts of the mind, we do so only as a metaphor, and will not allow it to be taken literally.
>
> The paradoxes of consciousness arise because a conscious being can be aware of itself, as well as of other things, and yet cannot really be construed as being divisible into parts. It means that a conscious being can deal with Gödelian questions in a way in which a machine cannot, because a conscious being can both consider itself and its performance and yet not be other than that which did the performance. A machine can be made in a manner of speaking to 'consider' its performance, but it cannot take this 'into account' without thereby becoming a different machine, namely the old machine with a 'new part' added. But it is inherent in our idea of a conscious mind that it can reflect upon itself and criticize its own performances, and no extra part is required to do this: it is already complete, and has no Achilles' heel.

Somehow—and I think understandably—Lucas was under the impression that human beings are endowed with powers that are equivalent to a self-watcher of infinite depth, someone who will detect and terminate any and all patterned behavior: the ultimate in antisphexishness. I call this hypothetical ability "Breaking Out Of Loops Everywhere"—"BOOLE" for short, in honor of George Boole, who wrote one of the most influential books of the nineteenth century, *The Laws of Thought,* surely a forerunner of today's artificial intelligence work.

Lucas seems to think that to be human is to be endowed with this "BOOLE" ability—this total and perfect antisphexishness—intrinsically. On reflection, however, one realizes this surely is not the case. Despite not being *Sphex* wasps or Airedales, we humans are all still vulnerable to getting caught in ruts, as I attempted to point out in the dozen-item list above. None of us is immune. Each of us—even the Mozarts among us—exhibits a "cognitive style" that in essence defines the ruts we are permanently caught in.

Far from being a tragic flaw, this is what makes us interesting to each other. If we limit ourselves to thinking about music, for instance, each composer exhibits a "cognitive style" in that domain—a musical style. Do we take it as a sign of weakness that Mozart did not have the power to break out of his "Mozart rut" and anticipate the patterns of Chopin? And is it because he lacked spark that Chopin could not see his way to inventing the subtle harmonic ploys of Maurice Ravel? And from the fact that in "Bolero" Ravel does not carry the idea of pseudo-sphexish music to the intoxicating extreme that Steve Reich has, should we conclude that Ravel was less than magical?

On the contrary. We celebrate individual styles, rather than seeing them negatively, as proofs of inner limits. What in fact is curious is that those people who are able to put on or take off styles in the manner of a chameleon seem to have no style of their own and are simply saloon performers, amusing imitators. We accord greatness to those people whose "limitations", if that is how you want to look at it, are the most apparent, the most blatant. If you are familiar with his style, you can recognize music by Maurice Ravel any time. He is powerful *because* he is so recognizable, because he is trapped in that inimitable "Ravel rut". Even if Mozart *had* jumped that far out of his Mozart system, he still would have been trapped inside the Ravel system. You simply *can't* jump infinitely far!

The point is that Mozart and Ravel, and you and I, are all highly antisphexish, but not perfectly so, and it is at that fuzzy boundary where we can no longer quite maintain the self-watching to a high degree of reliability that our own individual styles, characters, begin to emerge to the world.

Although Lucas has been roundly criticized, and rightly so, I believe, by many philosophers, logicians, and computer scientists for failing to see many important subtleties of the Gödel argument on which he bases his paper, most of his critics have failed to see the crucial aspect of mind that Lucas was one of the first to point out. Lucas correctly observes that the degree of nonmechanicalness that one perceives in a being is directly related to its ability to self-watch in ever more exquisite ways. Unfortunately, too many artificial-intelligence people are ready to pooh-pooh the Lucas article on the grounds that its central thesis—the impossibility of mechanizing mind—is wrong. What they miss is that it is pointing at very deep issues that have much to do with the very core of intelligence and creativity.

* * *

Earlier I stressed the importance of the organization of memory and the pressing need to come at the question "What is a concept?" Critical to the way our memory is organized is our automatic mode of storing and retrieving items, our knowledge of when we know and do not know, of how we know or why we wouldn't know. Such aspects of what is sometimes called "metaknowledge" are fluidly integrated into the way our concepts are meshed together. They are not some sort of "extra layer" added on top by a second-generation programmer who decided that metaknowledge is a good thing, over and above knowledge! No, metaknowledge and knowledge are simmering together in a single stew, totally fused and flavoring each other richly. This makes self-watching an automatic consequence of how memory is structured. How is this wondrous stew of antisphexishness realized in the human brain?

And how can we create a program that, like a human brain, is all "of a piece", a program that is not simply a stack of ever-higher "other-watchers", but is truly a seamless "*self*-watcher", where all levels are collapsed into one? If we wish to have a program that breaks out of the extremely sphexish mold that all programs seem to be in today, we have to figure out how a flexible perception program might exploit its own flexibility to look at itself. Of course, no such program will be written as I just stated. That is, it will *not* come into being in the following way:

Step 1. We write a flexible perception program.
Step 2. We turn that program back on itself as a self-watcher.

Rather, to achieve the results desired in Step 1, we must have incorporated the goals of Step 2 into the design from the start! In other words, these two goals are intertwined, more in the following sense:

Goal 1. Flexible perception.
Goal 2. Self-watching.

There is no chronological priority here, for the two goals are too intertwined to have one precede the other. This is a tricky foldback, quite a bit more elaborate than the one involved in the halting problem, yet in spirit related to it.

It is interesting that Lucas' argument was based on Gödel's Theorem, whose proof depends on making one of these seemingly impossible (or at least highly counterintuitive) foldbacks—this one where a mathematical system of reasoning folds back on itself and subsumes itself as an object of study. What is fascinating in that proof is how, in such a system, there is a kind of level-collapse that ensues from the ability of a system to see itself. Rather than there being towers of watchers, then towers of those towers, and so on *ad infinitum* in the worst possible sort of multiply infinite regress, all those degrees and levels of self-perception are achieved at once by the fact that the system can mirror itself. Not that it mirrors itself in every aspect, mind you—for that would entail contradiction—but it does so at all levels of complexity.

The seemingly distinct levels of watcher and watched are totally fused, in the Gödel construction, exactly as Lucas would have it occurring in the minds of all conscious beings. The only thing that Lucas failed to understand is that the ability to fold around and see oneself in the wonderfully circular Gödelian way does not—in fact, *cannot*—bring with it *total* antisphexishness. That, fortunately or unfortunately, depending on your point of view, is a chimera.

* * *

Back in 1952, the philosopher and composer John Myhill wrote a lyrical article entitled "Some Philosophical Implications of Mathematical Logic: Three Classes of Ideas". The three classes are borrowed from mathematical logic, and Myhill's names for them are the *effective,* the *constructive,* and the *prospective.* In logic, they are known more technically as the *recursive,* the *renotrec* (short for "recursively enumerable but not recursive"), and the *productive.* Their essence is described below.

A category is *effective* provided that there is a way, given a candidate for membership, of deciding without any doubt whether that object is or is not a member. Is Ronald Reagan a KGB agent? Is the Pope Catholic? Although these two questions are easy to answer, which would seem to imply that being a KGB agent and being Catholic are examples of the effective, this is slightly misleading. Was Lee Harvey Oswald a KGB agent? Is an excommunicated bishop Catholic? Examples like these show that these categories are not genuinely effective categories—but then nothing in the real world is as clean as it is in logic. I could have asked, "Is 29 prime?" but I wanted to show how these notions extend beyond the mathematical realm. In natural languages, grammaticality (syntactic well-formedness) is a rather fuzzy property, but in an idealized language or formal system, it would be a perfect example of an effective property.

We pass on to the *constructive.* A property that is constructive is more elusive than one that is effective. The idea here is that some means exists whereby members of the category can be churned out one by one, so that you will eventually see any particular member if you wait long enough, but no means exists for doing the complementary operation—namely, churning out *non*members, one by one. Unfortunately, although this kind of set in mathematics is an extremely important one, easily definable examples of it are rather hard to come by. The set of all theorems in any formal axiomatic system is always recursively enumerable, but very often its complement is also, which turns the set into an effective one rather than a constructive one. You have to be dealing with a formal system whose *non*theorems are not themselves producible by some complementary formal system. Only then do you have a renotrec, or constructive, set. The set of theorems of any formalized version of number theory turns out (by Gödel's theorem) to have this property.

So much for the "constructive". We finally come to the *prospective,* also known as the *productive.* Myhill's characterization of it is this: "A prospective character is one which we cannot either recognize or create by a series of reasoned but in general unpredictable acts." Thus it is neither effective nor constructive. It eludes production by *any* finite set of rules. However—and this is important—it can be *approximated* to a higher and higher degree of accuracy by a series of bigger and better sets of generative rules. Such rules tell you (or a machine) how to churn out members of this prospective category. In mathematical logic, works by Tarski and Gödel establish that *truth* has this open-ended, prospective character. This means that you can produce all sorts of examples of truths—unlimitedly many—but no set of rules is ever sufficient to characterize them *all.* The prospective character eludes capture in any finite net. (See Chapter 13 for a discussion of Platonic notions such as "chairness", 'A'-ness, etc.)

As his prime example outside of mathematical logic of this quality, Myhill suggests beauty. As he puts it:

> Not only can we not guarantee to recognize it [beauty] when we encounter it, but also there exists no formula or attitude, such as that in which the romantics believed, which can be counted upon, even in a hypothetical infinitely protracted lifetime, to create all the beauty that there is.

Thus beauty admits of a succession of ever-better approximations, but is never fully attainable. Beauty and irrationality are often linked. Is it coincidental that the first example of such a notion of something approximable but never attainable in a finite process is called an "irrational" number?

Myhill is bold enough to speculate as follows: "The analogue of Gödel's theorem for aesthetics would therefore be: There is no school of art which permits the production of all beauty and excludes the production of all ugliness." To each coin there are two sides; and the obverse side of beauty is ugliness. By a rather ironic coincidence, the complementary set to a productive (or prospective) set is called, in the jargon of mathematical logic, *creative.* It must be admitted that it would take a stupendously brilliant, if perverse, sort of creativity to produce all possible ugly objects.

If we see the aim of art as the production of all possible objects of beauty (which is doubtless an oversimplification, but let us adopt that view nonetheless), then each individual artist contributes objects in a particular style. That style is a product of the artist's heredity and formation, and becomes a hallmark. To the extent of having an individual style, any artist is sphexish—trapped within invisible, intangible, but inescapable boundaries of mental space. But that is nothing to lament. Artists in groups form movements or schools or periods, and what limits one artist need not limit another. Thus, by the fact that its boundaries are wider, a school is less sphexish—more conscious—than any of its members.

But even the collective movement of a school of art has its limits, shows its finitude, after a period of time. It begins to wind down, to lose fertility, to stagnate. And a new school begins to form. What no individual can make out clearly is perhaps seen collectively, on the level of a society. Thus art progresses towards an ever wider vision of beauty—a "prospective" vision of beauty—by a series of repeated "diagonalizations": processes of recognizing and breaking out of ruts. As I like to put it, this is the process of *jootsing* (jumping out of the system) to ever wider worlds.

This endless jootsing is a process whose totality (so says Gödel) cannot be formalized, either in a computer or in any finite brain or set of brains. Thus one need not fear that the mechanization of creativity, if ever it comes about, will mark the end of art. Quite the contrary: It is a day to look forward to, for on that day our eyes will open—as will those of computers, to be sure —onto whole new worlds of beauty. It will be a happy day when, hand in hand with our new computer friends, we take an unanalyzable leap out of the system and go get some more punch.

Post Scriptum.

Do you know the Saint-Saëns Violin Concerto No. 3? Its middle movement happens to be based on a ravishingly beautiful melody—long, sinuous, flowing, lyrical. I suggest you get a hold of it and listen to it! Where do such melodies come from? Did they always exist? Are some people just lucky to have picked them up, these pretty pebbles lying on the musical beach?

Well, I hardly want to get into the discovery-invention-existence quagmire here. I have my own opinions, to be sure, but what I am more concerned with is where such inspiration comes from. One can point with a fair degree of objectivity to certain composers as being the most melodically gifted. These names come to my mind, for instance: Chopin, Rachmaninoff, Saint-Saëns, Tchaikovsky, Brahms, Bach, Mendelssohn, Handel, Puccini—and, switching gears somewhat, Cole Porter, Richard Rodgers, Jerome Kern, and George Gershwin. Obviously there are others. Some people undoubtedly would strike some off this list and would suggest others—

perhaps Schubert, Dvořák, Prokofiev, Scott Joplin, Fats Waller, Frederick Loewe, Kurt Weill, the Beatles, Carole King . . . It's hard to draw the line.

The main point is that certain rare people seem to be able to tap into some magic vein in which flow incredibly catchy patterns, deeply intoxicating to the human spirit. Leonard Bernstein once wrote a lively dialogue encatchily titled "Why Don't You Run Upstairs and Write a Nice Gershwin Tune?". In it, he talks about why that vein is so hard to tap. Bernstein should know, of course, since he too is one of the great melodic inventors of our time.

The problem is that melody invention, like every other art, looks so easy after the fact. In fact, in many ways it looks easier than creating other kinds of beauty, because melodies are such small, easily described structures. Making a beautiful turn on skis at least involves a *continuum* of possibilities, whereas a melody usually involves a very restricted, discrete alphabet (the notes within a two-octave range or so), and isn't even very long!

It is tempting, therefore, to imagine that good melodies are producible from some sort of recipe or mathematical formula, or, what comes to nearly the same thing, to think that the *amount* of beauty in a melody could be measured by some sort of machine, just as the amount of radioactivity in a sample of ore can be measured by a scintillation counter. You would stick your proposed string of notes into a machine and out would come a number called its "CQ" ("catchiness quotient").

If you doubt that the very idea of such a number is coherent, just remember that attached to every piece of existent music there really *is* a measure of its catchiness—namely, how often it actually is listened to, at the present time. Pieces can be rank-ordered according to this very cold, linear measure. This is not to suggest that the top piece is the best, but only to point out that the idea of a single, one-dimensional "catchiness index" applying to every possible string of notes is by no means absurd. Admittedly, under the present circumstances, it seems to take an entire society of millions of people to calculate the value for any string of notes, but could all that not be simulated? Perhaps the catchiness-quotient machine could be built to accept a set of parameters characterizing the target culture and its general musical mood at the time, and then it would predict how the given tune would fare in the given society under the specified musical circumstances. Is that not an engaging notion?

Are the musical receptivities of a culture truly characterizable in purely mathematical terms relating only to the syntactical structures of melodies? Ultimately, of course, the answer has *got* to be "yes", if by "syntactical structures" you mean structures whose recognition might require bringing in arbitrary amounts of external information. Sufficiently deep syntactic probing is tantamount to semantic probing, a motto from Chapter 1's *P.S.* The question is, then, just how complex a "syntax machine" that creates, or at least measures, melodic beauty would be. (Let's assume that it contains adjustable parameters for culture and mood.) Need it be as complex as a human society or a human brain? Can wonderful, lyrical, sinuous, and rapturous melodies come pouring out of a black box that can do nothing but that? Readers of *Gödel, Escher, Bach* (especially pages 676–680) might recall that I am extremely skeptical on that score. Yet how solid is the ground I am standing on? Could music not yield to brute computational power as swiftly as chess skill has (something which, in the same passages in *GEB*, I also was very skeptical about)?

* * *

It is funny how certain fads catch on, seemingly for no reason, while other things die, again for no clear reason. We all laugh at the Edsel today—yet what exactly is there to laugh at, except the fact that it did so poorly? What exactly was *wrong* with the Edsel? What is wrong with those thousands upon thousands of melodies that are composed every year and go nowhere? What made Michael Jackson and Pachelbel's simple Canon all the rage? Why did the typeface Helvetica catch on like wildfire when it was first invented, when a dozen extremely similar ones died on the vine? Why did the typographical gimmick of symmetrically capitalizing both the first and the last letter of a word or title, as in

GATEWAY
INN

PRINCE
SPAGHETTI

become a sudden vogue about four years ago?

Why is it now faddish to write run-on words such as "Intelligenetics" or "PEOPLExpress"? What makes words like "Da-glo", "Turbomatic", and "Rayon" seem slightly dated? Why is "Qantas" still modern-sounding? What is poor about brand names like "Luggo" and "Flimp"? Why are 'x's now so popular in brand names? And yet why would "Goxie" be a weak name compared with, say, "Exigo" or "Xigeo"? Why are the ordinary-seeming names that nasal-voiced comedians Bob and Ray come up with—for example, "Wally Ballou", "Hudley Pierce", "Bodin Pardew", and "John W. Norbis"—apt to evoke snickers? How come Norma Jean Baker changed her name to "Marilyn Monroe"? Why would it not do for a movie star to be named "Arnold Wilberforce"? Why is the name "Tiffany" popular today, and why was "Lisa" so popular a few years earlier? Is something wrong with "Agnes", "Edna", or "Thelma"? With "Clyde", "Lance", or "Bartholomew"? Mere length certainly cannot be the answer (think of "Elizabeth"). Nor can the sound, in any simple sense. (Why is "Lance" bad if "Vance" is okay?)

All this may seem a far, far cry from sphexishness and self-watching computers and brains. But what I am getting at is the unbelievable number of forces and factors that interact in our unconscious processing of even very tiny structures composed of discrete parts, such as words and names only a few letters long, let alone melodies several dozen notes long. Most of us could not put our finger on the answers to any of these questions. In fact, nobody could really answer these questions definitively. If we are going to try to get machines to do the subtlest of cognitive tasks, we had jolly well better be able to explain how mere words are appealing or repelling!

* * *

There are currently some efforts in artificial intelligence to imbue programs with a certain type of introspective capacity. Such a capacity is usually termed "reflection", a self-explanatory name that harks back to mathematical logic. A formal system is said to be capable of reflection if it can reason about itself. Gödel was the first person to discuss such things in detail. Nowadays reflective systems are the bread and butter of many a logician. However, computer modeling of logic is just now reaching the point where reflection is being seriously explored.

The idea is very enticing, but I think it has less to do with genuine progress in AI than it does with progress in elegant formal systems. It all has to do with one's ultimate view of what thought is. If you believe that thought is intimately tied up with some strict notion of truth and reasoning, and that exquisitely honed deductive capacities are the centerpiece of mentality, then you will naturally be drawn toward reflective reasoning systems. If, on the other hand, you believe, as I do, that reasoning is a far, far cry from the core of thought, then you will not be too inclined to jump toward such systems.

One way of looking at things is this. Imagine you have a set of rules that are supposed to capture the way people think in some domain—say that of melody composition. Now you try them out, and you find that most of the time they fail for complex reasons, but reasons that you have some intuitions about. How should you proceed now? There are two main rival avenues, the way I see it.

One avenue says, "Add meta-rules! Then add meta-meta-rules! Then . . . *ad infinitum*!" This might be called the "meta-meta" school of AI. The strategy is to improve the performance of a given set of rules by having higher-order meta-rules that help determine when and how to apply the ordinary rules. And this process knows no bounds, even to the point that one can formalize the progression from one level to its meta-level, so that in principle, an infinite number of meta-levels now are "there" to be consulted if needed.

The alternate avenue is to sidestep the topless tower of bureaucracies and meta-bureaucracies *above* by making rule-like behavior emerge out of a multi-level bubbling broth of activity *below*. This means that you give up the idea of trying to explicitly tell the system as a whole how to run itself. Instead, you content yourself with defining explicit micro-behaviors that will interact in vast numbers, and then you just let them go, carefully watching what ensues and noting what you like and what you don't like. After the run, you theorize about what might have made the system's top-level behavior more closely resemble your ultimate goals, and you go back and tinker around with the micro-elements whose micro-behavior you have explicit control over, using your best guess as to what sorts of changes will improve overall performance. Then you run the system again.

I remember a long time ago seeing a television show—perhaps you have seen it, too—in which someone set up a bathtub full of spring-loaded mousetraps holding ping-pong balls. Then they threw a single ping-pong ball in, and WHAM! The whole thing exploded madly, in parallel chain reactions. It was all over in a few seconds, but you can imagine running a film of it in slow motion. There are numerous large-scale features of the explosion that one could aim at creating, such as how long the pop takes, how high the average ping-pong ball flies, what the envelope of the flying balls looks like, and so on. If there were more types of micro-element and their interactions were more variegated, then you can imagine how multi-dimensional the system's macrobehavior would be, and how hard it would be to predict even its most basic features.

Yet when certain vast ensembles grow sufficiently big, the statistical principle called "the law of large numbers" sets in, in essence guaranteeing that there will be so much cancellation in the chaos that ultimately, a kind of order will emerge. It is for reasons like this that the National Safety Council can predict fairly accurately how many deaths there will be on a Labor Day weekend, even though they have no idea where any particular one will occur. Somehow, amazingly, the drivers cooperate and produce just about the predicted number each time, usually even on the state-by-state level, although less accurately.

The difference between such statistically emergent macrobehavior and rigidly constrained macrobehavior is best made by contrasting the mousetrap system with a huge domino-chain network, involving branching and rejoining paths, paths that climb hills and go back down, anything you can imagine as long as it's entirely self-determined (*i.e.*, no unanticipated external events start chains falling). In this kind of system, you know how everything is going to work beforehand. It's true that you may not be able to predict which of two "rival" pathways will reach a certain point first, but this kind of unpredictability is not nearly as hard to correct as that of the mousetrap system. If on one run the result is not what you want, you can just set it up again the same way, change some specific region, and you know what will happen. You can *program* this kind of system, but you cannot program a statistical system in the same sense. You can only tailor its micro-elements, and then release them and see what happens.

Which approach to mind is superior? Is the mind more like a fancy system of domino chains or a bathtub full of spring-loaded mousetraps? I'm betting on the latter. More will be found on this topic in Chapters 25 and 26 and their postscripts.

* * *

I received a letter from Thomas P. Laubert, in which he expressed considerable perplexity over a paragraph he had come across containing the following sentence: "Experience had taught the du Pont engineers to provide . . . . flexibility in the design, wherever possible, to meet unforeseen problems that were sure to arise." Laubert mused: "But if the nature of the problems was unforeseen, then what parameters were used to determine these built-in flexibilities?" Another reader, whose letter I have unfortunately misplaced, brought up a similar point about engineering. What I remember vividly is his term "UNK-UNK"s—meaning the *unknown unknowns* that plague all complex systems. He was asking, rather skeptically, as I recall, how one can ever hope to build a system that anticipates all possible problems.

These simple-seeming questions hit the nail on the head. An intelligence is, by definition, a system supposed to be able to deal with the unpredictable. But how can any set of rules "frozen" into a machine's design do that? Doesn't the very fact of being frozen make any foreordained system/program/machine/organism vulnerable in some way that actually follows from the rules themselves? This, of course, is the Gödelian point that J. R. Lucas was trying to make in the article I quoted from in the column. And the only satisfactory answer that I can see is to admit that, yes, all intelligences are indeed vulnerable—including biological ones, and that means people no less than *Sphex* wasps. Natural selection has looked favorably upon organisms with highly abstract kinds of vulnerability, highly abstract kinds of sphexishness. And so for the time being, humans are doing all right. But as for there being a fixed recipe that would allow an organism to cope with all the curves that the universe at large might throw at it, that is a vain and crazy hope!