26

Waking Up
from the Boolean Dream,
or, Subcognition as Computation

July, 1982

Introduction

THE philosopher John Searle has recently made quite a stir in the cognitive-science and philosophy-of-mind circles with his celebrated article "Minds, Brains, and Programs", in which he puts forth his "Chinese room" thought experiment. Its purpose is to reveal as illusory the aims of artificial intelligence, and particularly to discredit what he labels *strong AI*—the belief that a programmed computer can, in principle, be conscious. Various synonymous phrases could be substituted for "be conscious" here, such as:

* *think*;
* *have a soul* (in a humanistic rather than a religious sense);
* *have an inner life*;
* *have semantics* (as distinguished from "mere syntax");
* *have content* (as distinguished from "mere form");
* *have intentionality*;
* *be something it is like something to be* (a weird phrase due to T. Nagel);
* *have personhood*;

and others. Each of these phrases has its own peculiar set of connotations and imagery attached to it, as well as its own history and proponents. For our purposes, however, we shall consider them all as equivalent, and lump them all together, so that the claim of strong AI now becomes very strong indeed.

SPIRIT & SUBSTRATE

January 20, 1981

In the natural environment, one can never achieve a zero error rate. So nature, in her delicate brilliance, finds a way to make errors work. It's sort of like this: outrageous errors are deadly, but slight errors keep things loose enough for the new to appear. If the steps of the error can be kept small enough, a near-perfect approximation can gradually be built.

It's hard for me to put into words, but qualities like "quiveriness" and "vulnerability" come to mind when I think of creativity. Maybe another image is something like the doping used in the manufacture of semiconductors. Introducing small bits of something that doesn't belong there can dramatically change the structure of a large-scale array. It's not really randomness that is the vital ingredient in creativity, but rather the slight tint of the subtle truths that creativity finds that changes the random from white to pink. Michelangelo, standing before the marble that was to become David, had in his mind the image of a young man, but also allowed himself to be invaded by the block of marble, the whole universe itself. Creativity requires a sense of smell, a palate to taste the scents that make brilliance.

"Chance favors the prepared mind." —Louis Pasteur. I like that quote. All life feeds upon the random. Creativity is simply the *haute cuisine*.

Little ripples that exist or rather persist long enough to be observed are the lowest forms that possess what could be called personality. Rocks and snow tend to have very little personality because the ripples are very small-scale and dominated by randomness. Smoke has more. Indeed I think that's why Bach and Magritte liked pipe smoke and why I, too, confess to enjoy the habit.

A puff of smoke is sort of like a cousin of ours. Little eddies, the loosely coupled systems, shear and spin, and we can observe the gentle drift of the whole ensemble. Bubbling streams have this quality, as do breaking ocean waves. The boiling and roiling, the little pieces, each with their own life, each wavelet connected, interacting, and yet participating in the whole.

*   *   *

February 23, 1981

BUTTERFLIES

The best thoughts are the most delicate,
fastest, trickiest to capture.
Lepidoptera so different on the wing,
than when caught, killed,
and proudly displayed.

*   *   *

On April 10, 1983, Randy Read took his own life. I don't know why. Perhaps these musings, dancing and sparking in the neurons of a few thousand readers out there, will keep alive, in scattered form, a tiny piece of his soul.

Waking Up from the Boolean Dream

At the same time, various AI workers have been developing their own philosophies of what AI is, and have developed some useful terms and slogans to describe their endeavor. Some of them are: "information processing", "cognition as computation", "physical symbol system", "symbol manipulation", "expert system", and "knowledge engineering". There is some confusion as to what words like "symbol" and "cognition" actually mean, just as there is some confusion as to what words like "semantics" and "syntax" mean.

It is the purpose of this article to try to delve into the meanings of such elusive terms, and at the same time to shed some light on the views of Searle, on the one hand, and Allen Newell and Herbert Simon, on the other hand —visible AI pioneers who are responsible for several of the terms in the previous paragraph. The thoughts expressed herein were originally triggered by a paper called "Artificial Intelligence: Cognition as Computation", by Avron Barr. However, they can be read completely independently of that paper.

The questions are obviously not trivial, and certainly not resolvable in a single article. Most of the ideas in this article, in fact, were stated earlier and more fully in my book *Gödel, Escher, Bach: an Eternal Golden Braid*. However, it seems worthwhile to extract a certain stream of ideas from that book and to enrich it with some more recent musings and examples, even if the underlying philosophy remains entirely the same. In order to do justice to these complex ideas, many topics must be interwoven, and they include the nature of symbols, meaning, thinking, perception, cognition, and so on. That explains why this article is not three pages long.

Cognition versus Perception: The 100-millisecond Dividing Line

In Barr's original paper, AI is characterized repeatedly by the phrase "information-processing model of cognition". Although when I first heard that phrase years ago, I tended to accept it as defining the nature of AI, something has gradually come to bother me about it, and I would like to try to articulate that here. Now what's in a word? What's to object to here? I won't attempt to say what's wrong with the phrase so much as try to show what I disagree with in the ideas of those who have promoted it; then perhaps the phrase's connotations will float up to the surface so that other people can see why I am uneasy with it.

I think the disagreement can be put in its sharpest relief in the following way. In 1980, Simon delivered a lecture that I attended (the Procter Award Lecture for the Sigma Xi annual meeting in San Diego), and in it he declared (and I believe I am quoting him nearly verbatim):

Everything of interest in cognition happens above the 100-millisecond level— the time it takes you to recognize your mother.

Well, our disagreement is simple; namely, I take exactly the opposite viewpoint:

Everything of interest in cognition happens below the 100-millisecond level— the time it takes you to recognize your mother.

To me, the major question of AI is this: "What in the world is going on to enable you to convert from 100,000,000 retinal dots into one single word 'mother' in one tenth of a second?" Perception is where it's at!

The Problem of Letterforms: A Test Case for AI

The problem of intelligence, as I see it, is to understand the fluid nature of mental categories, to understand the invariant cores of percepts such as your mother's face, to understand the strangely flexible yet strong boundaries of concepts such as "chair" or the letter 'a'. Years ago, long before computers, Wittgenstein had already recognized the centrality of such questions, in his celebrated discussion of the nonpindownability of the meaning of the word "game". To emphasize this and make the point as starkly as I can, I hereby make the following claim:

The central problem of AI is the question: *What is the letter 'a'?*

Donald Knuth, on hearing me make this claim once, appended, "And what is the letter 'i'?"—an amendment that I gladly accept. In fact, perhaps the best version would be this:

The central problem of AI is: *What are 'a' and 'i'?*

By making these claims, I am suggesting that, for any program to handle letterforms with the flexibility that human beings do, it would have to possess full-scale general intelligence.

Many people in AI might protest, pointing out that there already exist programs that have achieved expert-level performance in specialized domains without needing general intelligence. Why should letterforms be any different? My answer would be that specialized domains tend to obscure, rather than clarify, the distinction between strengths and weaknesses of a program. A familiar domain such as letterforms provides much more of an acid test.

To me, it is strange that AI has said so little about this classic problem. To be sure, some work has been done. There are a few groups with interest in letters, but there has been no all-out effort to deal with this quintessential problem of pattern recognition. Since letterform understanding is currently an important target of my own research project in AI, I would like to take a moment and explain why I see it as contrasting so highly with domains at the other end of the "expertise spectrum".

Each letter of the alphabet comes in literally thousands of different "official" versions (typefaces), not to mention millions, billions, trillions, of "unofficial" versions (those handwritten ones that you and I and everyone else produces all the time). There thus arises the obvious question: "How are all 'a's like each other?" The goal of an AI project would be, of course, to give an exact answer in computational terms. However, even taking advantage of the vagueness of ordinary language, one is hard put to find a satisfactory intuitive answer, because we simply come up with phrases such as "They all have the same shape." Clearly, the whole problem is that they *don't* have the same shape. And it does not help to change "shape" to "form", or to tack on phrases such as "basically", "essentially", or "at a conceptual level".

There is also the less obvious question: "How are all the various letters in a single typeface related to each other?" This is a grand analogy problem if ever there were an analogy problem. One is asking for a 'b' that is to the abstract notion of 'b'-ness as a given 'a' is to the abstract notion of 'a'-ness. You have to take the qualities of a given 'a' and, so to speak, "hold them loosely in the hand", as you see how they "slip" into variants of themselves as you try to carry them over to another letter. Here is the very hingepoint of thought, the place where one thing slips into alternate, subjunctive, variations on itself. Here, that "thing" is a very abstract concept—namely, "the way that this particular shape manifests the abstract quality of being an 'a'". The problem of 'a' is thus intimately connected with the problems of 'b' through 'z', and with that of stylistic consistency.

The existence of optical character readers, such as the reading machines invented by Ray Kurzweil for blind people, might lead one to believe at first that the letter-recognition problem has been solved. If one considers the problem a little more carefully, however, one sees that the surface has barely been scratched. In truth, the way that most optical character recognition programs work is by a fancy kind of template matching, in which statistics are done to determine which character, out of a fixed repertoire of, say, 100 stored characters, is the "best match". This is about like assuming that the way I recognize my mother is by comparing the scene in front of me with stored memories of the appearances of tigers, cigarettes, hula hoops, gambling casinos, and can openers (and of course all other things in the world simultaneously), and somehow instantly coming up with the "best match".

The Human Mind and Its Ability to Recognize and Reproduce Forms

The problem of recognizing letters of the alphabet is no less deep than that of recognizing your mother, even if it might seem so, given that the number of Platonic prototype items is on the small side (26, if one ignores all characters but the lowercase alphabet). One can even narrow it down

further—to just a handful. As a matter of fact, Godfried Toussaint, editor of the pattern recognition papers for the *IEEE Transactions*, has said to me that he would like to put up a prize for the first program that could tell correctly, given twenty characters that people easily can identify, which ones are 'a's and which are 'b's. To carry out such a task, a program cannot just recognize that a shape is an 'a'; it has to see *how* that shape embodies 'a'-ness. And then, as a test of whether the program really knows its letters, it would have to carry "that style" over to the other letters of the alphabet. This is the goal of my research: To find out how to make letters slip in "similar ways to each other", so as to constitute a consistent artistic style in a typeface—or simply a consistent way of writing the alphabet.

By contrast, most AI work on vision pertains to such things as aerial reconnaissance or robot guidance programs. This would suggest that the basic problem of vision is to figure out how to recognize textures and how to mediate between two and three dimensions. But what about the fact that although we are all marvelous face-recognizers, practically none of us can draw a face at all well—even of someone we love? Most of us are flops at drawing even such simple things as pencils and hands and books. I personally have learned to recognize hundreds of Chinese characters (shapes that involve neither three dimensions nor textures) and yet, on trying to reproduce them from memory, find myself often drawing confused mixtures of characters, leaving out basic components, or worst of all, being unable to recall anything but the vaguest "feel" of the character and not being able to draw a single line.

Closer to home, most of us have read literally millions of, say, 'u's with serifs, yet practically none of us can draw a 'u' with serifs in the standard places, going in the standard directions. (This holds even more for the kind of 'g' you just read, but it is true for any letter of the alphabet.) I suspect that many people—perhaps most—are not even consciously aware of the fact that there are two different types of lowercase 'a' and of lowercase 'g', just as many people seem to have a very hard time drawing a distinction between lowercase and uppercase letters, and a few have a hard time telling letters drawn forward from letters drawn backward.

How can such a fantastic "recognition machine" as our brain be so terrible at rendition? Clearly there must be something very complex going on, enabling us to *accept* things as members of categories and to perceive *how* they are members of those categories, yet not enabling us to reproduce those things from memory. This is a deep mystery.

In his book *Pattern Recognition*, the late Mikhail Bongard, a creative and insightful Russian computer scientist, concludes with a series of 100 puzzles for a visual pattern recognizer, whether human, machine, or alien, and to my mind it is no accident that he caps his set off with letterforms. In other words, he works his way up to letterforms as being at the pinnacle of visual recognition ability. There exists no pattern recognition program in the world today that can come anywhere close to doing those Bongard problems. And yet, Barr cites Simon as writing the following statement:

The evidence for that commonality [between the information processes that are employed by such disparate systems as computers and human nervous systems] is now overwhelming, and the remaining questions about the boundaries of cognitive science have more to do with whether there also exist nontrivial commonalities with information processing in genetic systems than with whether men and machines both think. Wherever the boundary is drawn, there exists today a science of intelligent systems that extends beyond the limits of any single species.

I find it difficult to understand how Simon can believe this, in an era when computers still cannot do basic kinds of *subcognitive* acts (acts that we feel are unconscious, acts that underlie cognition).

In another lecture in 1979 (the opening lecture of the inaugural meeting of the Cognitive Science Society, also in San Diego), I recall Simon proclaiming that, despite much doubting by people not in the know, there is no longer any question as to whether computers can think. If he had meant that there should no longer be any question about whether machines may *eventually* become able to think, or about whether we humans are machines (in some abstract sense of the term), then I would be in accord with his statement. But after hearing and reading such statements over and over again, I don't think that's what he meant at all. I get the impression that Simon genuinely believes that today's machines are intelligent, and that they really do think (or perform "acts of cognition"—to use a bit of jargon that adds nothing to the meaning but makes it sound more scientific). I will come back to that shortly, since it is in essence the central bone of contention in this article, but first a few more remarks on AI domains.

Toy Domains, Technical Domains, Pure Science, and Engineering

There is in AI today a tendency toward flashy, splashy domains—that is, toward developing programs that can do such things as medical diagnosis, geological consultation (for oil prospecting), designing of experiments in molecular biology, molecular spectroscopy, configuring of large computer systems, designing of VLSI circuits, and on and on. Yet there is no program that has common sense; no program that learns things that it has not been explicitly taught how to learn; no program that can recover gracefully from its own errors. The "artificial expertise" programs that do exist are rigid, brittle, inflexible. Like chess programs, they may serve a useful intellectual or even practical purpose, but despite much fanfare, they are not shedding much light on human intelligence. Mostly, they are being developed simply because various agencies or industries fund them.

This does not follow the traditional pattern of basic science. That pattern is to try to isolate a phenomenon, to reduce it to its simplest possible manifestation. For Newton, this meant the falling apple and the moon; for Einstein, the thought experiment of the trains and lightning flashes and,

later, the falling elevator; for Mendel, it meant the peas; and so on. You don't tackle the messiest problems before you've tackled the simpler ones; you don't try to run before you can walk. Or, to use a metaphor based on physics, you don't try to tackle a world with friction before you've got a solid understanding of the frictionless world.

Why do AI people eschew "toy domains"? Once, about ten years back, the MIT "blocks world" was a very fashionable domain. Roberts and Guzmán and Waltz wrote programs that pulled visions of three-dimensional blocks out of two-dimensional television-screen dot matrices; Winston, building on their work, wrote a program that could recognize instantiations of certain concepts compounded from elementary blocks in that domain ("arch", "table", "house", and so on); Winograd wrote a program that could "converse" with a person about activities, plans, past events, and some structures in that circumscribed domain; Sussman wrote a program that could write and debug simple programs to carry out tasks in that domain, thus effecting a simple kind of learning. Why, then, did interest in this domain suddenly wane?

Surely no one could claim that the domain was exhausted. Every one of those programs exhibited glaring weaknesses and limitations and specializations. The domain was phenomenally far from being understood by a single, unified program. Here, then, was a nearly ideal domain for exploring what cognition truly is—and it was suddenly dropped. MIT was at one time doing truly basic research on intelligence, and then quit. Much basic research has been supplanted by large teams marketing what they vaunt as "knowledge engineering". Firmly grounded engineering is fine, but it seems to me that this type of engineering is not built upon the solid foundations of a science, but upon a number of recipes that have worked with some success in limited domains.

In my opinion, the proper choice of domain is the critical decision that an AI researcher makes, when beginning a project. If you choose to get involved in medical diagnosis at the expert level, then you are going to get mired in a host of technical problems that have nothing to do with how the mind works. The same goes for the other earlier-cited ponderous domains that current work in expert systems involves. By contrast, if you are in control of your own domain, and can tailor it and prune it so that you keep the essence of the problem while getting rid of extraneous features, then you stand a chance of discovering something fundamental.

Early programs on the nature of analogy (Evans), sequence extrapolation (Simon and Kotovsky, among others), and so on, were moving in the right direction. But then, somehow, it became a common notion that these problems had been solved. Simply because Evans had made a program that could do some very restricted types of visual analogy problem "as well as a high school student", many people thought the book was closed. However, one need only look at Bongard's 100 to see how hopelessly far we are from dealing with analogies. One need only look at any collection

of typefaces (look at any magazine's advertisements for a vast variety) to see how enormously far we are from understanding letterforms. As I claimed earlier, letterforms are probably the quintessential problem of pattern recognition. It is both baffling and disturbing to me to see so many people working on imitating cognitive functions at the highest level of sophistication when their programs cannot carry out cognitive functions at much lower levels of sophistication.

AI and the True Nature of Intelligence

There are some notable exceptions. The Schank group at Yale, whose original goal was to develop a program that could understand natural language, has been forced to "retreat", and to devote at least a bit of its attention to the organization of memory, which is certainly at the crux of cognition (because it is part of subcognition, incidentally)—and the group has gracefully accommodated this shift of focus. I will not be at all surprised, however, if eventually the group is forced into yet further retreats—in fact, all the way back to Bongard problems or the like. Why? Simply because their work (on such things as how to discover what "adage" accurately captures the "essence" of a story or episode) already has led them into the deep waters of abstraction, perception, and classification. These are the issues that Bongard problems illustrate so perfectly. Bongard problems are idealized ("frictionless") versions of these critical questions.

It is interesting that Bongard problems are in actuality nothing other than a well-worked-out set of typical IQ-test problems, the kind that Terman and Binet first invented 50 or more years ago. Over the years, many other less talented people have invented similar visual puzzles that had the unfortunate property of being filled with ambiguity and multiple answers. This (among other things) has given IQ tests a bad name. Whether or not IQ is a valid concept, however, there can be little question that the original insight of Terman and Binet—that carefully constructed simple visual analogy problems probe close to the core mechanisms of intelligence—is correct. Perhaps the political climate created a kind of knee-jerk reflex in many cognitive scientists to shy away from anything that smacked of IQ tests, since issues of cultural bias and racism began raising their ugly heads. But one need not be so Pavlovian as to jump whenever a visual analogy problem is placed in front of one. In any case, it will be good when AI people are finally driven back to looking at the insights of people working in the 1920's, such as Wittgenstein and his "games", Koehler and Koffka and Wertheimer and their "gestalts", and Terman and Binet and their IQ-test problems.

I was saying that some AI groups seem to be less afraid of "toy domains", or more accurately put, they seem to be less afraid of stripping down their domain in successive steps, to isolate the core issues of intelligence that it involves. Aside from the Schank group, N. Sridharan and Thorne McCarty

at Rutgers have been doing some very interesting work on "prototype deformation", which, although it springs from work in legal reasoning in the quite messy real-world domain of corporate tax law, has been abstracted into a form in which it is perhaps more like a toy domain (or, perhaps less pejorative-sounding, an "idealized domain") than at first would appear.

At the University of California at San Diego, a group led by psychologist Donald Norman has been for years doing work on understanding errors, such as grammatical slips, typing errors, and errors in everyday physical actions, for the insights it may offer into the underlying (subcognitive) mechanisms. (For example, one of Norman's students unbuckled his watch instead of his seatbelt when he drove into his driveway. What an amazing mental slippage!) A group led by Norman and his colleague David Rumelhart has developed a radically different model of cognition largely based on parallel subcognitive events termed "schema activations". The reason that this work is so different in flavor from mainstream AI work is twofold: firstly, these are psychologists who are studying genuine cognition in detail and who are concerned with reproducing it; and secondly, they are not afraid to let their vision of how the *mind* works be inspired by research and speculation about how the *brain* works.

Then there are those people who are working on various programs for perception, whether visual or auditory. One of the most interesting was Hearsay II, a speech-understanding program developed at Carnegie-Mellon, Simon's home. It is therefore very surprising to me that Simon, who surely was very aware of the wonderfully intricate and quite beautiful architecture of Hearsay II, could then make a comment indicating that perception and, in general, subcognitive (under 100 milliseconds) processes, "have no interest".

There are surely many other less publicized groups that are also working on humble domains and on pure problems of mind, but from looking at the proceedings of AI conferences one might get the impression that, indeed, computers must really be able to think these days, since after all, they are doing anything and everything cognitive—from ophthalmology to biology to chemistry to mathematics—even discovering scientific laws from looking at tables of numerical data, to mention one project ("Bacon") that Simon has been involved in. However, there's more to intelligence than meets the AI.

Expert Systems versus Human Fluidity

The problem is, AI programs are carrying out all these *cognitive activities* in the absence of any *subcognitive* activity. There is no substrate that corresponds to what goes on in the brain. There is no fluid *recognition* and recall and reminding. These programs have no common sense, little sense of similarity or repetition or pattern. They can perceive some patterns as long as they have been anticipated—and particularly, as long as the *place* where they will occur has been anticipated—but they cannot see patterns

where nobody told them explicitly to look. They do not learn at a high level of abstraction.

This style is in complete contrast to how people are. People perceive patterns anywhere and everywhere, without knowing in advance where to look. People learn automatically in all aspects of life. These are just facets of common sense. Common sense is not an "area of expertise", but a general—that is, domain-independent—capacity that has to do with fluidity in representation of concepts, an ability to sift what is important from what is not, an ability to find unanticipated analogical similarities between totally different concepts ("reminding", as Schank calls it). We have a long way to go before our programs exhibit this cognitive style.

Recognition of one's mother's face is still nearly as much of a mystery as it was 30 years ago. And what about such things as recognizing family resemblances between people, recognizing a "French" face, recognizing kindness or earnestness or slyness or harshness in a face? Even recognizing age—even sex!—these are fantastically difficult problems. As Donald Knuth has pointed out, we have written programs that can do wonderfully well at what people have to work very hard at doing consciously (e.g., doing integrals, playing chess, medical diagnosis, etc.)—but we have yet to write a program that remotely approaches our ability to do what we do *without* thinking or training—things like understanding a conversation partner with an accent at a loud cocktail party with music blaring in the background, while at the same time overhearing wisps of conversations in the far corner of the room. Or perhaps finding one's way through a forest on an overgrown trail. Or perhaps just doing some anagrams absentmindedly while washing the dishes.

Asking for a program that can discover new scientific laws without having a program that can, say, do anagrams, is like wanting to go to the moon without having the ability to find your way around town. I do not make the comparison idly. The level of performance that Simon and his colleague-Langley wish to achieve in Bacon is on the order of the greatest scientists. It seems they feel that they are but a step away from the mechanization of genius. After his Procter Lecture, Simon was asked by a member of the audience, "How many scientific lifetimes does a five-hour run of Bacon represent?" After a few hundred milliseconds of human information processing, he replied, "Probably not more than one." I don't disagree with that. However, I would have put it differently. I would have said, "Probably not more than one-millionth."

Anagrams and Epiphenomena

It's clear that I feel we're much further away from programs that do human-level scientific thinking than Simon does. Personally, I would just like to see a program that can do anagrams the way a person does. Why anagrams? Because they constitute a "toy domain" where some very significant subcognitive processes play the central role.

What I mean is this. When you look at a "Jumble" such as "telkin" in the newspaper, you immediately begin shifting around letters into tentative groups, making such stabs as "knitle", "klinte", "linket", "keltin", "tinkle" —and then you notice that indeed, "tinkle" is a word. The part of this process that I am interested in is the part that precedes the recognition of "tinkle" as a word. It's that part that involves experimentation, based only on the "style" or "feel" of English words—using intuitions about letter affinities, plausible clusters and their stabilities, syllable qualities, and so on. When you first read a Jumble in the newspaper, you play around, rearranging, regrouping, reshuffling, in complex ways that you have no control over. In fact, it feels as if you throw the letters up into the air separately, and when they come down, they have somehow magically "glommed" together in some English-like word! It's a marvelous feeling— and it is anything but cognitive, anything but conscious. (Yet, interestingly, *you* take credit for being good at anagrams, if you are good!)

It turns out that most literate people can handle jumbles (i.e., single-word anagrams) of five or six letters, sometimes seven or eight letters. With practice, maybe even ten or twelve. But beyond that, it gets very hard to keep the letters in your head. It is especially hard if there are repeated letters, since one tends to get confused about which letters there are multiple copies of. (In one case, I rearranged the letters "dinnal" into "nadlid"—incorrectly. You can try "raregarden", if you dare.) Now in one sense, the fact that the problem gets harder and harder with more and more letters is hardly surprising. It is obviously related to the famous "7 plus or minus 2" figure that psychologist George A. Miller first reported in connection with short-term memory capacity. But there are different ways of interpreting such a connection.

One way to think that this might come about is to assume that concepts for the individual letters get "activated" and then interact. When too many get activated simultaneously, then you get swamped with combinations and you drop some letters and make too many of others, and so on. This view would say that you simply encounter an explosion of connections, and your system gets overloaded. It does not postulate any explicit "storage location" in memory—a fixed set of registers or data structures—in which letters get placed and then shoved around. In this model, short-term memory (and its associated "magic number") is an *epiphenomenon* (or "innocently emergent" phenomenon, as Daniel Dennett calls it), by which I mean it is a consequence that emerges out of the design of the system, a product of many interacting factors, something that was not necessarily known, predictable, or even anticipated to emerge at all. This is the view that I advocate.

A contrasting view might be to build a model of cognition in which you have an explicit structure called "short-term memory", containing about seven (or five, or nine) "slots" into which certain data structures can be fitted, and when it is full, well, then it is full and you have to wait until an empty slot opens up. This is one approach that has been followed by

and associates in work on production systems. The problem with this approach is that it takes something that clearly is a very complex consequence of underlying mechanisms and simply plugs it in as an explicit structure, bypassing the question of what those underlying mechanisms might be. It is difficult for me to believe that any model of cognition based on such a "bypass" could be an accurate model.

When a computer's operating system begins thrashing (*i.e.*, bogging down in its timesharing performance) at around 35 users, do you go find the systems programmer and say, "Hey, go raise the thrashing-number in memory from 35 to 60, okay?"? No, you don't. It wouldn't make any sense. This particular value of 35 is not stored in some local spot in the computer's memory where it can be easily accessed and modified. In that way, it is very different from, say, a student's grade in a university's administrative data base, or a letter in a word in an article you're writing on your home computer. That number 35 emerges dynamically from a host of strategic decisions made by the designers of the operating system and the computer's hardware, and so on. It is not available for (twiddling. There is no "thrashing-threshold dial" to crank on an operating system, unfortunately.

Why should there be a "short-term-memory-size" dial on an intelligence? Why should 7 be a magic number built into the system explicitly from the start? If the size of short-term memory really were explicitly stored in our genes, then surely it would take only a simple mutation to reset the "dial" at 8 or 9 or 50, so that intelligence would evolve at ever-increasing rates. I doubt that AI people think that this is even remotely close to the truth; and yet they sometimes act as if it made sense to assume it is a close approximation to the truth.

It is standard practice for AI people to bypass epiphenomena ("collective phenomena", if you prefer) by simply installing structures that mimic the superficial features of those epiphenomena. (Such mimics are the "shadows" of genuine cognitive acts, as John Searle calls them in his paper cited above.) The expectation—or at least the hope—is for tremendous performance to issue forth; yet the systems lack the complex underpinning necessary.

The anagrams problem is one that exemplifies mechanisms of thought that AI people have not explored. How do those letters swirl among one another, fluidly and tentatively making and breaking alliances? Clomming together, then coming apart, almost like little biological objects in a cell. AI people have not paid much attention to such problems as anagrams. Perhaps they would say that the problem is "already solved". After all, a virtuoso programmer has made a program print out all possible words that anagrammize into other words in English. Or perhaps they would point out that in principle you can do an "alphabetize" followed by a "hash" and thereby retrieve, from any given set of letters, all the words they anagrammize into. Well, this is all fine and dandy, but it is really beside the point. It is merely a show of brute force, and has nothing to contribute to

our understanding of how we actually do anagrams ourselves, just as most chess programs have absolutely nothing to say about how chess masters play (as de Groot, and later, Simon and coworkers have pointed out).

Is the domain of anagrams simply a trivial, silly, "toy" domain? Or is it serious? I maintain that it is a far purer, far more interesting domain than many of the complex real-world domains of the expert systems, precisely because it is so playful, so unconscious, so enjoyable, for people. It is obviously more related to creativity and spontaneity than it is to logical derivations, but that does not make it—or the mode of thinking that it represents—any less worthy of attention. In fact, because it epitomizes the unconscious mode of thought, I think it more worthy of attention.

In short, it seems to me that something fundamental is missing in the orthodox AI "information-processing" model of cognition, and that is some sort of substrate from which intelligence emerges as an epiphenomenon. Most AI people do not want to tackle that kind of underpinning work. Could it be that they really believe that machines already can think, already have concepts, already can do analogies? It seems that a large camp of AI people really do believe these things.

Not Cognition, But Subcognition, Is Computational

Such beliefs arise, in my opinion, from a confusion of levels, exemplified by the title of Barr's paper: "Cognition as Computation". Am I really computing when I think? Admittedly, my neurons may be performing sums in an analog way, but does this pseudo-arithmetical hardware mean that the epiphenomena themselves are also doing arithmetic, or should be—or even *can be*—described in conventional computer-science terminology? Does the fact that taxis stop at red lights mean that traffic jams stop at red lights? One should not confuse the properties of objects with the properties of statistical ensembles of those objects. In this analogy, traffic jams play the role of thoughts and taxis play the role of neurons or neuron-firings. It is not meant to be a deep analogy, only one that emphasizes that what you see at the top level need not have anything to do with the underlying swarm of activities bringing it into existence. In particular, *something can be computational at one level, but not at another level.*

Yet many AI people, despite considerable sophistication in thinking about a given system at different levels, still seem to miss this. Most AI work goes into efforts to build rational thought ("cognition") out of smaller rational thoughts (elementary steps of deduction, for instance, or elementary motions in a tree). It comes down to thinking that what we see at the top level of our minds—our ability to think—comes out of rational "information-processing" activity, with no deeper levels below that.

Many interesting ideas, in fact, have been inspired by this hope. I find much of the work in AI to be fascinating and provocative, yet somehow I feel dissatisfied with the overall trend. For instance, there are some people who believe that the ultimate solution to AI lies in getting better and better

Passive Symbols and Formal Rules

One way to explore this disagreement is to look at some of the ways that Simon and Newell express themselves about "symbols".

At the root of intelligence are symbols, with their denotative power and their susceptibility to manipulation. And symbols can be manufactured of almost anything that can be arranged and patterned and combined. Intelligence is mind implemented by any patternable kind of matter.

From this quotation and others, one can see that to Simon and Newell, a *symbol* seems to be any token, any character inside a computer that has an ASCII code (a standard but arbitrarily assigned sequence of seven bits). To me, by contrast, "symbol" connotes something with representational power. To them (if I am not mistaken), it would be fine to call a bit (inside a computer) or a neuron-firing a "symbol". However, I cannot feel comfortable with that usage of the term.

To me, the crux of the word "symbol" is its connection with the verb "to symbolize", which means "to denote", "to represent", "to stand for", and so on. Now, in the quote above, Simon refers to the "denotative power" of symbols—yet elsewhere in his paper, Barr quotes Simon as saying that thought is "the manipulation of formal tokens". It is not clear to me which side of the fence Simon and Newell really are on.

It takes an immense amount of richness for something to represent something else. The letter 'I' does not in and of itself stand for the person I am, or for the concept of selfhood. That quality comes to it from the way that the word behaves in the totality of the English language. It comes from a massively complex set of usages and patterns and regularities, ones that are regular enough for babies to be able to detect so that they too eventually come to say 'I' to talk about themselves.

Formal tokens such as 'I' or "hamburger" are in themselves empty. They do not denote. Nor can they be made to denote in the full, rich, intuitive sense of the term by having them obey some rules. You can't simply push around some Pnames of Lisp atoms according to complex rules and hope to come out with genuine thought or understanding. (This, by the way, is probably a charitable way to interpret John Searle's point in his above-mentioned paper—namely, as a rebellion against claims that programs that can manipulate tokens such as "John", "ate", "a", "hamburger" actually have understanding. Manipulation of empty tokens is not enough to create understanding—although it is enough to imbue them with meaning in a *limited* sense of the term, as I stress in my book *Gödel, Escher, Bach*—particularly in Chapters II through VI.)

---

theorem-proving mechanisms in some predicate calculus. They have developed extremely efficient and novel ways of thinking about logic. Some people—Simon and Newell particularly—have argued that the ultimate solution lies in getting more and more efficient ways of searching a vast space of possibilities. (They refer to "selective heuristic search" as the key mechanism of intelligence.) Again, many interesting discoveries have come out of this.

Then there are others who think that the key to thought involves making some complex language in which pattern matching or backtracking or inheritance or planning or reflective logic is easily carried out. Now admittedly, such systems, when developed, are good for solving a large class of problems, exemplified by such AI chestnuts as the missionary-and-cannibals problem, cryptarithmetic problems, retrograde chess problems, and many other specialized sorts of basically logical analysis. However, these kinds of techniques of building small logical components up to make large logical structures have not proven good for such things as recognizing your mother, or for drawing the alphabet in a novel and pleasing way.

One group of AI people who seem to have a different attitude consists of those who are working on problems of perception and recognition. There, the idea of coordinating many parallel processes is important, as is the idea that pieces of evidence can add up in a self-reinforcing way, so as to bring about the locking-in of a hypothesis that no one of the pieces of evidence could on its own justify. It is not easy to describe the flavor of this kind of program architecture without going into multiple technical details. However, it is very different in flavor from ones operating in a world where everything comes clean and precategorized—where everything is specified in advance: "There are three missionaries and three cannibals and one boat and one river and ..." which is immediately turned into a predicate-calculus statement or a frame representation, ready to be manipulated by an "inference engine". The missing link seems to be the one between perception and cognition, which I would rephrase as the link between subcognition and cognition, that gap between the sub-100-millisecond world and the super-100-millisecond world.

Earlier, I mentioned the brain and referred to the "neural substrate" of cognition. Although I am not pressing for a neurophysiological approach to AI, I am unlike many AI people in that I believe that any AI model eventually has to converge to brainlike hardware, or at least to an architecture that at some level of abstraction is "isomorphic" to brain architecture (also at some level of abstraction). This may sound empty, since that level could be anywhere, but I believe that the level at which the isomorphism must apply will turn out to be considerably lower than (I think) most AI people believe. This disagreement is intimately connected to the question of whether cognition should or should not be described as "computation".

Active Symbols and the Ant Colony Metaphor

So what is enough? What am I advocating? What do I mean by "symbol"? I gave an exposition of my concept of *active symbols* in Chapters XI and XII of *Gödel, Escher, Bach*. However, the notion was first presented in the dialogue "Prelude . . . Ant Fugue" in that book, which revolved about a hypothetical conscious ant colony. The purpose of the discussion was not to speculate about whether ant colonies are conscious or not, but to set up an extended metaphor for brain activity—a framework in which to discuss the relationship between "holistic", or collective, phenomena, and the microscopic events that make them up.

One of the ideas that inspired the dialogue has been stated by E. O. Wilson in his book *The Insect Societies* this way: "Mass communication is defined as the transfer, among groups, of information that a single individual could not pass to another." One has to imagine teams of ants cooperating on tasks, and information passing from team to team that no ant is aware of (if ants indeed are "aware" of information at all—but that is another question). One can carry this up a few levels and imagine hyperhyperteams carrying and passing information that no hyperteam, not to mention team or solitary ant, ever dreamed of.

I feel it is critical to focus on collective phenomena, particularly on the idea that some information or knowledge or ideas can exist at the level of collective activities, while being totally absent at the lowest level. In fact, one can even go so far as to say that *no* information exists at that lowest level.

It is hardly an amazing revelation, when transported back to the brain: namely, that no ideas are flowing in those neurotransmitters that spark back and forth between neurons. Yet such a simple notion undermines the idea that thought and "symbol manipulation" are the same thing, if by "symbol" one means a formal token such as a bit or a letter or a Lisp Pname.

What is the difference? Why couldn't symbol manipulation—in the sense that I believe Simon and Newell and many writers on AI mean it—accomplish the same thing? The crux of the matter is that these people see symbols as lifeless, dead, *passive* objects—things to be manipulated by some overlying program. I see symbols—representational structures in the brain (or perhaps someday in a computer)—as *active*, like the imaginary hyperhyperteams in the ant colony. *That* is the level at which denotation takes place, not at the level of the single ant. The single ant has no right to be called "symbolic", because its actions stand for nothing. (Of course, in a real ant colony, we have no reason to believe that teams at *any* level genuinely stand for objects outside the colony (or inside it, for that matter) —but the ant-colony metaphor is only a thinly disguised way of making discussion of the brain more vivid.)

Who Says Active Symbols Are Computational Entities?

It is the vast collections of ants (read "neural firings", if you prefer) that add up to something genuinely symbolic. And who can say whether there exist rules—formal, computational rules—*at the level of the teams themselves* (read "concepts", "ideas", "thoughts") that are of full predictive power in describing how they will flow? I am speaking of rules that allow you to ignore what is going on "down below", yet that still yield perfect or at least very accurate predictions of the teams' behavior.

To be sure, there are phenomenological observations that can be formalized to sound like rules that will describe, very vaguely, how those highest-level teams act. But what guarantee is there that we can skim off the full fluidity of the top-level activity of a brain and encapsulate it—without any lower substrate—in the form of some computational rules?

To ask an analogous question, what guarantee is there that there are rules at the "cloud level" (more properly speaking, the level of cold fronts, isobars, trade winds, and so on) that will allow you to say accurately how the atmosphere is going to behave on a large scale? Perhaps there are no such rules; perhaps weather prediction is an intrinsically intractable problem. Perhaps the behavior of clouds is not expressible in terms that are computational *at their own level*, even if the behavior of the microscopic substrate—the molecules—*is* computational.

The premise of AI is that thoughts themselves are computational entities at their own level. At least this is the premise of the information-processing school of AI, and I have very serious doubts about it.

The difference between my active symbols ("teams") and the passive symbols (ants, tokens) of the information-processing school of AI is that the active symbols flow and act on their own. In other words, there is no higher-level agent (read "program") that reaches down and shoves them around. Active symbols must incorporate within their own structures the wherewithal to trigger and cause actions. They cannot just be passive storehouses, bins, receptacles of data. Yet to Newell and Simon, it seems, even so tiny a thing as a bit is a symbol. This is brought out repeatedly in their writings on "physical symbol systems".

A good term for the little units that a computer manipulates (as well as for neuron firings) is "tokens". All computers are good at "token manipulation"; however, only some—the appropriately programmed ones—could support active symbols. (I prefer not to say that they would carry out "symbol manipulation", since that gets back to that image of a central program shoving around some passive representational structures.) The point is, in such a hypothetical program (and none exists as of yet) the symbols themselves are acting!

A simple analogy from ordinary programming might help to convey the level distinction that I am trying to make here. When a computer is running a Lisp program, does it do function calling? To say "yes" would be unconventional. The conventional view is that *functions* call other functions, and the computer is simply the hardware that *supports* function-calling activity. In somewhat the same sense, although with much more parallelism, symbols activate, or trigger, or awaken, other symbols in a brain.

The brain itself does not "manipulate symbols"; the brain is the medium in which the symbols are floating and in which they trigger each other. There is no central manipulator, no central program. There is simply a vast collection of "teams"—patterns of neural firings that, like teams of ants, trigger other patterns of neural firings. The symbols are not "down there" at the level of the individual firings; they are "up here" where we do our verbalization. We feel those symbols churning within ourselves in somewhat the same way as we feel our stomach churning; we do not *do* symbol manipulation by some sort of act of will, let alone some set of logical rules of deduction. We cannot decide what we will next think of, nor how our thoughts will progress.

Not only are we not symbol manipulators; in fact, quite to the contrary, we are manipulated by our symbols! As Scott Kim once cleverly remarked, "free will" won't, perhaps it is more appropriate to speak of "free won't". This way of looking at things turns everything on its head, placing cognition—that rational-seeming level of our minds—where it belongs, namely as a consequence of much deeper processes of myriads of interacting subcognitive structures. The rational has had entirely too much made of it in AI research; it is time for some of the irrational and subcognitive to be recognized for its pivotal role.

The Substrate of Active Symbols Does Not Symbolize

"Cognition as computation" sounds right to me only if I interpret it quite liberally, namely, as meaning this: "Cognition is an activity that can be supported by computational hardware." But if I interpret it more strictly as "Cognition is an activity that can be achieved by a program that shunts around meaning-carrying objects called symbols in a complicated way", then I don't buy it. In my view, meaning-carrying objects won't submit to being shunted about (it's demeaning); meaning-carrying objects carry meaning only by virtue of being active, autonomous agents themselves. There can't be an overseer program, a pusher-around.

To paraphrase a question asked by neurophysiologist Roger Sperry, "Who shoves whom around inside the computer?" (He asked it of the cranium.) If some program shoves data structures around, then you can bet it's not carrying out cognition. Or more precisely, if the data structures are supposed to be *meaning-carrying*, representational things, then it's not cognition. Of course, in any computer-based realization of genuine

cognition, there will have to be, at *some* level of description, programs that shove formal tokens around, but it's only agglomerations of such tokens *en masse* that, above some unclear threshold of collectivity and cooperativity, achieve the status of genuine representation. At that stage, the computer is not shoving them around any more than our brain is shoving thoughts around! The thoughts themselves are causing flow. (This is, I believe, in agreement with Sperry's own way of looking at matters—see, for instance, his article "Mind, Brain, and Humanist Values".) Parallelism and collectivity are of the essence, and in that sense, my response to the title of Barr's paper is *no*, cognition is *not* computation.

At this point, some people might think that I myself sound like John Searle, suggesting that there are elusive "causal powers of the brain" that cannot be captured computationally. I hasten to say that this is not my point of view at all! In my opinion, AI—even Searle's "strong AI"—is still possible, but thought will simply not turn out to be the formal dream of people inspired by predicate calculus or other formalisms. Thought is not a formal activity whose rules exist *at that level*.

Many linguists have maintained that language is a human activity whose nature could be entirely explained at the linguistic level—in terms of complex "grammars", without recourse or reference to anything such as thoughts or concepts. Nowadays many AI people are making a similar mistake: They think that rational thought simply is composed of elementary steps, each of which has some interpretation as an "atom of rational thought", so to speak. That's just not what is going on, however, when neurons fire. On its own, a neuron firing has no meaning, no symbolic quality whatsoever. I believe that those elementary events at the bit level—even at the Lisp-function level (if AI is ever achieved in Lisp, something I seriously doubt)—will have the same quality of *having no interpretation*. It is a level shift as drastic as that between molecules and gases that takes place when thought emerges from billions of in-themselves-meaningless neural firings.

A simple metaphor, hardly demonstrating my point but simply giving its flavor, is provided by Winograd's program SHRDLU, which, using the full power of a very large computer (a DEC-10), could deal with whole numbers up to ten in a conversation about the blocks world. It knew nothing—at its "cognitive" level—of larger numbers. Turing invents a similar example, a rather sly one, in his paper "Computing Machinery and Intelligence", where he has a human ask a computer to do a sum, and the computer pauses 30 seconds and then answers incorrectly. Now this need not be a ruse on the computer's part. It might genuinely have tried to add the two numbers at the *symbol level*, and made a mistake, just as you or I might have, despite having neurons that can add fast.

The point is simply that the lower-level arithmetical processes out of which the higher level of any AI program is composed (the adds, the shifts, the multiplies, and so on) are completely shielded from its view. To be sure,

Winograd could have artificially allowed his program to write little pieces of Lisp code that would execute and return answers to questions in English such as "What is 720 factorial?", but that would be similar to your trying to take advantage of the fact that you have billions of small analog adders in your brain, some time when you are trying to check a long grocery bill. You simply don't have access to those adders! You can't reach them.

Symbol Triggering Patterns Are the Roots of Meaning

What's more, you *oughtn't* to be able to reach them. The world is not sufficiently mathematical for that to be useful in survival. What good would it do a spear thrower to be able to calculate parabolic orbits when in reality there is wind and drag, the spear is not a point mass—and so on? It's quite the contrary: A spear thrower does best by being able to imagine a cluster of approximations of what may happen, and anticipating some plausible consequences of them.

As Jacques Monod in *Chance and Necessity* and Richard Dawkins in *The Selfish Gene* both point out, the real power of brains is that they allow their owners to simulate a variety of plausible futures. This is to be distinguished from the *exact* prediction of eclipses by iterating differential equations step by step far into the future, with very high precision. The brain is a device that has evolved in a less exact world than the pristine one of orbiting planets, and there are always far more chances for the best-laid plans to "gang agley". Therefore, mathematical simulation has to be replaced by abstraction, which involves discarding the irrelevant and making shrewd guesses based on analogy with past experience. Thus the symbols in a brain, rather than playing out a scenario precisely isomorphic to what actually will transpire, play out a few scenarios that are probable or plausible, or even some scenarios from the past that may have no obvious relevance other than as metaphors. (This brings us back to the "adages" of the Yale group.)

Once we abandon perfect mathematical isomorphism as our criterion for symbolizing, and suggest that symbol triggering-patterns are just as related to their suggestive value and their metaphorical richness, this severely complicates the question of what it means when we say that a symbol in the brain "symbolizes" anything. This is closely related to perhaps one of the subtlest issues, in my opinion, that AI should be able to shed light on, and that is the question "What is meaning?" This is actually the crucial issue that John Searle is concerned with in his earlier-mentioned attack on AI; although he camouflages it, and sometimes loses track of it by all sorts of evasive maneuvers, it turns out in the end (see his reply to Dennett in the *New York Review of Books*) that what he is truly concerned with is the "fact" that "computers have no semantics"—and he of course means "Computers do not now have, and never will have, semantics." If he were talking only about the present, I would agree. However, he is making a point in principle, and I believe he is wrong there.

Where do the meanings of the so-called "active symbols", those giant "clouds" of neural activity in the brain, come from? To what do they owe their denotational power? Some people have maintained that it is because the brain is physically attached to sensors and effectors that connect it to the outside world, enabling those "clouds" to mirror the actual state of the world (or at least some parts of it) faithfully, and to affect the world outside as well, through the use of the body. I think that those things are *part of* denotational power, but not its crux. When we daydream or imagine situations, when we dream or plan, we are *not* manipulating the concrete physical world, nor are we sensing it. In imagining fictional or hypothetical or even totally impossible situations we are still making use of, and contributing to, the meaningfulness of our symbolic neural machinery. However, the symbols do not symbolize specific, real, physical objects. The fundamental active symbols of the brain represent *semantic categories*—classes, in AI terminology.

Categories do not point to specific physical objects. However, they can be used as "masters" from which copies—instances—can be rubbed, and then those copies are activated in various conjunctions; these activations then automatically trigger other instance-symbols into activations of various sorts (teams of ants triggering the creation of other teams of ants, sometimes themselves fizzling out). The overall activity will be semantic—meaningful—if it is isomorphic, not necessarily to some actual event in the real world, but to some event that is compatible with all the known constraints on the situation.

Those constraints are not at the molecular or any such fine-grained level; they are at the rather coarse-grained level of ordinary perception. They are to some extent verbalizable constraints. If I utter the Schankian cliché "John went to a restaurant and ate a hamburger", there is genuine representational power in the patterns of activated symbols that your brain sets up, not because some guy named John actually went out and ate a hamburger (although, most likely, this is a situation that has at some time occurred in the world), but because the symbols, with their own "lives" (autonomous ways of triggering other symbols) will, if left alone, cause the playing-out of an imaginary yet realistic scenario. [Note added in press: I have it on good authority that one John Findling of Floyds Knobs, Indiana, did enter a Burger Queen restaurant and did eat one (1) hamburger. This fact, though helpful, would not, through its absence, have seriously marred the arguments of the present article.]

Thus, the key thing that establishes meaningfulness is whether or not the semantic categories are "hooked up" in the proper ways so as to allow realistic scenarios to play themselves out on this "inner stage". That is, the triggering patterns of active symbols must mirror the general trends of how the world works as perceived on a macroscopic level, rather than mirroring the actual events that transpire.

Beyond Intuitive Physics: The Centrality of Slippability

Sometimes this capacity is referred to as "intuitive physics". Intuitive physics is certainly an important ingredient of the triggering patterns needed for an organism's comfortable survival. John McCarthy gives the example of someone able to avoid moving a coffee cup in a certain way, because they can anticipate how it might spill and coffee might get all over their clothes. Note that what is "computed" is a set of alternative rough descriptions for what might happen, rather than one exact "trajectory". This is the nature of intuitive physics.

However, as I stated earlier, there is much more required for symbols to have meaning than simply that their triggering patterns yield an intuitive physics. For instance, if you see someone in a big heavy leg cast and they tell you that their kneecap was acting up, you might think to yourself, "That's quite a nuisance, but it's nothing compared to my friend who has cancer." Now this connection is obviously caused by triggering patterns possessed by symbols representing health problems. But what does this have to do with the laws of motion governing objects or fluids? Precious little. Sideways connections like this, having nothing to do with causality, are equally much of the essence in allowing us to *place situations in perspective*— to compare what actually *is* with what, to our way of seeing things, "might have been" or might even come to be. This ability, no less than intuitive physics, is a central aspect of what meaning is.

This way that any perceived situation has of seeming to be surrounded by a cluster, a halo, of alternative versions of itself, of variations suggested by slipping any of a vast number of features that characterize the situation, seems to me to be at the dead center of thinking. Not much AI work seems to be going on at present to mirror this kind of "slippability". (There are some exceptions. Jaime Carbonell's group working on metaphor and analogy at Carnegie-Mellon is an example. Some other former members of Schank's Yale group have turned toward this as well, such as Michael Dyer and Margot Flowers at UCLA, and Jerry DeJong at Illinois. I would also include myself as another maverick investigating these avenues. Cognitive psychologists such as Stanford's Amos Tversky and Daniel Kahneman of the University of British Columbia have done some very interesting studies of certain types of slippability, though they don't use that term.) This is an issue that I covered in some detail in *Gödel, Escher, Bach*, under various headings such as "slippability", "subjunctive instant replays", "almost' situations", "conceptual skeletons and conceptual mapping", "alternity" (a term due to George Steiner), and so on.

If we return to the metaphor of the ant colony, we can envision these "symbols with halos" as hyperhyperteams of ants, many of whose members

are making what appear to be strange forays in random directions, like flickering tongues of flame spreading out in many directions at once. These tentative probes, which allow the possibility of all sorts of strange lateral connections as from "kneecap" to "cancer", have absolutely no detrimental effect on the total activity of the hyperhyperteam. In fact, quite to the contrary: the hyperhyperteam depends on its members to go wherever their noses lead them. The thing that saves the team—what keeps it coherent— is simply the regular patterns that are sure to emerge out of a random substrate when there are enough constituents. Statistics, in short.

Occasionally, some group of wandering scouts will cause a threshold amount of activity to occur in an unexpected place, and then a whole new area of activity springs up—a new high-level team is activated (or, to return to the brain terminology, a new "symbol" is awakened). Thus, in a brain as in an ant colony, high-level activity spontaneously flows around, driven by the myriad lower-level components' autonomous actions.

AI's Goal Should Be to Bridge the Gap between Cognition and Subcognition

Let me, for a final time, make clear how this is completely in contradistinction to standard computer programs. In a normal program, you can account for every single operation at the bit level, by looking "upward" toward the top-level program. You can trace a high-level function call downward: It calls subroutines that call other subroutines that call this particular machine-language routine that uses these words and in which this particular bit lies. So there is a high-level, global *reason* why this particular bit is being manipulated.

By contrast, in an ant colony, a particular ant's foray is not the carrying-out of some global purpose. It has no interpretation in terms of the overall colony's goals; only when many such actions are considered at once does their statistical quality then emerge as purposeful, or interpretable. Ant actions are not the "translation into machine language" of some "colony-level program". No one ant is essential; even large numbers of ants are dispensable. All that matters is the statistics: thanks to it, the information moves around at a level far above that of the ants. Ditto for neural firings in brains. Not ditto for most current AI programs' architecture.

AI researchers started out thinking that they could reproduce all of cognition through a 100 percent top-down approach: functions calling subfunctions calling subsubfunctions and so on, until it all bottomed out in some primitives. Thus intelligence was thought to be hierarchically decomposable, with high-level cognition at the top driving low-level cognition at the bottom. There were some successes and some difficulties —difficulties particularly in the realm of perception. Then along came such things as production systems and pattern-directed inference. Here, some bottom-up processing was allowed to occur within essentially top-d

context. Gradually, the trend has been shifting. But there still is a large element of top-down quality in AI.

It is my belief that until AI has been stood on its head and is 100 percent bottom-up, it won't achieve the same level or type of intelligence as humans have. To be sure, when that kind of architecture exists, there will still be high-level, global, cognitive events—but they will be epiphenomenal, like those in a brain. They will not in themselves be computational. Rather, they will be constituted out of, and driven by, many many smaller computational events. In other words, *subcognition at the bottom will drive cognition at the top*. And, perhaps most importantly, the activities that take place at that cognitive top level will neither have been written nor anticipated by any programmer. This is the essence of what I call *statistically emergent mentality*.

Statistically Emergent Mentality Supersedes the Boolean Dream

Let me then close with a return to the comment of Simon's: "Nothing below 100 milliseconds is of interest in the study of cognition." I cannot imagine a remark about AI with which I could more vehemently disagree. Simon seems to be most concerned with having programs that can imitate chains of serial actions that come from verbal protocols of various experimental subjects. Perhaps, in some domains, even in some relatively complex and technical ones, people have come up with programs that can do this. But what about the simpler, noncognitive acts that in reality are the substrate for those cognitive acts? Whose program carries those out? At present, no one's. Why is this?

It is because AI people have in general tended to cling to a notion that, just as formal rules at the thought level, just as in some sense, thoughts obey formal rules at the thought level, George Boole believed that "the laws of thought" amounted to formal rules for manipulating propositions. I believe that this Boolean dream is at the root of the slogan "Cognition as computation"—and I believe it will turn out to be revealed for what it is: an elegant chimera.

Post Scriptum.

Since writing this diatribe, I have found, to my delight, that there are quite a few fledgling efforts underway in AI that fall squarely under the "statistical emergence" banner. I mentioned the work by Norman and Rumelhart at the Institute for Cognitive Science at the University of California at San Diego. That institute is in fact a hotbed of subversive "PDP" (parallel distributed processing) activity. Paul Smolensky, a PDP researcher there, has developed a theory of perceptual activity directly based on an analogy to the branch

of physics known as statistical mechanics, and it includes a mental counterpart to the physical concept of *temperature*. In physics, temperature is a number that measures the degree of random thermal jumbling going on in a system composed of many similar parts. In Smolensky's work, a "computational temperature" controls how much randomness is injected into the system.

Imagine a system that is "looking" at a simple scene. (I mean it has a television camera providing input to a computer.) This system's job is to figure out the most plausible interpretation of what is "out there". Is it the word "READ"? Is it the system's grandmother? Is it Smolensky's dog Mandy? When the system is first faced with a fresh situation, the temperature is high, indicating that the system is in a completely open-minded state, ready to have any ideas activated. As randomly chosen concept fragments (not full concepts) are tried on for size, the system gradually starts developing a sense for what sorts of things "fit". Thus the temperature is lowered a bit, lessening the chances of stray concept fragments floating in and destroying the fragile order that is just beginning to coalesce. As fragments start to fit together coherently, the system continues to turn down its randomness knob.

Gradually, larger conceptual structures begin to form and to confirm each other in a benign, self-reinforcing loop. Furthermore, these high-level structures now bias the probabilities of random activation of lower-level fragments, so that the thermal activity, though still random, is more directed. The system is settling into a stable state that captures, in some internal code, the salient external realities. When it is completely "happy" (or "harmonious", in Smolensky's terminology), then the system's temperature reaches zero: it is "freezing". It is no coincidence that the moment of freezing coincides with the attainment of maximal computational bliss, for the temperature gets lowered only when the system is seen to have made some upward jump in its happiness level.

This idea of stochastically guided convergence to what is called a *globally optimum state* seems to have arisen (as do so many good ideas) in the minds of several people at once, spread around the globe. For all I know, it is an ancient idea (though I will not go so far as to credit the ancient Buddhists with it), but it seems that the atmosphere has to be just right for this kind of spark to "catch". People not involved in AI sometimes have expressed the spirit of this sort of thing very poetically. Here is Henri Poincaré writing in the early part of this century about the genesis of mathematical inspirations:

Permit me a rough comparison. Figure the future elements of our combinations [full-fledged ideas] as something like the hooked atoms of Epicurus. During the complete repose of the mind, these atoms are motionless, they are, so to speak, hooked to the wall; so this complete rest may be indefinitely prolonged without the atoms meeting, and consequently without any combination between them.

a bit, occasionally "mate", then couples continue the search for other compatible couples as well as for more letters they could gobble up to make triples or quaduples. (See Figure 27-3.) Syllables build, sniff at each other's ends, occasionally unite, making word candidates. Then those large "gloms" can undergo internal transformations, break down into their natural subunits or even into elemental smithereens. For instance, "pan-gloss" could become "pang-loss" by *regrouping*, which could then by *spoonerism* become "lang-poss", and so on. Forkerism and kniferism (like spoonerism, only different) are other types of recombination mechanisms, as are sporkerism and foonerism. A typical low-temperature route, meandering through a portion of logological space using these mechanisms, might visit, in sequence, "lang-poss", "lass-pong", "lasp-song", "song-lasp", "son-glasp", and so on. And if, as a consequence of global tension, the temperature rises, the entire bubble may burst apart and we will be left with isolated letters scattered all over the place, with occasional surviving duplets ("ng", maybe) here and there, souvenirs of what it was like before the blast. Sigh . . . Oh, but why suffer pangs of loss? After all, isn't this world, of all possible worlds, the very best?

* * *

Given the passages from Poincaré and Thomas, I will not claim that these ideas are totally new—but then, why would I want to? Part of my thesis on creativity is that even the best ideas are simply variations on themes already enunciated, discovered by unconscious and random processes of recombination, filtering, and association. In fact, the "fit" between statistical mechanics and "statistical mentalics" is not yet exact, and it is to be hoped that the collective mental temperature of cognitive scientists is high enough that the jiggling-about of ideas in our brains will finally bring the right ones into contact with each other, thus bringing us closer to an accurate view of the physics-cognition connection, allowing the temperature to go down, bringing us even closer to truth, which will lower the temperature still further—and on and on.

Besides Paul Smolensky, there are many other people sniffing about in roughly the same territory. David Rumelhart (mentioned above), James McClelland, and co-workers in the "PDP" group at San Diego have modeled several types of perceptual and cognitive behavior using a system of this sort. Geoffrey Hinton and Scott Fahlman (like Simon and Newell, at Carnegie-Mellon University) and Terrence Sejnowski (of Johns Hopkins) are exploring, via what they call the "Boltzmann machine", "pseudo-neural" models of learning, based on ideas closely resembling those of Smolensky. (The prognosis is good, for "neural" rearranges into "u learn".) J.J. Hopfield of Caltech has studied the statistical properties of neural nets, to see what one can say about the substrate of associative memory. Pentti Kanerva, a highly original and autonomous philosopher-programmer at

---

On the other hand, during a period of apparent rest and unconscious work, certain of them are detached from the wall and put in motion. They flash in every direction through the space (I was about to say the room) where they are enclosed, as would, for example, a swarm of gnats or, if you prefer a more learned comparison, like the molecules of gas in the kinematic theory of gases. Then their mutual impacts may produce new combinations . . . .

Now our will did not choose them at random; it pursued a perfectly determined aim. The mobilized atoms are therefore not any atoms whatsoever; they are those from which we might reasonably expect the desired solution. Then the mobilized atoms undergo impacts which make them enter into combinations among themselves or with other atoms at rest which they struck against in their course. Again I beg pardon, my comparison is very rough, but I scarcely know how otherwise to make my thought understood.

And more recently the biologist Lewis Thomas, in his book *The Medusa and the Snail*, wrote this:

At any waking moment the human head is filled alive with molecules of thought called notions. The mind is made up of dense clouds of these structures, flowing at random from place to place, bumping against each other and caroming away to bump again, leaving random, two-step tracks like the paths of Brownian movement. They are small round structures, featureless except for tiny projections that are made to fit and then lock onto certain other particles of thought possessing similar receptors. Much of the time nothing comes of the activity. The probability that one notion will encounter a matched one, fitting closely enough for docking, is at the outset vanishingly small.

But when the mind is heated a little, the movement speeds up and there are more encounters. The probability is raised.

The receptors are branched and complex, with configurations that are wildly variable. For one notion to fit with another it is not required that the inner structure of either member be the same; it is only the outside signal that counts for docking. But when any two are locked together they become a very small memory. Their motion changes. Now, instead of drifting at random through the corridors of the mind, they move in straight lines, turning over and over, searching for other pairs. Docking and locking continue, pairs are coupled to pairs, and aggregates are formed. These have the look of live, purposeful organisms, hunting for new things to fit with, sniffing for matched receptors, turning things over, catching at everything. As they grow in size, anything that seems to fit, even loosely, is tried on, stuck on, hung from the surface wherever there is room. They become like sea creatures, decorated all over with other creatures as living symbionts.

At this stage of its development, each mass of conjoined, separate notions, remembering and searching at the same time, shifts into its own fixed orbit, swinging in long elliptical loops around the center of the mind, rotating slowly as it goes. Now it is an idea.

This poetic passage reminds me of nothing more than my Jumbo system for doing anagrams, which I developed in 1982. There, in what I call the "cytoplasm", letters bash at random into other letters, check each other out

Stanford, has done related theoretical work aimed at suggesting plausible substrates underlying the fluidity of memory, and his findings dovetail beautifully with recent observations about the anatomical structure of various areas of the brain. This may be a coincidence and it may not, but there is certainly plenty there to speculate about. Related work has been done by T. Kohonen in Finland, and O. P. Buneman and D. Willshaw in England. James Anderson and Stuart Geman at Brown University have developed theories and models of how collective activity of many individual processing units can have emergent character. Jerome Feldman and colleagues at the University of Rochester have developed what they call a "connectionist" theory of perception and cognition, in which neurons can assemble into stable and not-so-stable aggregates called "coalitions". These shifting alliances are presumed to form the subcognitive basis of fluid cognition. And finally, my group's active projects—Jumbo, Seek-Whence, and Copycat—are all thoroughly permeated with an independently conceived vision of a temperature-controlled randomness at the subcognitive level, out of which emerges, at the cognitive level, a fluid but hopefully not wildly meandering train of thought.

Marsha Meredith, who has been working on implementing a Seek-Whence program, seems to really have taken the idea of "fluid" cognition to heart. In writing up what she has implemented so far, she spoke of the cytoplasm of her system:

The cytoplasm might be viewed as a soup bubbling with gloms, the bubbles which rise to the top being the system's current view of the sequence. If neighboring bubbles have enough mutual attraction (strong enough bonds) they will combine; otherwise they will either exist independently or burst to permit new bubbles to take their place.

In addition to her cytoplasm, Marsha has created a "Platoplasm" (where Platonic concepts are stored) and a "Socratoplasm" (to mediate between the down-to-earth cytoplasm and the ethereal Platoplasm). Marsha's bubbling, boiling, churning, roiling "Seek-Whence soup" is thus very much like alphabet soup, the only difference being that the good old ABC's have been replaced by 123's.

*   *   *

I think it would be silly to try to attach credit to any one person for these "soup-cognitive" ideas, for they are in the air, as it were, and the time is simply ripe. This is not to say that they are being roundly welcomed by the whole AI and cognitive science community. There are definite "pro" and "con" camps, and some more neutral observers. There are people who cling to the Boolean dream like it was going out of style! Daniel Dennett has recently coined another term for the same concept: "High Church Computationalism", to which he contrasts what he calls "The New

Connectionism". I like the vision of orthodoxy implied by the former term, but I think the latter term overstresses the role of neural modeling in the new approaches. A model of thought in the new style need not be based so literally on brain hardware that there are neuron-like units and axon-like connections between them. The essence of the dissenting movement lies, it seems to me, in three notions:

(1) asynchronous parallelism;
(2) temperature-controlled randomness;
(3) statistically emergent active symbols.

Actually, for those who understand this intuition well, line 3 alone says it all. How? Well, the phrase "statistically emergent" clearly implies that collective phenomena are involved, in which many independent uncorrelated micro-events, chaotically spread all about in some physical medium, are happening all the time, forming and breaking patterns. This is the imagery attached to lines 1 and 2.

I am reminded, whenever I visualize this kind of thing really clearly, of one fairly old but still influential theory about how water's fluidity emerges out of all the frenetic molecular bumping and banging "down there". This is the theory that goes by the poetic name of *flickering clusters* (referred to also in Chapter 10). The idea is that water molecules can form small and highly ephemeral hydrogen-bonded clusters (with a lifespan even shorter than a mayfly's!). Within microseconds, a group will form and break down again, and its constituent molecules will regroup with other free ones. This is going on, over and over, day and night, second by second, in every tiny drop of water, gadzillions of times. The statistically emergent phenomenon, in that case, is the macroscopic nature of water. In particular, such familiar physical properties of water as its boiling point, density, viscosity, compressibility, and so on are deducible—at least in theory—from such a model.

If one is concerned with minds, however, the phenomena to be explained are less tangible and far more elusive. What seems to most people a primary goal to aim for—and here John Searle and I agree, for once—is that of explaining where *meaning* really comes from, or in other words, a theory of the basis of semantics, or reference. Put in a nutshell, the question is, "What makes mental activity *symbolic*?"

*   *   *

There seems to be a genuine conundrum about how mere matter could possess *reference*. How could a lump of stuff be *about* anything else (let alone about *itself*)? Searle conveniently exempts bio-stuff (or at least neuro-stuff) from this query, assigning to it special "causal powers" that he mysteriously declines to identify but that magically (it would seem) allow brains, or

seem to experience pain and fear and curiosity." Men? "Well, maybe . . . despite the fact that they don't know what it's like to menstruate."

Such reactions are somewhat natural, but it is curious to me that what seems to be the most convincing is the moving-about in the world, and the perceptual and motor interface. Systems that are not interfaced with our tangible, three-dimensional world via perceptors and motor capacities, no matter how sophisticated their innards, seem to be un-identifiable-with, by most people. I have in mind a certain kind of program that most people would probably find it ludicrous to ever consider conscious: a program that does symbolic mathematical manipulations. Take the famous system called Macsyma, for instance, which can do calculus and algebra problems of a very high order of difficulty. Its performance would have been so unimaginable in the days of Gauss or Euler that many smart people would have gasped and many brilliant people might have worshiped it. No one could pooh-pooh it—but today we do. Today we are "sophisticated". In a way, this is good, but in a way it is bad.

What bothers me is a kind of "hardware chauvinism" that we humans evince. This chauvinism says, "Real Things live in three dimensions; they are made of atoms. Photons bounce off Real Things. Real Things make noises when you drop them. Real Things are material, not insubstantial mental ghosts." The idea that numbers or functions or sets or any other kind of mathematical construct might be Real would provoke guffaws in many if not most intellectual quarters today. The idea that being able to maneuver about in a "space" or "universe" of pure abstractions might entitle a robot to be called "sentient" would be ridiculed to the skies, no matter if the maneuvering in that abstruse high-dimensional space were as supple and graceful as that of the most skilled Olympic ice-skating champion or the greatest jazz pianist.

Speaking of which, the musical universe provides another wonderful testbed. Would a robot able to devise incredibly beautiful, lyrical, flowing passages that brought tears to your eyes be entitled to a bit of empathy? Suppose it were otherwise immobile, its only conception of "reality" being inward-directed rather than something accessible through hands or eyes or ears. How would you feel then?

I personally don't think that such a program could come to exist in actuality, but as a thought experiment it asks something interesting about our conception of sentience. Does access to the "real world" count for a lot? Why should the intangible world of the intellect be any less real than the tangible world of the body? Does it have less structure? No, not if you get to know it. Every type of complexity in the physical world has its mirror image in the world of mathematical constructs, including time. What kind of prejudice is it, then, that biases us in favor of our kind so strongly? As questions of mind and matter grow ever more subtle, we must watch out for tacit assumptions of this sort ever more vigilantly, for they affect us at the deepest level and provide pat answers to exceedingly non-pat questions.

---

something in them, to refer. This is as thoroughly *ad hoc* as the Boolean dreamers' chutzpah in simply proclaiming that there is no problem at all there, for Lisp atoms *do* refer. The fact of the matter is that an analysis of what reference is has proved a little too tough for both sides so far, and so it degenerates into polemics. Each side already *knows* what "aboutness" is all about, and is most impatient with the other side for its obtuseness. I certainly am just as guilty of this syndrome as any other party, for I too feel I *know* (intuitively and nonverbalizably) just what reference really is, and how it *can* come out of "mere matter" and its patterns. I devoted a very large portion of *Gödel, Escher, Bach* to trying to get across some of those intuitions, and since then I have continued to try to spell them out better (most notably in a paper called "Shakespeare's Plays Weren't Written by Him, But by Someone Else of the Same Name", not co-authored by Gray Clossman and Marsha Meredith but by people of the same names, and in the developing work on roles and analogies, described in Chapter 24 of this book). The questions still seem to stymie the best minds, however.

Does the expression "—p—q—-" intrinsically mean anything? Does the expression "(SS0+SS0)=SSSS0" intrinsically mean anything? How about "(equals 4 (plus 2 2))" or "2+2=4" or "bpbqd"? What would imbue *one* of them with meaning, if not *all*? If none of these has meaning, then do printed symbols *ever* have meaning? Does an entire set of the *Encyclopaedia Britannica* tumbling out of control in interstellar space have any intrinsic meaning, or is it just an empty lump of nonsymbolic matter? Would it help if we lifted the entire Library of Congress into that selfsame interstellar orbit? If not, why not?

What about a cute little robot that scampers about in your living room, seeking to plug itself into any locatable electric outlet and avoiding banging into furniture? Has it got anything inside it that truly *represents* anything else? If so, why? If not, why not? What about a human-sized robot that roams the world in search of beauty and truth and along the way "emits" strange pieces of weird and garbled "syntactic behavior" such as "This sentence no verb"—might that type of robot possess any shreds of *aboutness?* Or would you have to know precisely what it was made of, down to the the most microscopic fibers of its circuitry? What if it objected to such examination? Would your prior knowledge that it was a robot tell you that it was merely "artificially signaling" such objections, and entitle you (as a *bona fide* sentient being) to override its *ersatz* objections without compunction, and to open it up and dissect it?

*

*

*

In a way it is natural but in another way it is curious that most people's threshold for changing their tune on whether or not an organism has a mind and feelings (and "aboutness") seems to lie at just about the point where they can easily identify with the organism. Microbes? "Naah, they're too small." Mosquitos? "Maybe, but they're just mechanical." Mice? "They sure

In a way, the AI hope up till recently has been to get away with just one level. This is not dissimilar to the hopes of the brain-research people, who in their own way have wanted to locate everything in just one level: that of neurons. Well, AI people are loosening up and so are brain people, and some meaningful dialogue is beginning. This is a hopeful sign, but some people resent the implications that their long-held views are being challenged. They particularly resent anyone's writing about such matters in a general and philosophical way, full of imagery, meant to stir up the intuitions rather than to present well-known facts dryly and impartially.

My aim in the preceding article, which was solicited expressly for the purpose of interdisciplinary communication (it was published in *The Study of Information: Interdisciplinary Messages*, edited by Fritz Machlup and Una Mansfield), was to spark new intuitions about places where progress is needed—not so much specific new experiments, but new areas for musing and theorizing. I was hoping to stimulate not only AI people but also cognitive psychologists, philosophers of mind, and brain researchers. That is why I used so much imagery and appealed to the intuition.

Allen Newell, whose ideas were criticized in the article, did not take too kindly to it. In his reply (solicited by the book's editors), he dismissed my ideas as nonscientific, despite the fact that all the articles solicited were expressly requested to be personal viewpoints rather than scientific papers. In fact, he treated my article with as much disdain as one would treat a pesky fly that one wanted to swat. I had expected, and would of course have warmly welcomed, a reply discussing the issues in a substantive way. He

---

The question that launched this digression was what kinds of entities deserve attribution of genuine meaning, genuine symbolicness. Some people, Searle for one, seem to feel that nothing any computer system might do could ever be genuinely symbolic. It might well capture the "shadows" of symbolic activity, but it would never have the "right stuff", that is, the "causal powers of the brain", whether or not it passed the Turing Test. Now, I don't agree at all with Searle about there being an unbreachable machine-mind gap, but I do agree with his skepticism toward orthodox AI's view that we have just about got to the point where computers are using words and symbols with genuine meanings, in the full sense of the term.

The problem is, as I emphasized in the article, that computers' concepts thus far lack slippability (and therefore, their "aboutness" is very weak). The blurry boundaries between human concepts are not well captured by models that try to do blurring explicitly. Such models range from so-called "fuzzy set theory", in which an unblurry amount of blurriness is inserted into the most precise of logical calculi (actually a rather comical idea), to memory models with concepts strung together in complex kinds of webs, with hierarchical and lateral connections galore, even including explicit "hierarchies of variability". Somehow human fluidity is not even approached, though.

The alternate school's recipe is to build symbolic activity up from nonsymbolic activity, rather than presuming that the objects one begins with (Lisp atoms, for instance) can be imbued with all the fluidity one wants by making ever-larger piles of complex rules to push them around in the right ways. I am a strong believer in the idea that symbolicness, like greenness, disintegrates. E. O. Wilson's idea of "mass communication" being "the transfer, among groups, of information that a single individual could not pass to another" seems to me to be at the heart of the idea of statistically emergent active symbols. Somehow, in any genuinely cognitive system, there must be layers upon layers of organization, allowing fluid semantics to emerge at the top level out of rigid syntax at the bottom level. Symbolic events will be broken down into nonsymbolic ones. In the ant-colony metaphor, the top-level hyperhyperteams will be symbolic, hyperteams will be subsymbolic, mere teams will be subsubsymbolic (whatever that means!), and the lowly ants will be totally devoid of symbolicness. Obviously, the number of levels need not be four, but this is enough to make a point: Symbolic events are *not* the primitives of thought.

If you believe in this notion of different layers of collectivity having different degrees of symbolicness and fluidity, then you might ask, "What can we learn from trying to make a system with a small number of such layers?" This is an excellent scientific question. In fact, simply to make a two-layer system in which the upper layer is simultaneously more collective, more symbolic, and more fluid than the lower layer would be the key step —and that is precisely what the statistical-emergence camp is trying to do.

Fortunately, Newell did spend a page or so doing that kind of thing. He pointed out that in his and Simon's writings, the word "symbol" has always had the meaning of "something that denotes", as distinguished from mere tokens, such as the bits at the bottom level of a computer. He gave several excerpts from articles by Simon and himself, including the following one, referring to the 0's and 1's in a typical computer:

These entities are not symbols in the sense of our symbol system. They satisfy only part of the requirements for a symbol, namely being the tokens in expressions. It is of course possible to give them full symbolic character by programming an accessing mechanism that gets from them to some data structure.

Newell claims that in my article I have seriously misrepresented his and Simon's well-known views on physical symbol systems. A typical passage where he feels I do so is this one:

To me . . . 'symbol' connotes something with representational power. To them (if I am not mistaken), it would be fine to call a bit (inside a computer) or a neuron-firing a 'symbol'.

than bits be somewhat consistent with my view? My answer is *no*, and here's why. The rules governing Lisp structures are strictly computational in and of themselves, and implementing a Lisp system in 0's-and-1's hardware adds nothing enriching to the Lisp atoms whatsoever. The logic of a Lisp system does not emerge from the details of levels below it; it is present in full in the written program even without any computer that can run Lisp. In that sense, Lisp programs are Platonic, which is so well demonstrated by Gödel's original "Lisp program", written way back in 1931, before computers existed. In fact, the only distinction between bits and atoms is in number: There are only two types of bit, whereas there can be an arbitrarily large variety of atoms. But as for *fluidity*, nothing is gained by moving from the bit level to the atom level. Either level is 100 percent formal in operation.

What we are looking for, however, in explaining cognition, is *a bridge between the formal and the informal*. Now it may be that Newell does not believe in cognition's informality, and I probably would not be able to convince him of it. Indeed, it would be hard to convince anyone who doesn't see it already that it is reasonable to think of human cognition in those terms, but that is how I see it. And statistical emergence seems to me to be not merely a shot in the dark, but the obvious route to explore. The brain certainly does an immense amount in parallel, with different parts operating completely independently from others. There is known to be a lot of "noise", or randomness, in the brain, and moreover, the world itself is acting on the brain in so many different ways at once that it is like being bombarded simultaneously with the output of a thousand different random number generators. So temperature there's plenty of. All we need to figure out is what kinds of collective entities could evolve in such a rich medium, how they would interact, and how they could be symbolic.

This is the challenge I was posing to Newell and other staunch believers in the Boolean dream. The debate will continue, but meantime research must be done. And there, everyone must be guided by personal intuitions about what the right path is. Newell and Simon have theirs, and I have mine. We both think we're right. As Wanda Landowska, the famous harpsichordist, once remarked, "You play Bach *your* way, and I'll play him *his* way." How can one reply to that? No way. So let the game go on!

Newell comments bluntly: "Hofstadter is indeed mistaken, absolutely and unequivocally." Now here is an opportunity for substantive discussion! I am glad to reply at that level.

*   *   *

Firstly, I plead guilty to one count of misrepresentation of the Newell-Simon view of symbols. I now realize that they place the symbolic level above the bit level; effectively, they place it at the level of Lisp structures. However, I wish to point out that there is a curious vacillation on Newell's part in the paper from which he draws the quote given above. In the first part of the paper, he repeatedly uses the word "symbol" to refer to the 0's and 1's in a Turing machine. In fact, he does it so often that a naive reader *might* conclude that Newell considers them to *be* symbols. But no! It turns out that after more than a dozen such usages, he turns right around and repudiates any such usage, in the passage quoted above. That, I submit, is hardly clarity in writing, and I would request that it be considered by the jury as constituting mitigating circumstances, possibly providing grounds for a reduced sentence for my client.

But there is a more substantive area of disagreement. Newell repeatedly makes the point that for him, a physical symbol is virtually identical to a Lisp atom with an attached list (usually called its "property list"). He says as much: "That Lisp is a close approximation to a pure symbol system is often not accorded the weight it deserves." And later on, he refers to his paradigmatic physical symbol system as "a garden variety, Lisp-ish sort of beast". (It is no coincidence that the name of one company making Lisp machines is "Symbolics".) Throughout his article, Newell refers to the *manipulation of symbols* by programs (although strangely, he avoids the word "program"). I may have been "mistaken, absolutely and unequivocally" in attributing to Newell and Simon the view that bits are symbols, but I am certainly not mistaken in attributing to them the view that a Lisp atom with attached property list has all the prerequisites of being a genuine symbol, as long as the right program is manipulating it. That much is crystal-clear. And that is the view I was opposing, no less than the view of bits as symbols.

As a sidelight, it is an amusing coincidence that John Searle was quite upset when, in *The Mind's I*, I misquoted him, saying he had said "a few bits of paper" when he had actually said "slips of paper". Now I find myself in a similar situation: I accused someone of having said "bits" when they meant something else. Searle meant "slips"; Newell meant "lisps" (Lisp atoms or lists). And in both cases, although I admit I was wrong in detail, I feel I was entirely right in principle. My arguments remain unchanged even after the misquotation is corrected.

To some, the build-up of atoms from bits might seem to resemble the first layer of emergence of fluid semantics from rigid syntax that I was speaking of earlier. So couldn't a view that sees Lisp structures as slightly more fluid