Cognitive Science And The Search For Intelligence

Jason Eisner
Invited talk at the Socratic Society
University of Cape Town, South Africa
20 May 1991

The human mind is a sort of mysterious, amorphous substance, like a handful of clay from a fossil-rich gorge. We are told that it is a mixture of thoughts, emotions, memories, perspectives, and habits, half-blended and bound together stickily by something called consciousness. It contains fossils from millions of years ago; yet if we so much as press a thumb into it today, it retains the imprint. Though it is not quite physical, under extreme physical conditions it will vanish – as clay under pressure ceases to be clay, and becomes stone and water.

From the time of the ancient Greeks, through the Enlightenment and Kant and up to the present day, we have been asking questions about our mental existence. Some of these questions now seem naïve: How large is a thought? Where does it go when we're not thinking it? Can new ideas be created in the mind, or, as Socrates argued, must they all be present at birth? Other questions are still with us. How do we interpret the sensory world? What is the nature of knowledge, and where does it come from? In what sense, if any, are we rational? How does our intelligence differ from that of animals, and are the differences merely ones of degree?

Only in the nineteenth century did anyone began to study the mind scientifically – a task that many had thought impossible. The philosopher Johann Herbart pointed out that while ideas might not have measurable spatial dimension, they did have duration, quality, and intensity, which could be measured. This suggestion triggered a spate of research. Soon after, Hermann von Helmholtz successfully determined the speed of nerve impulses in animals and humans, and F. C. Donders found ways to time low-level mental operations themselves, such as the classification of a sensory stimulus. Gustav Fechner showed that across all the human senses, the perceived intensity of a stimulus was logarithmically related to its physical intensity.[1] In the early twentieth century, psychologists like Jean Piaget even began studying the content of ideas; they especially wanted to know whether people made mistakes in a systematic way.

Such measurements have become the tools of cognitive psychology in this century. The standard approach is to study people's performance on an artificial task, under varying conditions. This allows us to theorize about *how the task is being accomplished.* A

[1]This account is summarized from Gardner (see Further Reading), pp. 99-101.

classic example is Saul Sternberg's paradigm for studying memory. The experimenter reads a list of numbers – 5, 7, 2, 9, 12 – then asks you whether some "probe," such as 9, was in the list. You answer yes or no as quickly as you can. Now, there are many interesting questions to ask, and many of the answers are surprising. What happens to your speed and accuracy as the list of numbers gets longer? When the probe is in the list, does it matter whether it appears early or late? When it's not in the list, does the particular choice of probe make any difference? What if the list is a mixture of one-digit and three-digit numbers? What if it is organized in some obvious way (2, 4, 6, 8, 10, 12, 14)? What if some numbers are repeated within the list? What if the experimenter doesn't use numbers at all, but common nouns, or sentences, or the names of your friends, or pieces of advice, even all of these jumbled together? What happens if the list is read very quickly? If you hear it twice? If you're quizzed about it the next day? Suppose you happen to be brain-damaged in one of a hundred ways – what kinds of difference might that make?

The Big Questions

> If the work of God could be comprehended by reason, it would be no longer wonderful.
>
> –Pope Gregory I, 6th century

> A year spent working on artificial intelligence is enough to make one believe in God.
>
> –Anon.

Such investigations shed light on the organization of memory and the recall process. Not everyone finds those subjects deeply interesting in themselves. After all, one could perform similar experiments on the "search" facility of a word processor. But these studies of human memory – or, to put it more intriguingly, the representation and retrieval of knowledge – fit into a much broader study of what the brain does and how it does it. Those are big questions. The tasks that people perform on a daily basis are really astounding.

Take visual perception. If I hold up an object, you can tell me what it is: "A plastic comb." This ability seems perfectly ordinary – until you try to program it into a computer. Then the difficulty of the task becomes apparent. Depending on how I hold the comb and you hold your head, the light-sensitive cells at the back of your eye will be stimulated in one of infinitely many different ways. When I move the comb slightly, every photocell is affected. Yet somehow you recognize "plastic combness" in all these configurations of light. I can show you two objects that are superficially unlike, and you will recognize them both as combs. The problem is even more perplexing when I show you a glass jar and you can identify it. After all, you have never really *seen* glass at all – glass merely distorts the scene behind it in characteristic ways!

If I now take you into a roomful of people – at a cocktail party, say – your abilities are so phenomenal as to almost defy mortal explanation. There are a thousand identifiable objects in the room: people, people's limbs, articles of clothing, glasses, alcoholic beverages inside the glasses, and so on. Few of these objects are wholly visible. Yet you can identify all of them, and tell me the physical relations they probably have to each other: "That hand holding the whiskey glass? It's attached to the arm inside the red sweater, which must be Phyllis's arm, although it's a little hard to tell with that fat guy standing in front of her." Even more remarkable, if all the professors are on one side of the room and all the students on the other, you are quite sure to notice that fact. Note that such an observation requires you to correlate the age of dozens of individuals with their spatial location, for no apparent reason. (Guessing people's ages is itself so difficult that a foreigner often cannot do it – let alone a computer!)

For a different example, consider the phenomenon of language understanding. As I speak to you, all I am doing is making vibrations in the air. Your ears are equipped to pick these vibrations up: at any moment in time, your ears register the amount of energy on each audio frequency from 100 to 20,000 Hz. You analyze this sound spectrograph on several levels:

1. *Phonetics.* At the lowest level, you classify bits of sound as various vowels and consonants. (Even if someone synthesizes a continuum of þ sounds ranging between "b" and "p," you will always perceive these sounds to be one consonant or the other, never something in between.)

2. *Lexicalization.* At the level above phonetics, you must segment the sound sequence into meaningful words. Most pauses in speech fall within words, not between words, so this is no trivial task. Yet you do it without even realizing it. The difficulty is only apparent for languages one speaks badly. For example, when I listen to a French conversation, I am often unable to pick any French words out of the rushing stream of sound.

3. *Syntax.* Once you have all the words of a sentence, you can impose a syntactic structure on it, relating the words to one another. The set of possible structures is constrained by complex linguistic principles. If I tell you,

> Koos is scared that the judge will convict himself,

the word "himself" necessarily refers to the judge, not to Koos, no matter how implausible this makes my sentence. You might question whether I really have my facts straight, or whether Koos does; but the meaning of the sentence stands.

4. *Semantics.* The syntactic structure of the sentence gives you a way to interpret its overt meaning. Once you have identified the relationships of the words, once you have distinguished the subject from the predicate, you can tell who is scared and why he is scared. Once you know that the auxiliary word "will" modifies the tense of "convict," you can conclude that the feared conviction is yet to come.

5. *Pragmatics.* The overt meaning of a sentence is not always its complete or even its true meaning. Language is used to communicate; its meaning is dependent on the context of the situation. The following examples should make this clear:

> Tourist: Stellenbosch train, please?

Spoornet Worker: Track 15. And you'd better hurry.

Speaker 1: I hear that Phyllis is coming to this cocktail party.
Speaker 2: Phyllis is an ugly, spiteful bag of bones who would eat her own grandmother without salt.
Speaker 1: Well! Nice weather we're having, isn't it?

Clearly, the last two statements have nothing to do with either Phyllis's grandmother or the weather. Phyllis may not have a grandmother, and it may very well be snowing outside.

Each of these processing levels offers an agenda of challenging theoretical questions. How challenging? Well, thousands of linguists have been trying for twenty-five years to pin down the principles of English syntax, with only moderate success. It is worth noting that if English was your first language, you'd grasped 90% of those principles by the time you were three years old, simply by hearing an arbitrary set of spoken English sentences; and no one knows how you did *that*, either.

What's worse, these processing levels for language are not separate stages. They influence each other intimately. Syntax helps determine meaning; but at the same time, considerations of meaning may "reach down" a level or two and sway the interpretation of syntax. Who are "they" in the following sentences?

- The city council refused to grant the women a permit because they feared violence.
- The city council refused to grant the women a permit because they advocated violence.

The higher levels may even "reach down" to influence the lexical and phonetic analyses. For example, we can usually understand distorted tape recordings, or people with unusual accents. In conversation, no one has any trouble understanding the following mumbled sentences, where þ represents a sound that could be b or p:

John dumped some trash in the þin.
John mounted a butterfly on the þin.
John cast a þall over the fence.
John cast a þall over the party.

We do so well at integrating these multiple influences that we are unaware, on a conscious level, that language is riddled with ambiguities. A famous example is the apparently innocuous proverb, "Time flies like an arrow." Who but a syntactician would suspect that this sentence is five ways ambiguous? But it is. As some linguist once quipped: "Time flies like an arrow, but fruit flies like a banana." And then there was the grad student whose advisor admonished, "Time flies like an arrow, if you must; but time experiments like a scientist!"

Perhaps these two areas of research, vision and language understanding, give you a sense of how complex and remarkable mental processes really are, and why one might try to study them scientifically.

Tackling the Big Questions: The Cognitive Science Enterprise

> If physiology were simpler and more obvious than it is, no one would have felt the need for psychology. –Richard Rorty

> Great fleas have little fleas upon their backs to bite 'em,
> And little fleas have lesser fleas, and so *ad infinitum*. –Augustus Morgan

The study of such mental processes is known, these days, as cognitive science. (The term actually dates back to 1960 or so.) If cognitive scientists have one grand question to answer, it goes like this:

> The Grand Question: What exactly is the mind doing, and how does it manage to do it with a one-kilo hunk of neurons?

Most researchers would agree that this grand question is the right one to ask. In practice, however, it falls apart into two questions.

> The Top-Down Question: What exactly is the mind doing, and how could *anything* do it?

> The Bottom-Up Question: How is the brain organized?

As the names imply, these questions are pursued in different ways. One starts with the high-level phenomena of intelligent behaviour – vision, language, memory, etc. The other starts with the low-level structure and operation of neural tissue.

It may help to invent an analogy. Suppose we don our lab coats and approach that mysterious and powerful artifact, the Bremner Building fees computer. We have little idea how computers work. But since it is nighttime and no one else is around, we are free to experiment with this one, or even dismantle it. We would keenly like to understand it.

We might take a top-down approach, studying the printouts. With a little work, we could formulate some general laws about the computer's behaviour. It seems to perform operations like addition, subtraction, alphabetization. The last of these is a little mysterious, because although we can recognize alphabetization when we see it, we're not sure how to accomplish it. So we scratch our heads and try to imagine a satisfactory method. Or we take out our stopwatches and perform some keyboard experiments, to figure out what the computer's method is. Maybe if we are very clever, we can formulate some plausible theories of what happens when the computer alphabetizes. Then we can try to fill in the details of *those* theories, and so on, until we have explained everything to our satisfaction.

A bottom-up investigation would be very different. We would wrench off the back of the machine, trace the microcircuitry etched on the chip, measure the flux at every point in the memory grid, test the changing polarity of magnetic oxide particles on the surface of the hard disk, study the sensory connections that the computer makes with the outside world via its keyboard and printer cables . . . After decades of secret nighttime labor, we might be able to make some high-level predictions about this physical system. Not that we'd know which kinds of electrical activity are important. But if someone asked us: "If fees of R4000 and R8 come in over the keyboard wires, what goes out over the printer wires?," we'd be able to do some calculations and answer, "R4008."

Now, let's leave the Bremner Building and visit the laboratory of some modern-day Frankensteins. (MIT's department of Brain and Cognitive Science will do.) Here, it is the psychologists and linguists who work primarily from the top down. Their theories tend to be *mentalistic* and *representational*; their abstractions are beliefs, goals, rules, categories, and the like. Cognitive science can hardly afford to dismiss these everyday concepts, for they have explanatory power.

The MIT neurobiologists, by contrast, work from the bottom up. They cannot actually ignore the work of their colleagues. (Imagine studying the visual system without knowing that it distinguishes objects in space, or that it is prone to optical illusions!) But their primary concern is the behaviour of individual neurons and small systems of neurons. For example, the brain contains many small visual "detectors" that respond specifically to edges, bars, color spots, directional movement, and the like. Each detector is a simple system of neurons; their operation is quite well understood by now.

In the long run, everyone hopes, the top-down and bottom-up approaches will meet in the middle.[2] After all, what we want is to explain all of mental functioning in terms of neural activity.

The Computational Perspective

Because cognitive science is indeed a science, we want our explanations to be good scientific explanations. Not only must they account for the facts; they must be plausible, elegant, and sharply defined. The goal is to answer the grand questions with mathematical rigor.

Perhaps it would be more appropriate to say computational rigor. For virtually everyone agrees that if you genuinely understand a mental process, you should be able to write

[2]Some present-day researchers are actively trying to bring the two approaches together. They want to bridge the gap by modelling high-level behaviour in a neurobiologically plausible way. Such "neural network" models are built out of many simple, neuron-like computational elements. Already it is possible to build neural networks that learn how to associate patterns, regularize incomplete or distorted patterns, predict incoming data, produce familiar sequences of behaviour, or generalize their responses for new inputs. Not all of these networks are psychologically plausible – for example, some of them learn far too slowly. The approach is promising, however.

your theory down as an explicit computer program. This guarantees that you really have thought everything through, since computers don't understand hand-waving. More important, it allows the rest of us to test the theory, even if it is very complex. We can check: Does the program really perform the behaviour in question? Does it have the same abilities, exhibit the same idiosyncrasies, make the same kinds of mistakes? In short, does it act like the mind does?

There are historical reasons for having chosen the computer as a model of mind. Other metaphors are possible: Freud imagined hydraulics, Locke a blank slate, Socrates a storehouse, and Plato a cave-dweller gazing at shadows. In our era, computers happened to arrive as part of a new Platonism, one that separated certain abstractions – such as program, process, information, and organization, or thought and knowledge – from any particular physical context. This perspective seems straightforward today. It could even have been conceived in a pre-electronic century: e.g., a musical theme retains its identity whether it is bowed, tooted, plucked, or written down. But at the time it was revolutionary. Electrical engineering and neurobiology were developing in parallel. Mathematicians of the 1930's and 1940's, building on each others' work, proved equivalences between electrical circuits in machines, neural circuits in the brain, and formulas of Boolean logic. Alan Turing invented the idea of the programmable computer, though no such machine had been built yet. Claude Shannon formalized the notion of information, that abstract stuff we measure in bits and bytes. And *all* these researchers, and more, wrote about the connections between programs and behaviour, between circuits and brains. As a result, artificial intelligence was conceived before anyone had even written a decent sorting algorithm. It was one of the first ideas of computer science.

Metaphors in science should be treated with suspicion, but some prove useful. Although the mind-as-computer metaphor may have grown from a particular conjunction of circumstances in the history of science, it seems a shrewd choice in retrospect. It has two tremendous advantages. First of all, it correctly treats mental functioning as a *process*. We can easily dismiss the old mind-body problem, or mind-brain problem, by explaining that the brain is a sort of computer (physical) and the mind its active program (non-physical). This stance certainly seems preferable to Descartes' dualism, where matter and mind are two distinct kinds of *substance*.[3]

Second, whether or not the mind does "information processing" or "data analysis" in any meaningful sense, computers are extremely general devices. Turing suspected that they are, in fact, the most general devices imaginable. He suggested that *any* precisely defined process, if it can be carried out at all, can be carried out by a vanilla computer of the sort he described. (This view is now widely held.) Thus, even if the mind does not act much like a typical computer, an appropriately large computer can always be programmed to act like a mind.

---

[3]Communication between the worlds of mind and matter was a vexing problem. Descartes once speculated that the physical body and non-physical mind might influence each other through the base of the pineal gland!

Simulation

This notion that a computer can "act like something else" is really at the heart of the artificial intelligence enterprise. It deserves some discussion. When I was a child, I had a friend who was a terrific mimic. We would beg him to imitate our teachers and friends: "Do Kevin!" "Do Mr. Smith!" "Do Miss Piggy!" And he would do it, often brilliantly. He evidently had a kind of theory of each person – not only did he understand the inner workings of Miss Piggy's accent, but he could assume the personality of Miss Piggy and say what she would be likely to say. In people, we call such behaviour acting. In machines, we call it simulation.

Modern weather forecasters run programs that simulate the atmosphere. The simulation program receives data about many high- and low-pressure zones in the forecast area. Its job is to calculate how those zones will move over the terrain, change shape, and interact with each other. The program might conclude that in three days, humid warm front 543 is likely to meet cold front 192 over Cape Town, causing a thunderstorm. And if the program acts sufficiently like the atmosphere, and cold front 192 really moves the way the program thinks it will, the prediction is likely to be right. What makes this possible is that data structures in the computer are induced to act like real objects in the world.

The principle of simulation is this:

> *Suppose* we really understand why a system behaves as it does. *Then* we can build a parallel system on a computer, which functions by the same laws and hence has all the same properties as the original.

Think about what this powerful idea means. An all-knowing economist could write a computer program that acts just like the South African economy. Perhaps the program would need millions of data structures representing self-interested agents. If the economist lowers the program's interest rate, some of the little data structures representing people go and "borrow money" from data structures representing banks. And now that there is more money to be spent, the simulated shop owners raise their prices. Result: Inflation! So lowering the interest rate results in inflation. This fact is a property of the South African economy, and for exactly the same reasons, it is also a property of the program.

Similarly, an all-knowing meteorologist could write a program that acts just like the atmosphere. *Any* process – as soon as it is completely understood – can be exactly duplicated by a program.

No one would claim, of course, that a real thunderstorm ever takes place *inside the computer*. Even if the program were so detailed as to track individual molecules of oxygen and water vapor, with a perfect little model atmosphere fashioned out of variables, the inside of the computer would not be any wetter than before. Water is by definition physical; simulated water is not.

But imagine a clever program that composes musical themes. Bad musical themes, if that is easier to imagine. Are these "simulated themes" made of different stuff from equally bad "real themes" composed by a human? Presumably not. Music is information, not matter. The water represented inside a computer won't get you wet; but the music represented inside a computer can be played.

Finally, imagine an incredibly complex program that simulates thinking. Such a program would keep track of beliefs and habits and feelings, or perhaps neurons and neurotransmitters. And it would appear to behave intelligently. Is the program's simulated intelligence something different from real, old-fashioned, human intelligence? Or are they merely two instances of the same thing?

The Turing Test

> Dad, why do we call this spaghetti?
> Well, it looks like spaghetti, doesn't it?
> Yes.
> And it's long and thin like spaghetti, isn't it?
> Yes.
> And it falls off your fork like spaghetti, doesn't it?
> Yes, but –
> (triumphantly) *So why not call it spaghetti?*

With this question, we begin moving beyond the bounds of science, or at least current science, and on into the domain of philosophy. From here on, we will be concerned with thinking computers and the nature of intelligence.

First a disclaimer. It is likely that our concept "intelligence" does not have mathematical rigor and clarity. There are undoubtedly entities that we would hesitate to call either intelligent or non-intelligent. Even so, "What is intelligence?" is not a fuzzy question; it is a very precise question about a possibly fuzzy concept.[4] It asks us to clarify the ideas we attach to the word intelligence. In other words, if we had to decide whether an entity was genuinely intelligent, what would we consider?

In 1950, Alan Turing wrote a clear and cogent article that quickly became famous. The article, published fourteen years after Turing's mathematical formulation of computing machinery, was called "Computing Machinery and Intelligence." It posed the question: "Can machines think?"

---

[4]An analogy may help here. Pearl and I are playing at flipping coins. Someone asks: "How do you define the winner of a toss?" That is a precise analytic question with a very straightforward answer – if the coin comes up heads, I win, and if it comes up tails, Pearl wins. Of course determining the winner may still be difficult in certain cases. The surface of the coin may be corroded and nearly unreadable. For certain foreign coins without heads or tails, no winner will even be defined. The point is merely that "Was Pearl the winner?" always has the same answer, or non-answer, as "Did the coin come up tails?"

Turing proposed the following practical test, modeled after a party game. Let an skeptical observer converse by teletype with both a human and a computer. If the skeptic cannot manage to tell the two apart, it is reasonable to speak of the computer as intelligent.

It is important to understand the kind of broad interrogation Turing had in mind. He provides the following sample:[5]

> Q: Please write me a sonnet on the subject of the Forth Bridge.
> A [either a human or a computer]: Count me out on this one. I never could write poetry.
> Q: Add 34957 to 70764.
> A: (Pause about 30 seconds and then give an answer) 105621.
> Q: Do you play chess?
> A: Yes.
> Q: I have my K at my K1, and no other pieces. You have only K at K6 and R at R1. It is your move. What do you play?
> A: (After a pause of 15 seconds) R-R8 mate.

And to test the intelligence of a sonnet-writing machine:

> Q: In the first line of your sonnet which reads "Shall I compare thee to a summer's day," would not "a spring day" do as well or better?
> A: It wouldn't scan.
> Q: How about "a winter's day"? That would scan all right.
> A: Yes, but nobody wants to be compared to a winter's day.
> Q: Would you say Mr. Pickwick reminded you of Christmas?
> A: In a way.
> Q: Yet Christmas is a winter's day, and I do not think Mr. Pickwick would mind the comparison.
> A: I don't think you're serious. By a winter's day one means a typical winter's day, rather than a special one like Christmas.

As Turing notes, this test successfully excludes many factors that have nothing to do with intelligence. The computer is not penalized for its inability to play soccer, or for being made of silicon rather than carbon compounds. But it must have a command of language; and it must be able to converse, on any topic, as well as a human could.

You are free, of course, to dispute the exact boundaries of the test. Perhaps you think that other behaviours are important – e.g., that the ability to judge human facial expressions is a vital part of intelligence. This ability could easily be incorporated into the test. The only change necessary would be to let the computer and the human see their interrogator

---

[5] These quotes come from Turing's article, reprinted in Hofstadter and Dennett (see Further Reading).

through a one-way window. A computer system would now require a videocamera and improved software to succeed; but the test's basic principle is the same.

In whatever form, the Turing test provides an operational approach to the question of intelligence. That is, it leaves open the question of what intelligence really is, or how it might be simulated. It simply gives us a sensible way to identify it. If someone programs a machine to produce intelligent behaviour, and the behaviour is actually *indistinguishable* from a human's, we can take that performance as a sure mark of intelligence.

Common Objections to the Turing Test

Turing's article really addresses three questions. First, under what conditions might we speak of a computer as intelligent? (Answer: When it has passed the imitation test.) Second, could a machine pass this test in principle? (Yes.) Third, will a machine ever pass this test? (Yes, in about 50 years.)

Most people have no objection to the way Turing answers the first question. They are willing to accept that *if* a machine could argue with them about sonnets, tease them playfully, act insulted, commit errors of judgment, come up with original ideas, learn, take criticism, etc., it would be hard to treat the machine as anything other than intelligent.

Most people simply claim that Turing is wrong on the second question, and that a machine could never, ever do this or that. *If* a machine could do such-and-such, then maybe it could be considered intelligent, but the possibility is too silly to even discuss. After all, how could a machine make mistakes? Make jokes? Feel depressed? No way.

There is one basic answer to such objections. *If no machine can do such-and-such, how does the brain do it?* That question should give any skeptic pause for thought; and if the skeptic cannot answer it, he or she should drop the objection. After all, your brain is a machine. It is such an amazingly complex and successful machine that its existence is very improbable, but it exists nonetheless. Your emotional responses are governed by the limbic system, at the base of the brain; if you do not believe this, a surgeon can make a convincing demonstration by removing part of yours. Any of the intellectual capacities of which you are so proud could be dispelled with a well-placed blow to the head. Your originality seems strange by comparison with most familiar machines, but it is not unheard of. As early as the 1950's, Arthur Samuel at IBM wrote a simple program that beat him regularly at checkers. He knew everything that the program did, in a sense, but the program was complicated enough that Samuel could not possibly predict how it would respond to new situations. In our society, most machines are not even as original as that program. But there is no reason *in principle* that other machines cannot be just as original as brains. Especially if they are built very much like brains.

You can take refuge in the idea that people are somehow special. They have souls, or sparks of life, or something ineffable that machines can by definition never get. These are comforting notions, and perhaps they are true. But they are not really scientific. Nor do they bear on the question at hand. The question is "Could a machine in principle pass the Turing test?," not "Could a machine in principle go to heaven?"

Turing made this observation:

> "I grant you that you can make machines do all the things you have mentioned but you will never be able to make one do X." . . . No support is usually offered for these statements. I believe they are mostly founded on the principle of scientific induction. A man has seen thousands of machines in his lifetime. From what he sees of them he draws a number of general conclusions. They are ugly, each is designed for a very limited purpose, when required for a minutely different purpose they are useless, the variety of behaviour of any one of them is very small, etc., etc. Naturally he concludes that these are necessary properties of machines in general. Many of these limitations are associated with the very small storage capacity of most machines. . . .

If we do accept that a machine could in principle pass Turing's test, the third question – "Will a machine ever pass the test?" – becomes a matter of engineering. If cognitive scientists can manage to describe the laws of the mind or brain in sufficient detail, they should be able to simulate mental processes on a computer, as meteorologists simulate the weather.

Of course, it remains an open question whether we will ever understand the mind or brain well enough. Certainly Turing's 50-year estimate has proved too optimistic. And although progress has been encouraging, there is no way to be sure that we will eventually understand as much as we'd like.

John Searle and the Chinese Room

> A human body that functions as if it were a machine and a machine that duplicates human functions are equally fascinating and frightening. Perhaps they are so uncanny because they remind us that the human body can operate without a human spirit, that body can exist without soul.
>
> –Bruno Bettelheim[6]

Many researchers did adopt Turing's view of the situation. The ultimate goal of the whole artificial intelligence enterprise, or AI, is to get a computer to pass the Turing test. It is generally agreed that this should be possible someday. The successful computer's program might or might not be modeled on the human brain, but it would have to act just the same, emotions and all.

According to a "strong AI" philosophy, such a machine should be considered just as intelligent as you or I, in every sense of the word. A Turing-test machine is not merely a

---

[6] Bruno Bettelheim, "Joey: A 'Mechanical Boy,'" p. 294. In Nancy Comley et al., *Fields of Writing* (New York: St. Martin's Press, 1984). Originally published in 1959.

device to test a psychological theory. Nor is it merely a useful labor-saving contraption, the sort to put poets and chessmasters out of work. It is a genuine sentient creature. On the evidence of its conversational skills, we may conclude that it understands; that it feels; that it has thoughts, opinions, and aspirations.

In 1980, the philosopher John Searle published a scathing critique of this philosophy. Searle imagined that he had been locked in a room with a batch of written squiggles and some obscure instructions for manipulating the squiggles, which he followed faithfully. Every so often, some further squiggles arrived on sheets of paper under the door. Searle dealt with these exactly as the instructions told him to. He would do some intermediate manipulations, end up by writing some new squiggles in response, and slip those back under the door.

Life was lonely in the little room. Fortunately, the experimenters outside the room would sometimes pass Searle little notes under the door, making conversation. Searle would answer these in English.

The catch, of course, is that the squiggles turned out to be conversation in Chinese. As far as anyone outside the room could tell, Searle was answering the Chinese messages in perfect Chinese, and the English messages in perfect English. From the outside, he appeared to understand Chinese and English equally well. But the funny thing is that Searle does not really understand a word of Chinese. He was only pushing squiggles around.

In his article, Searle concluded that a person or a computer could manipulate meaningless symbols till the cows come home, and pass the Turing test in 15 languages, without understanding anything at all. Formal rules alone, he said, do not confer the power to understand. Just following content-free instructions, whether they deal with slips of paper or variables in a computer, will never give anyone or anything the power to understand Chinese.

Of course, Searle continued, this fact does not mean no *machine* can understand. The brain is a machine, and it understands. So there must be something special about the physical makeup of the brain. Not something special about its organization, which could be duplicated stupidly by a computer or an elaborate network of water pipes, but its *biochemistry*. Brains are evidently made of the right sort of stuff to have mental states. Cars and rooms and computers and plumbing are not, because they deal in squiggles at best.

What are we to make of these claims? Are there really two fundamentally different kinds of stuff in the universe – brain tissue, which can have genuine understanding, and everything else, which can only fake it? Or is something wrong with the argument?

It is important to see just what Searle is saying. He is not opposing Turing; he cheerfully agrees that cognitive science might someday produce a machine that is indistinguishable from a person. His argument is only against the proponents of strong AI. "Sure, you may be able to put on a perfect magic show," he says. "You'd fool me along with everyone else. But nonetheless it will all be an illusion."

It is equally important to understand why he believes this. He is not arguing from principles of faith, e.g., some hypothetical spark of life that computers will never have. His objection is purely logical. It goes as follows: You could theoretically take someone's brain and reproduce it in another form, as slips of paper plus instructions. When I sit in a room and rearrange the slips of paper according to the instructions, I carry out the same processes as the brain. Aided by the papers, I can answer all questions exactly as the brain would have. Yet there is something odd. The brain understood certain things, such as Chinese. I understand none of those things. Evidently, in copying the structure of the brain onto paper, we have lost something about the brain – its consciousness! Ergo, consciousness must be a physical property of the brain.

The argument has several serious mistakes, but the most important mistake is that Searle casts himself as a privileged observer. His argument hinges on the fact that *he* never starts understanding Chinese. But remember, the room also contains one hundred billion neurons worth of information, enough to encode the abilities, the memories, and the entire personality of some simulated Chinese speaker. Those squiggles slipping out under the door answer questions about what it was like for her to grow up in Nanking during the war. Perhaps the room *as a whole* is understanding Chinese. Searle is no more significant than a little clerk labouring in a vast corporation. There may or may not be some understanding going on; but if there *is*, why should Searle be aware of it?

Douglas Hofstadter complains that Searle has hoodwinked us – that the experiment has been set up in a deceptive way.[7] We have trouble identifying with a putative "intelligent room," whose answers come in Chinese at the rate of one per century. It is much easier to identify with Searle, who happens to be our own size and speed, and who is already known to be intelligent! Nonetheless, there really are two points of view in that room, not one. Searle complains that he still doesn't understand Chinese. He contends that if anyone there really understands Chinese, it ought to be him. Who else could it possibly be? After all, he happens to be intelligent, and no one else is around . . . But if we are coaxed into accepting this last assumption, we have already denied that the room might have its own perspective.

In a miniaturized, speeded-up version of the scenario – where there is no human inside the room to identify with, but only mechanical devices – we have rather different intuitions. Hofstadter quotes Zenon Pylyshyn's parody of Searle:

> If more and more of the cells in your brain were to be replaced by integrated circuit chips, programmed in such a way as to keep the input-output *function* of each unit identical to that of the unit being replaced, you would in all likelihood just keep right on speaking exactly as you are

[7]"Reflections" on Searle's article, in Hofstadter and Dennett (see Further Reading).

> doing now except that you would eventually stop *meaning* anything by it. What we outside observers might take to be words would become for you just certain noises that circuits caused you to make.

This time, the flaws in the argument are easy to see. Most people identify with you, the person whose brain is being patched up by medical technology. When exactly does your consciousness leave you? And if one of the integrated circuits named micro-Searle wrote a tell-all article revealing that it, the little circuit, didn't mean anything at all by the noises you were making, why should anyone care?

The Consciousness Problem

I think that while Searle's conclusions are indeed wrong, his concerns are well-founded. It is genuinely not clear how consciousness and understanding could emerge from the manipulation of formal symbols. Just as there is an unbridgeable gap between "is" and "should," so that no fact about the physical universe will ever imply any moral theory, there appears to be an unbridgeable gap between "it" and "I," between objective behaviour and subjective experience.

I can easily agree that your brain categorizes and schematizes sensory input, integrates large quantities of information into coherent wholes, coordinates novel and intricately modulated sequences of muscle contractions, and modifies its activity in response to experience. I may also agree that your brain contains a detailed representation of you yourself in relation to the world, and that it accurately represents aspects of its own thought processes. But why any of this should make you *feel conscious*, indeed *feel* anything at all, I don't know. Nor do I know what additional information about the brain could possibly shed light on the question. This is a genuine puzzle.

Searle claims that only certain substances like brain tissue can enjoy subjective experience. Though such a claim is conceivably correct, I do not understand the basis for it. It appears to be an empirical statement – yet there is no way to test it empirically. For subjective states are fundamentally private. Searle only knows for sure that he is conscious; he can't tell whether I am conscious, or whether my computer system is conscious. So he would make a plausible agnostic on the consciousness question . . . but to be a believer with respect to all brains, and an atheist with respect to all computers, seems like wild speculation.

Does the Turing Test Define Intelligence?

Let us now lay aside the subject of subjective experience, since we must admit we don't know quite what it is, or how to recognize it in another. We return to discussing the nature of intelligence. Does the Turing test adequately capture the notion of an intelligent creature? If someone brought us a robot that made good dinner conversation, we would obviously be unable to tell whether it was conscious, just as you are unable to tell whether I am conscious. But would it at least be reasonable to call such a robot intelligent, in the same sense that a human is intelligent?

I used to believe firmly that the Turing test was a useful *definition* of intelligence. The test derives from the idea that if two systems behave indistinguishably, then either both are intelligent or neither is intelligent. But I now suspect that this operational position is incorrect – that the criterion of *intelligent behaviour*, conversational or otherwise, does not fully capture our usual notion of intelligence.

It should already be clear that a system can be intelligent without passing the standard Turing test. An autistic child is presumably intelligent, but cannot or will not converse. An English-speaking Martian or Vulcan might be intelligent, and still fail the test because we can always distinguish her responses from a human's responses. So the ability to pass the Turing test is clearly not *necessary* for intelligence. What I want to argue is that it's not *sufficient* either. In other words, certain systems might pass the Turing test without being intelligent at all. I'll use a series of analogies to explain why.

Suppose we move from the domain of intelligence to the domain of physics. Consider a video of a bouncing ball. The image appears to behave just as a real ball would. On every bounce, it bounces just as high as a real ball would bounce, and takes just as long. The laws of physics are impeccably observed by that image. Shall we say that the television is a simulation of the physical system?

Most of us would say no. After all, the laws of physics are not represented anywhere in the television. There is nothing inside the television representing the ball's mass, its bulk modulus, the force of gravity, the local air resistance, etc. We get the right answers only because the original ball – the one that was filmed – happened to obey the laws of physics. The original ball's movements are faithfully reproduced, but the *principles* by which it moved are nowhere captured.

(A clarification: When I say, "The laws of physics are not represented or captured in the system," I don't mean that they are somehow suspended. I only mean that neither the television nor the videotape in question has any commitment to simulating physics. The laws of physics govern the operation of the apparatus, of course, but they only act upon the apparatus; they're not explicitly built into the TV set or the videotape, any more than they're built into a bowl of mealie-pap or any other physical object. Moreover, the laws that make the apparatus work certainly aren't the ones deciding how high the image bounces. If we showed the video backwards, the ball would appear to bounce higher every time, thus violating the laws of thermodynamics. This would not in the least imply that the VCR was violating the laws of thermodynamics.)

You might argue that this is not a good analogy, because I can't interact with the image of the ball. I can't test its physical properties by asking the VCR to bounce it under low gravity, or squeeze it, or pick it up and throw it. So I have no way of knowing whether the image really simulates a physical ball in all respects. But in the Turing test, I don't just watch the computer print out clever remarks; I can ask it new questions about anything at all. Surely this allows me to probe the depth and correctness of the simulation?

Canned Responses

Well, not really, because there is still the chance that the responses might be canned. In theory, the people who built the AI program might have anticipated all 500 zillion zillion dialogues I could possibly have with the system in the next 70 years. So even if the system exhibits intelligent behaviour under a wide variety of circumstances, we have no way of being sure that it is any more principled about it than the VCR, whose image exhibits physical behaviour under a single circumstance.

To clarify the problem, let me introduce a computer program M, which does multiplication problems by looking up the answers in an enormous multiplication table. If you ask it for 57,324 times 99 billion, it checks its table and tells you the answer. We might hesitate to say that this program is really multiplying. Yes, it behaves *as if* it is multiplying, but somehow it seems to miss the point. Someone else has done all the multiplying for it and put the answers into the table. The only job the program does is *remembering*, or *looking up*. The multiplication itself took place before the program was run.

There are at least four reasonable explanations for our feeling that M doesn't "really" multiply:

(1) It goes about it differently from the way we do.

(2) It doesn't work hard enough.

(3) The table is not infinite, so the process only captures part of the idea of multiplication.

(4) It relies on a large table. The underlying rule of the table is nowhere expressed; but it is that rule, and not the table or the rest of the program, that captures the idea of multiplication.

Explanation 1 seems a little silly. Sure, the program's method is different from mine. But that alone doesn't disqualify it from being multiplication. There are several reasonable ways to multiply. Does one take the digits left to right, or right to left? Or can genuine multiplication only be performed by repeated addition ($5 \times 3 = 5 + 5 + 5$)? These distinctions seem unimportant. Although you and I might use different methods, we can still agree that both of us are genuinely multiplying numbers. If my calculator works in base 2 instead of base 10, it too might be multiplying – so long as it doesn't rely on a big table.

As fussy as explanation 1 sounds in its unvarnished form, do recognize that it is the usual argument given by anti-Turing-test philosophers. It's not enough for the computer to get the right answers, they say; the computer has to get them *in the right way*, i.e., by the same methods that humans use. The problem with the argument is that "in the right way" and "by the same methods" are nowhere defined. How closely must the computer mimic us? Taken to extremes, the argument results in Searle's claim: no computer can ever truly understand, because computers use microchips instead of brain tissue!

The problem with M seems to reside not in the novelty of its procedure, but in the procedure itself. Using a prefabricated table somehow skips over something that is essential to our idea of multiplication. So explanation 2 sounds vaguely reasonable: The program isn't working hard enough to get the answer. But that is not precise enough to be correct. M might work very hard indeed, and still fail to be multiplying. Suppose, for example, that the enormous table is not in the computer's immediate memory. It is actually written on an inconceivably huge scroll of paper in the attic of the computer center. M controls a little robot, which must navigate around the building, painstakingly look up the answer, and bring it back. M is doing far more work than a conventional multiplication program. However, nothing in the system composed of M, the robot, and the scroll of paper, is multiplying numbers.

Explanation 3 has more force. It points out that M is actually limited in a way that our idea of multiplication is not. We know (in principle) how to multiply *any* two numbers, no matter how large. But M has no such general knowledge. It can only do a finite number of multiplication problems, because its table has only a finite number of entries. It can't do all of multiplication. So it can't really multiply.

This argument is close to correct, but it is not quite enough. Even for the finite set of problems that M *can* answer, something is missing. I would deny that M even does "multiplication of 12-digit integers." There is something wrong with saying that a computer that can look up the answers to 1024 different multiplication problems is actually multiplying *anything*.

So the real difference between our multiplication and M's pseudo-multiplication is not that we humans have an infinite rule. The difference is that we have a rule at all. This is explanation 4. The program M includes an explicit table of all the answers, but not the simple procedure that gets those answers. Evidently, when we talk about multiplication, we're talking about that procedure.

Now, there is an immediate objection to explanation 4. Granted, we multiply using a small number of rules. But M uses rules too. Zillions of rules, one for every potential problem. The difference is merely quantitative. Why does it matter? What is *wrong* with multiplying from a big table? After all, we've already agreed that different definitions of multiplication are possible; and M's definition gets the right answers, doesn't it?

Multiplication is a relationship among numbers: $5 \times 1 = 5$, $5 \times 2 = 10$, $5 \times 3 = 15$ . . . If we say M is a bad definition of 12-digit integer multiplication, we must mean it does a poor job of defining that relationship. It provides a poor theory of what multiplication actually is. M misses the essence of multiplication; it just follows a clumsy procedure that happens to get the same answers.

It should be clear why the program M is a poor theory of 12-digit integer multiplication. It strikes us as a far, far bigger theory than necessary. Few principles in science are as widely respected as Occam's Razor: *The most concise theories that account for the facts are most likely to be right.* Good science recognizes that order is rarely coincidental, and furthermore, that order on a massive scale is utterly improbable without some underlying principle to account for it.

Implicit Theories

So good theories are supposed to be both accurate and "parsimonious," i.e., roughly as small as possible. All right. But why should a principle about scientific theories have anything to do with the meaning of the word "multiplication"?

Interestingly, most process words in English refer to actual methods. We do not usually embrace operationalism at all. Thus, "doing housework" means something much more specific than "getting the housework done," as the absurdity of this dialogue makes clear:

> Speaker 1: Today I washed the dishes, scrubbed the floors, and sewed the curtains.
> Speaker 2: Gee! You must be tired.
> Speaker 1: Not really. I paid my brother to do the work.

Many actions can have the same effect, but hiring someone can't be described as "doing housework"! Similarly, when we talk about evolution, we don't just mean "the increase of adaptiveness by any means." The very term *specifies a particular mechanism* by which adaptiveness is increased.

No single mechanism is specified when we talk about multiplication or thinking. In one case we don't care exactly how it's done, and in the other we just don't know. But I suspect that when we use those words, we are still committed to the idea of some "reasonable" mechanism at work.

Furthermore, I think we can say what a "reasonable" mechanism is: anything that provides a good theory of the behaviour that is being accomplished. A good theory, again, is a description that is both accurate and parsimonious.

Implicit Theories of Multiplication – Good and Bad

Thus we cannot accept that M is really doing multiplication, because M doesn't do a good job of describing what multiplication really is. M's "theory" is that multiplication is just a truckload of numbers – particular numbers! – and a means for pulling some of them out when necessary. You have to know all of the numbers to understand what multiplication is, because with different numbers, it might turn out to be addition instead.

We have trouble swallowing this truck-sized account as reasonable. We have much more straightforward ways of describing "multiplying behaviour." Several such ways, in fact. It is much more parsimonious, and just as accurate, to define the result of multiplication as the result of repeated addition. Or the result of the procedure you learned in school. Unless M implements a reasonable procedure like one of these, we hesitate to say it is really doing multiplication.

Just to indicate that the problem with M really is a quantitative matter, a question of parsimony and the relative size of theories, let me mention something you may already have noticed. We humans *do* perform a restricted amount of table lookup in multiplication. Most of us know by heart our multiplication table from 0×0 to 9×9, and we will consult it readily when finding 703×495. Even if we were doing binary multiplication, we would still need to know the tables up through 1×1.

I don't think that relying on these *small* tables does too much violence to the idea of multiplication. Indeed, one might even be prepared to allow that the tables are themselves reasonably parsimonious theories of one-digit multiplication. If an earlier version of M only dealt with integers from 0 to 9, and solved the problems by table lookup alone, might it have been doing genuine one-digit multiplication? What if it only dealt with integers from 0 to 1?

A New Definition of "Multiplication"

> Science is facts; just as houses are made of stones, so is science made of facts; but a pile of stones is not a house and a collection of facts is not necessarily science.
>
> –Henri Poincaré

We're now equipped to tell whether an arbitrary device really does multiplication or only fakes it. Assume we are given a device that can solve multiplication problems. It may do other things as well, but we are only concerned with this one ability.

First, we analyze multiplication from the top down. We construct all possible parsimonious theories of how to answer multiplication problems correctly. These are our *theories of behaviour*. Note that these theories are purely formal and mathematical. They make no mention of brains, circuits, physics, and the like. Of course, we ignore any huge-table theories of multiplication, because these are not very parsimonious!

Next, we take the device in question and analyze it from the bottom up. The goal here is to explain how the physical device achieves its behaviour. If the device is a computer, we start at some level that is already well understood (electrons, circuits, or programs) and start making generalizations about causal relationships. Eventually, we construct one or more complete and parsimonious accounts of why this machine answers multiplication problems as it does. We call these *theories of operation*. They deal with physical objects like individual circuits, and more abstractly, common types of circuits such as logic gates. They may also deal with even more abstract objects, such as integers, whose properties are respected by the machine's behavior – an observation that is useful to state in the theory because it provides significant though incomplete information about the arrangement of the circuits.

For some devices, it may turn out that some parsimonious theory of the physical operation includes some parsimonious theory of the formal behaviour, namely multiplication. It is in exactly these cases, I believe, that we're comfortable saying that the device *truly multiplies*, or *embodies multiplication*.

For example, suppose we're studying my friend Duma the mathematician. We commission several concise analyses of his brain. One of these concludes that he has the unusual habit of multiplying everything in base 2, using a shift-add procedure. It goes on to describe how the shift-add procedure is managed with neurons – but we have already found out what we need to know. Binary shift-add happens to be a parsimonious description of multiplication. So Duma's brain embodies multiplication.

If the device in question does not embody multiplication, but gets the right answers anyway, we can apply some different words to it. We say that it *pseudo-multiplies*, or *mimics multiplication*. M is an example of a pseudo-multiplier. It is not a true multiplier, because the most concise theories concerning M fail to include any simple theory of multiplying behaviour. All the concise theories of M's operation only record that M goes to an enormous table.

Some less concise theories of operation might also mention the simple rule underlying the table, as a kind of footnote. But such a footnote doesn't help explain how M gets the answers; it only increases the size of the theory. So these theories of operation have irrelevant elements. They are not parsimonious theories; we do not consider them.

A New Criterion for Intelligence

> Logic is like the sword – those who appeal to it shall perish by it. –Samuel Butler

I have held to the multiplication example because it is easy to describe. But by strict analogy, we can decide whether a machine that passes the Turing test is truly intelligent or just pseudo-intelligent. Remember that robot who came to dinner? We simply open up our guest during coffee and biscuits, study how he works, and ponder whether his internal mechanisms embody a sufficiently parsimonious description of intelligent conversational behaviour.

If the robot is driven by a digital computer with a short program of a billion instructions, that is a good sign. Even if the program is uncomfortably long, perhaps we can establish that most of it deals with vision and locomotion, and that conversational behaviour is accomplished in a reasonable space. What we are really worried about is that the robot might be packed top to bottom with microfiche full of dinner conversation on every likely topic. That would *not* be a concise or even a correct embodiment of human conversational behaviour, even if it somehow managed to fool us for a while.

A truly intelligent program need not be modeled on principles of human intelligence, though that is arguably the safest way to write one. It is vaguely imaginable that the robot might embody intelligence via a parsimonious theory that is not even *mentalistic*. That is, perhaps some good theory can account for intelligent behaviour without explicit reference to beliefs, categories, goals, a self-concept, or other objects of psychological interest. But I doubt such an account exists. Mentalistic abstractions have great explanatory power. It would be quite difficult to explain human conversation without them; ditto for the conversation of an Turing-test computer.

If we later attend a dinner at the robot's home, should we expect to be dissected in return? Probably. There is no guarantee that we humans are above investigation. If we adopt this new test of intelligence, which requires intelligent devices to be built in a particular way, we should recognize that people just might fail it! Such disappointments are among the risks of philosophy. If our common notion of intelligence really does go beyond operational tests, as I believe it does, then it necessarily makes reference to facts about our brains. We know very little about our brains, and perhaps the true facts will not be to our liking. But that is simply too bad.

Whatever our notion of intelligence may have to say about the adequacy of the brain, it specifies at the same time that humans *are* intelligent. So if our brains somehow turn out to do very simple operations by improbably complex seat-of-the-pants tricks, or by using big tables, we will be both truly intelligent and pseudo-intelligent. We should then have to admit that our notion of intelligence is not very consistent.

Alien Intelligences

> I would be very ashamed of my civilization if we did not try to find out if there is life in outer space.
> –Carl Sagan

By way of closing, I should point out that the new test for intelligence might be extended to detect forms of intelligence in Martians, Vulcans, and autistic children, not to mention autistic AI programs. This is a helpful feature, because the Turing test cannot detect intelligence in such individuals.

The problem with Martians is defining their behaviour. We may suspect that they are saying something to us as they blow gently in the wind. And perhaps the swirling of ammonia through their many tubules is their way of thinking. But we have no way of telling. This makes it difficult for us to build a theory of operation for them. We are supposed to explain something in terms of their biology, but what?

If we had identified relevant Martian activities and had a theory of operation for them, we could pull out our bag labeled Parsimonious Theories of Intelligent Human Behaviour, sit down on a red rock, and start sifting through for a good match. The theory of operation describes various abstract properties of the Martian's physical system. Perhaps some theory in the bag boasts theoretical elements with roughly similar properties. Theory X, for example. Perhaps the Martian has no language subsystem like the one in theory X. But it does have something like what theory X calls short-term and long-term memory, and – yes, there it goes! – a self-symbol . . .

In the absence of a definitive theory of operation for Martians, the obvious solution is to permit any parsimonious theory of operation that works. This is a reasonable idea. Does

Martian physiology have *any* patterns of events that we can map onto human psychological constructs? Perhaps we could show that some undistinguished pool of ammonia swirls assumes the same configurations again and again, rather like a long-term memory. And furthermore, as other bubbles and eddies pass through, they necessarily nudge the pool toward one familiar pattern or another . . .

If we did find such a pool of ammonia in the Martian body, a pool that acts much like human memory, we could treat it as an abstract "memory device" and build a theory of operation for it. Such a theory wouldn't have to explain all of Martian behaviour. It would just have to explain how the pool worked. It would have to prove that this collection of ammonia molecules necessarily behaves according to certain formal laws of behaviour, laws similar to those that characterize human memory.

Now, if the parsimonious theory of the pool's operation turns out to be a parsimonious account of the formal laws, then we can say that the pool embodies long-term memory. We still don't know whether the Martians are truly intelligent, if that word has meaning here, and we certainly don't know whether they're conscious, but at least we know they have true long-term memories.

Some people (Searle among others) are uncomfortable with such procedures. "If you start finding minds in pools of ammonia," they ask, "what's to stop minds from showing up in every bottle of champagne? With a sufficiently complicated theory of operation, can't you make *anything* look like a mental process?"

This objection should be taken seriously. But the answer is that we do not consider "sufficiently complicated" theories, but only parsimonious ones. We can claim that the champagne bottle is a model of the mind of Bertrand Russell, if we like. We can point to six bubbles in the bottle and claim that these represent the concept of the Eiffel Tower, and we can claim that they are going sideways in order to prove a syllogism. Well and good. But close your eyes for a few seconds, and look again. Uh-oh . . . Russell's incisive mind appears to be going insane. Unfortunately, the causal laws that govern the bubbles are very different from the causal laws that govern mental representations. No mind here.

A scheme for mapping champagne onto mind would have to be far more complicated. Iif it exists at all, it would have to identify concepts with extremely tortured classes of configurations of bubbles (not six here and six there). If anyone doubts this, let him come see me. I will hand him a bottle of champagne, and ask him to name a finite representational code under which the physical properties of the champagne necessarily generate the Fibonacci sequence forever at one term per second. Compared to mental processes, this one ought to be trivial. But if he can do it I'll gladly let him keep the bottle.

Further Reading

This completes our short tour of cognitive science and cognitive philosophy. I hope it has provided a sense of the cognitivist perspective – both toward the human mind and toward the possibility of non-human minds.

I am convinced that cognitive science is the most exciting field in theoretical science today. The human mind is the most complex, intriguing and (to us) important system we know, and it is still largely unexplored. At the present moment, the best history of the field is *The Mind's New Science: A History of the Cognitive Revolution*, by Howard Gardner (New York: Basic Books, 1985). Gardner spends about half the book discussing specific contributions from philosophy, psychology, computer science, linguistics, anthropology, and neuroscience.

An excellent introduction to problems in cognitive philosophy is provided by *The Mind's I: Fantasies and Reflections on Self and Soul* (Sussex: Harvester Press, 1981). This eclectic collection, which includes the papers by Turing and Searle, also offers many wonderfully readable essays, dialogues, thought experiments and short stories touching on the nature of mind. It is edited by Douglas Hofstadter and Daniel Dennett; they provide their own reflections on each piece.

One of the footnotes to this paper mentions the exciting area of neural networks. This topic is not covered in Gardner. The standard introduction to the subject is a two-volume collection of papers, edited by David Rumelhart and Jay McClelland, called *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (Cambridge MA: MIT Press, 1986). A third volume comes with a disk of computer simulations.