Introducing the Reading

In these two selections from *Minds, Brains, and Science*, John Searle takes on the mind/body problem and the question of artificial intelligence. Like most contemporary philosophers, Searle is a materialist. But Searle thinks that many contemporary philosophers have not adequately understood the problem of mind and its relation to the brain, and so confusion reigns. He attempts to address this confusion through his two theses. As you think about Searle’s account of these two theses, think about some of the following questions:

1. How does Searle explain the nature of the mind?
2. Searle claims that all mental phenomena are caused by processes going on in the brain. Would Descartes agree with this? Explain why or why not.
3. Do you think that Searle would agree with the belief of many Christians that there is life after death? Explain why or why not.

In the second selection, Searle gives an argument that attempts to show that no computer could be intelligent: for even if a computer could pass the Turing test, it would not really be thinking. It would, at best, be *simulating* thinking, but this doesn’t mean that there would really be any thinking going on at all.

Searle is responding to a research program, inspired by Turing, that seeks to build an artificial, symbol-manipulating machine that could genuinely think. This is the research program that Searle calls “Strong AI.” For Strong AI, a mind is just like a computer that can manipulate symbols, or information, and come out with a response. Searle is going to suggest that this assumption, and the whole theory on which it is based, is wrong: for we could have a system that could manipulate signs and symbols just in the way that a human being does – and so could pass the Turing test – but which would not at all be intelligent. To show this, Searle gives us a thought experiment that envisions such a system. The system is called the “Chinese room.” It’s designed to pass a Turing test in Chinese: to carry on a conversation just like a native Chinese speaker would. People go up to the room and submit questions in Chinese, perhaps on little slips of paper. After a while an answer comes out on another slip of paper. We can imagine such a system “behaving” just like a human Chinese speaker – giving all the right answers to questions about the weather, about sports, about poetry, etc – in short, behaving verbally just like a human being would. This system would be a reliable simulation of a Chinese speaker.

But inside the room, there need not be anything or anyone that can actually understand Chinese. Imagine that you or I are locked inside the room. There are baskets of Chinese symbols, and also a rule-book for manipulating them. You follow the rules, just as the book says, to go from the signs that are put into the room to the signs that are sent out.
The rule book acts like a “program” for the manipulation of the signs, just as an AI program instructs a computer to manipulate signs in a particular set of ways. The system as a whole acts just like a native, thinking Chinese speaker. But nothing in the system can actually understand Chinese! Nothing is actually thinking the thoughts that appear to be “behind” the verbal responses. How does Searle move from this point to the conclusion that computers can’t think? See if you can provide the remaining moves in his argument.

pp. 13-15

ONE
THE MIND-BODY PROBLEM

For thousands of years, people have been trying to understand their relationship to the rest of the universe. For a variety of reasons many philosophers today are reluctant to tackle such big problems. Nonetheless, the problems remain, and in this book I am going to attack some of them.

At the moment, the biggest problem is this: We have a certain commonsense picture of ourselves as human beings which is very hard to square with our overall ‘scientific’ conception of the physical world. We think of ourselves as conscious, free, mindful, rational agents in a world that science tells us consists entirely of mindless, meaningless physical particles. Now, how can we square these two conceptions? How, for example, can it be the case that the world contains nothing but unconscious physical particles, and yet that it also contains consciousness? How can a mechanical universe contain intentionalistic human beings – that is, human beings that can represent the world to themselves? How, in short, can an essentially meaningless world contain meanings?

Such problems spill over into other more contemporary sounding issues: How should we interpret recent work in computer science and artificial intelligence – work aimed at making intelligent machines? Specifically, does the digital computer give us the right picture of the human mind? And why is it that the social sciences in general have not given us insights into ourselves comparable to the insights that the natural sciences have given us into the rest of nature? What is the relation between the ordinary, commonsense explanations we accept of the way people behave and scientific modes of explanation?

In this first chapter, I want to plunge right into what many philosophers think of as the hardest problem of all: What is the relation of our minds to the rest of the universe? This, I am sure you will recognize, is the traditional mind-body or mind-brain problem. In its contemporary version it usually takes the form: how does the mind relate to the brain?

I believe that the mind-body problem has a rather simple solution, one that is consistent both with what we know about neurophysiology and with our commonsense conception of the nature of mental states – pains, beliefs, desires and so on. But before presenting that solution, I want to ask why the mind-body problem seems so intractable. Why do we still have in philosophy and psychology after all these centuries a ‘mind-body problem’ in a way that we do not have, say, a ‘digestion-stomach problem’? Why does the mind seem more mysterious than other biological phenomena?
I am convinced that part of the difficulty is that we persist in talking about a twentieth-century problem in an outmoded seventeenth-century vocabulary. When I was an undergraduate, I remember being dissatisfied with the choices that were apparently available in the philosophy of mind: you could be either a monist or a dualist. If you were a monist, you could be either a materialist or an idealist. If you were a materialist, you could be either a behaviourist or a physicalist. And so on. One of my aims in what follows is to try to break out of these tired old categories. Notice that nobody feels he has to choose between monism and dualism where the "digestion-stomach problem" is concerned. Why should it be any different with the 'mind-body problem'?

But, vocabulary apart, there is still a problem or family of problems. Since Descartes, the mind-body problem has taken the following form: how can we account for the relationships between two apparently completely different kinds of things? On the one hand, there are mental things, such as our thoughts and feeling; we think of them as subjective, conscious, and immaterial. On the other hand, there are physical things; we think of them as having mass, as extended in space, and as causally interactiing with other physical things. Most attempted solutions to the mind-body problem wind up by denying the existence of, or in some way downgrading the status of, one or the other of these types of things. Given the successes of the physical sciences, it is not surprising that in our stage of intellectual development the temptation is to downgrade the status of mental entities. So, most of the recently fashionable materialist conceptions of the mind – such as behaviourism, functionalism, and physicalism – end up by denying, implicitly or explicitly, that there are any such things as minds as we ordinarily think of them. That is, they deny that we do really intrinsically have subjective, conscious, mental states and that they are as real and as irreducible as anything else in the universe.

Now why do they do that? Why is it that so many theorists end up denying the intrinsically mental character of mental phenomena? If we can answer that question, I believe that we will understand why the mind-body problem has seemed so intractable for so long.

pp. 17-23

The first thesis I want to advance toward 'solving the mind-body problem' is this:

_Mental phenomena, all mental phenomena whether conscious or unconscious, visual or auditory, pains, tickles, itches, thoughts, indeed, all of our mental life, are caused by processes going on in the brain._

To get a feel for how this works, let's try to describe the causal processes in some detail for at least one kind of mental state. For example, let's consider pains. Of course, anything we say now may seem wonderfully quaint in a generation, as our knowledge of how the brain works increases. Still, for form of the explanation can remain valid even though the details are altered. On current views, pain signals are transmitted from sensory nerve endings to the spinal cord by at least two types of fibres – there are Delta A fibres, which are specialized for prickling sensations, and C fibres, which are specialized for burning and aching sensations. In the spinal cord, they pass through a region called the...
tract of Lissauer and terminate on the neurons of the cord. As the signals go up the spine, they enter the brain by two separate pathways: the prickling pain pathway and the burning pain pathway. Both pathways go through the thalamus, but the prickling pain is more localised afterwards in the somato-sensory cortex, whereas the burning pain pathway transmits signals, not only upwards into the cortex, but also laterally into the hypothalamus and other regions of the brain. Because of these differences, it is much easier for us to localize a prickling sensation — we can tell fairly accurately where someone is sticking a pin into our skin, for example — whereas burning and aching pains can be more distressing because they activate more of the nervous system. The actual sensation of pain appears to be caused both by the stimulation of the basal regions of the brain, especially the thalamus, and the stimulation of the somato-sensory cortex.

Now for the purposes of this discussion, the point we need to hammer home is this: our sensations of pains are caused by a series of events that begin at free nerve endings and end in the thalamus and in other regions of the brain. Indeed, as far as the actual sensations are concerned, the events inside the central nervous system are quite sufficient to cause pains—we know this both from the phantom-limb pains felt by amputees and the pains caused by artificially stimulating relevant portions of the brain. I want to suggest that **what is true of pain is true of mental phenomena generally**. To put it crudely, and counting all of the central nervous system as part of the brain for our present discussion, everything that matters for our mental life, all of our thoughts and feelings, are caused by processes inside the brain. As far as causing mental states is concerned, the crucial step is the one that goes on inside the head, not the external or peripheral stimulus. And the argument for this is simple. If the events outside the central nervous system occurred, but nothing happened in the brain, there would be no mental events. But if the right things happened in the brain, the mental events would occur even if there was no outside stimulus. (And that, by the way, is the principle on which surgical anaesthesia works: the outside stimulus is prevented from having the relevant effects on the central nervous system.)

But if pains and other mental phenomena are caused by processes in the brain, one wants to know: what are pains? What are they really? Well, in the case of pains, the obvious answer is that they are unpleasant sorts of sensations. But that answer leaves us unsatisfied because it doesn’t tell us how pains fit into our overall conception of the world.

Once again, I think the answer to the question is obvious, but it will take some spelling out. To our first claim — that pains and other mental phenomena are caused by brain processes, we need to add a second claim:

**Pains and other mental phenomena just are features of the brain (and perhaps the rest of the central nervous system).**

One of the primary aims of this chapter is to show how **both** of these propositions can be true together. How can it be both the case that brains cause minds and yet minds just are features of brains? I believe it is the failure to see how both these propositions can be true together that has blocked a solution to the mind-body problem for so long. There are different levels of confusion that such a pair of ideas can generate. If mental and physical phenomena have cause and effect relationships, how can one be a feature of the other?
Wouldn’t that imply that the mind caused itself – the dreaded doctrine of *causa sui*? But at the bottom of our puzzlement is a misunderstanding of causation. It is tempting to think that whenever A causes B there must be two discrete events, one identified as the cause, the other identified as the effect; that all causation functions in the same way as billiard balls hitting each other. This crude model of the causal relationships between the brain and the mind inclines us to accept some kind of dualism; we are inclined to think that events in one material realm, the ‘physical’, cause events in another insubstantial real, the ‘mental’. But that seems to me a mistake. And the way to remove the mistake is to get a more sophisticated concept of causation. To do this, I will turn away from the relations between mind and brain for a moment to observe some other sorts of causal relationships in nature.

A common distinction in physics is between micro- and macro-properties of systems – the small and large scales. Consider, for example, the desk at which I am now sitting, or the glass of water in front of me. Each object is composed of micro-particles. The micro-particles have features at the level of molecules and atoms as well as at the deeper level of sub-atomic particles. But each object also has certain properties such as the solidity of the table, the liquidity of the water, and the transparency of the glass, which are surface or global features of the physical systems. Many such surface or global properties can be causally explained by the behaviour of elements at the micro-level. For example, the solidity of the table in front of me is explained by the lattice structure occupied by the molecules of which the table is composed. Similarly, the liquidity of the water is explained by the nature of the interactions between the H$_2$O molecules. Those macro-features are causally explained by the behaviour of elements at the micro-level.

I want to suggest that this provides a perfectly ordinary model for explaining the puzzling relationships between the mind and the brain. In the case of liquidity, solidity, and transparency, we have no difficulty at all in supposing that the surface features are *caused by* the behaviour of elements at the micro-level, and at the same time we accept that the surface phenomena *just are* features of the very systems in question. I think the clearest way of stating this point is to say that the surface feature is both *caused by* the behaviour of micro-elements, and at the same time is *realised in* the system that is made up of the micro-elements. There is a cause and effect relationship, but at the same time, the surface features are just higher level features of the very system whose behaviour at the micro-level causes those features.

In objecting to this someone might say that liquidity, solidity, and so on are identical with features of the micro-structure. So, for example, we might just define solidity as the lattice structure of the molecular arrangement, just as heat often is identified with the mean kinetic energy of molecule movements. This point seems to me correct but not really an objection to the analysis that I am proposing. It is a characteristic of the progress of science that an expression that is originally defined in terms of surface features, features accessible to the senses, is subsequently defined in terms of the micro-structure that causes the surface features. Thus, to take the example of solidity, the table in front of me is solid in the ordinary sense that it is rigid, it resists pressure, it supports books, it is not easily penetrable by most other objects such as other tables, and so on. Such is the commonsense notion of solidity. And in a scientific vein one can define solidity as whatever micro-structure causes these gross observable features. So one can then say either that solidity just is the lattice structure of the system of molecules and that solidity
so defines causes, for example, resistance to touch and pressure. Or one can say that solidity consists of such high level features as rigidity and resistance to touch and pressure and that it is caused by the behaviour of elements at the micro-level.

If we apply these lessons to the study of the mind, it seems to me that there is no difficulty in accounting for the relations of the mind to the brain in terms of the brain’s functioning to cause mental states. Just as the liquidity of the water is caused by the behaviour of elements at the micro-level, and yet at the same time it is a feature realised in the system of micro-elements, so in exactly that sense of ‘caused by’ and ‘realised in’ mental phenomena are caused by processes going on in the brain at the neuronal or modular level, and at the same time they are realised in the very system that consists of neurons. And just as we need the micro/macro distinction for any physical system, so for the same reasons we need the micro/macro distinction for the brain. And though we can say of a system of particles that it is 10°C or it is solid or it is liquid, we cannot say of any given particle that this particle is solid, this particle is liquid, this particle is 10°C. I can’t for example reach into a glass of water, pull out a molecule and say: ‘This one’s wet’.

In exactly the same way, as far as we now anything at all about it, thought we can say of a particular brain: ‘This brain is conscious’, or: ‘This brain is experiencing thirst or pain’, we can’t say of any particular neuron in the brain: ‘This neuron is in pain, this neuron is experiencing thirst’. To repeat this point, though there are enormous empirical mysteries about how the brain works in detail, there are no logical or philosophical or metaphysical obstacles to accounting for the relation between the mind and the brain in terms that are quite familiar to us from the rest of nature. Nothing is more common in nature than for surface features of a phenomenon to be both caused by and realised in a micro-structure, and those are exactly the relationships that are exhibited by the relation of mind to brain.

To summarise: on my view, the mind and body intact, but they are not two different things, since mental phenomena just are features of the brain. One way to characterise this position is to see it as an assertion of both physicalism and mentalism. Suppose we define ‘naive physicalism’ to be the view that all exists in the world are physical particles with their properties and relations. The power of the physical model of reality is so great that it is hard to see how we can seriously challenge naive physicalism. And let us define ‘naive physicalism’ to be the view that mental phenomena really exist. There really are mental states; some of them are conscious; many have intentionality; they all have subjectivity; and many of them function causally in determining physical events in the world. The thesis of this first chapter can now be stated quite simply. Naive mentalism and naive physicalism are perfectly consistent with each other. Indeed, as far as we know anything about how the world works, they are not only consistent, they are both true.
TWO
CAN COMPUTERS THINK?

In the previous chapter, I provided at least the outlines of a solution to the so-called ‘mind-body problem’. Though we do not know in detail how the brain functions, we do know enough to have an idea of the general relationships between brain processes and mental processes. Mental processes are caused by the behaviour of elements of the brain. At the same time, they are realised in the structure that is made up of those elements. I think this answer is consistent with the standard biological approaches to biological phenomena. Indeed, it is a kind of commonsense answer to the question, given what we know about how the world works. However, it is very much a minority point of view. The prevailing view in philosophy, psychology, and artificial intelligence is one which emphasizes the analogies between the functioning of the human brain and the functioning of digital computers. According to the most extreme version of this view, the brain is just a digital computer and the mind is just a computer program. One could summarise this view – I call it ‘strong artificial intelligence’, or ‘strong AI’ – by saying that the mind is to the brain, as the program is to the computer hardware.

This view has the consequence that there is nothing essentially biological about the human mind. the brain just happens to be one of an indefinitely large number of different kinds of hardware computers that could sustain the programs which make up human intelligence. On this view, any physical system whatever that had the right program with the right inputs and outputs would have a mind in exactly the same sense that you and I have minds. So, for example, if you made a computer out of old beer cans powered by windmills; if it had the right program, it would have to have a mind. And the point is not that for all we know it might have thoughts and feelings, but rather that it must have thoughts and feelings, because that is all there is to having thoughts and feelings: implementing the right program.

Most people who hold this view think we have not yet designed programs which are minds. But there is pretty much general agreement among them that it’s only a matter of time until computer scientists and workers in artificial intelligence design the appropriate hardware and programs which will be the equivalent of human brains and minds. These will be artificial brains and minds which are in every way the equivalent of human brains and minds.

Many people outside of the field of artificial intelligence are quite amazed to discover that anybody could believe such a view as this. So, before criticizing it, let me give you a few examples of the things that people in this field have actually said. Herbert Simon of Carnegie-Mellon University says that we already have machines that can literally think. There is no question of waiting for some future machine, because existing digital computers already have thoughts in exactly the same sense that you and I do. Well, fancy that! Philosophers have been worried for centuries about whether or not a machine could think, and now we discover that they already have such machines at Carnegie-Mellon.

Simon’s colleague Alan Newell claims that we have now discovered (and notice that Newell says ‘discovered’ and not ‘hypothesised’ or ‘considered the possibility’, but we have discovered) that intelligence is just a matter of physical symbol manipulation; it has
no essential connection with any specific kind of biological or physical wetware or hardware. Rather, any system whatever that is capable of manipulating physical symbols in the right way is capable of intelligence in the same literal sense as human intelligence of human beings. Both Simon and Newell, to their credit, emphasise that there is nothing metaphorical about these claims; they mean them quite literally. Freeman Dyson is quoted as having said that computers have an advantage over the rest of us when it comes to evolution. Since consciousness is just a matter of formal processes, in computers these formal processes can go on in substances that are much better able to survive in a universe that is cooling off than beings like ourselves made of our wet and messy materials. Marvin Minsky of MIT says that the next generation of computers will be so intelligent that we will ‘be lucky if they are willing to keep us around the house as household pets’. My all-time favourite in the literature of exaggerated claims on behalf of the digital computers is from John McCarthy, the inventor of the term ‘artificial intelligence’. McCarthy says even ‘machines as simple as thermostats can be said to have beliefs’. And indeed, according to him, almost any machine capable of problem-solving can be said to have beliefs. I admire McCarthy’s courage. I once asked him: ‘What beliefs does your thermostat have?’ And he said: ‘My thermostat has three beliefs – it’s too hot in here, it’s too cold in here, and it’s just right in here.’ As a philosopher, I like all these claims for a simple reason. Unlike most philosophical theses, they are reasonably clear, and they admit of a simple and decisive refutation. It is this refutation that I am going to undertake in this chapter.

The nature of the refutation has nothing whatever to do with any particular stage of computer technology. It is important to emphasise this point because the temptation is always to think that the solution to our problems must wait on some as yet uncreated technological wonder. But in fact, the nature of the refutation is completely independent of any state of technology. It has to do with the very definition of a digital computer, with what a digital computer is.

It is essential to our conception of a digital computer that its operations can be specified purely formally; that is, we specify the steps in the operation of the computer in terms of abstract symbols – sequences of zeroes and ones printed on a tape, for example. A typical computer ‘rule’ will determine that when a machine is in a certain state and it has a certain symbol on its tape, then it will perform a certain operation such as reading the symbol or printing another symbol and then enter another state such as moving the tape one square to the left. But the symbols have no meaning; they have no semantic content; they are not about anything. They have to be specified purely in terms of their formal or syntactical structure. The zeroes and ones, for example, are just numerals; they don’t even stand for numbers. Indeed, it is this feature of digital computers that makes them so powerful. One and the same type of hardware, if it is appropriately designed, can be used to run an indefinite range of different programs. And one and the same program can be fun on an indefinite range of different types of hardwares.

But this feature of programs, that they are defined purely formally or syntactically, is fatal to the view that mental processes and program processes are identical. And the reason can be stated quite simply. There is more to having a mind than having formal or syntactical processes. Our internal mental states, by definition, have certain sorts of contents. If I am thinking about Kansas City or wishing that I had a cold beer to drink or wondering if there will be a fall in interest rates, in each case my mental state has a
certain mental content in addition to whatever formal features it might have. That is, even if my thoughts occur to me in strings of symbols, there must be more to the thought than the abstract strings, because strings by themselves can’t have any meaning. If my thoughts are to be about anything, then the strings must have a meaning which makes the thoughts about those things. In a word, the mind has more than a syntax, it has a semantics. The reason that no computer program can ever be a mind is simply that a computer program is only syntactical, and minds are more than syntactical. Minds are semantical, and the sense that they have more than a formal structure, they have a content.

To illustrate this point I have designed a certain thought experiment. Imagine that a bunch of computer programmers have written a program that will enable a computer to simulate the understanding of Chinese. So, for example, if the computer is given a question in Chinese, it will match the question against its memory, or data base, and produce appropriate answers to the questions in Chinese. Suppose for the sake of argument that the computer’s answers are as good as those of a native Chinese speaker. Now then, does the computer, on the basis of this, understand Chinese, does it literally understand Chinese, in the way that Chinese speakers understand Chinese? Well, imagine that you are locked in a room, and in this room are several baskets full of Chinese symbols. Imagine that you (like me) do not understand a word of Chinese, but that you are given a rule book in English for manipulating these Chinese symbols. The rules specify the manipulations of their symbols purely formally, in terms of their syntax, not their semantics. So the rule might say: ‘Take a squiggle-squiggle sign out of basket number one and put it next to a squoggle-squoggle sign from basket number two.’ Now suppose that some other Chinese symbols are passed into the room, and that you are given further rules for passing back Chinese symbols out of the room. Suppose that unknown to you the symbols passed into the room are called ‘questions’ by the people outside the room, and the symbols you pass back out of the room are called ‘answers to the questions’. Suppose, furthermore, that the programmers are so good at designing the programs and that you are so good at manipulating the symbols, that very soon your answers are indistinguishable from those of a native Chinese speaker. There you are locked in your room shuffling your Chinese symbols and passing out Chinese symbols in response to incoming Chinese symbols. On the basis of the situation as I have described it, there is no way you could learn any Chinese simply by manipulating these formal symbols.

Now the point of the story is simply this: by virtue of implementing a formal computer program from the point of view of an outside observer, you behave exactly as if you understood Chinese, but all the same you don’t understand a word of Chinese. But if going through the appropriate computer program for understanding Chinese is not enough to give you an understanding of Chinese, then it is not enough to give any other digital computer an understanding of Chinese. And again, the reason for this can be stated quite simply. If you don’t understand Chinese, then no other computer could understand Chinese because no digital computer, just by virtue of running a program, has anything that you don’t have. All that the computer has, as you have, is a formal program for manipulating uninterpreted Chinese symbols. To repeat, a computer has a syntax, but not semantics. The whole point of the parable of the Chinese room is to remind us of a fact that we knew all along. Understanding a language, or indeed, having mental states at
all, involves more than just having a bunch of formal symbols. It involves having an interpretation, or a meaning attached to those symbols. And a digital computer, as defined, cannot have more than just formal symbols because the operation of the computer, as I said earlier, is defined in terms of its ability to implement programs. And these programs are purely formally specifiable – that is, they have no semantic content.

We can see the force of this argument if we contrast what it is like to be asked and to answer questions in English, and to be asked and to answer questions in some language where we have no knowledge of any of the meanings of the words. Imagine that in the Chinese room you are also given questions in English about such things as your age or your life history, and that you answer these questions. What is the difference between the Chinese case and the English case? Well, again, if like me you understand no Chinese and you do understand English, then the difference is obvious. You understand the questions in English because they are expressed in symbols whose meanings are known to you. Similarly, when you give the answers in English you are producing symbols which are meaningful to you. But in the case of the Chinese, you have none of that. In the case of the Chinese, you simply manipulate formal symbols according to a computer program, and you attach no meaning to any of the elements.