The Problem of Consciousness*

John R. Searle

Abstract: This paper attempts to begin to answer four questions. 1. What is consciousness? 2. What is the relation of consciousness to the brain? 3. What are some of the features that an empirical theory of consciousness should try to explain? 4. What are some common mistakes to avoid?

The most important scientific discovery of the present era will come when someone -- or some group -- discovers the answer to the following question: How exactly do neurobiological processes in the brain cause consciousness? This is the most important question facing us in the biological sciences, yet it is frequently evaded, and frequently misunderstood when not evaded. In order to clear the way for an understanding of this problem. I am going to begin to answer four questions: 1. What is consciousness? 2. What is the relation of consciousness to the brain? 3. What are some of the features that an empirical theory of consciousness should try to explain? 4. What are some common mistakes to avoid?

I. What is consciousness?

Like most words, `consciousness' does not admit of a definition in terms of genus and differentia or necessary and sufficient conditions. Nonetheless, it is important to say exactly what we are talking about because the phenomenon of consciousness that we are interested in needs to be distinguished from certain other phenomena such as attention, knowledge, and self-consciousness. By `consciousness' I simply mean those subjective states of sentience or awareness that begin when one awakes in the morning from a dreamless sleep and continue throughout the day until one goes to sleep at night or falls into a coma, or dies, or otherwise becomes, as one would say, `unconscious'.

Above all, consciousness is a biological phenomenon. We should think of consciousness as part of our ordinary biological history, along with digestion, growth, mitosis and meiosis. However, though consciousness is a biological phenomenon, it has some important features that other biological phenomena do not have. The most important of these is what I have called its `subjectivity'. There is a sense in which each person's consciousness is private to that person, a sense in which he is related to his pains, tickles, itches, thoughts and feelings in a way that is quite unlike the way that others are related to those pains, tickles, itches, thoughts and feelings. This phenomenon can be described in various ways. It is sometimes described as that feature of consciousness by way of which there is something that it's like or

something that it feels like to be in a certain conscious state. If somebody asks me what it feels like to give a lecture in front of a large audience I can answer that question. But if somebody asks what it feels like to be a shingle or a stone, there is no answer to that question because shingles and stones are not conscious. The point is also put by saying that conscious states have a certain qualitative character; the states in question are sometimes described as `qualia'.

In spite of its etymology, consciousness should not be confused with knowledge, it should not be confused with attention, and it should not be confused with self-consciousness. I will consider each of these confusions in turn.

Many states of consciousness have little or nothing to do with knowledge. Conscious states of undirected anxiety or nervousness, for example, have no essential connection with knowledge.

Consciousness should not be confused with attention. Within one's field of consciousness there are certain elements that are at the focus of one's attention and certain others that are at the periphery of consciousness. It is important to emphasize this distinction because `to be conscious of' is sometimes used to mean `to pay attention to'. But the sense of consciousness that we are discussing here allows for the possibility that there are many things on the periphery of one's consciousness -- for example, a slight headache I now feel or the feeling of the shirt collar against my neck -- which are not at the centre of one's attention. I will have more to say about the distinction between the center and the periphery of consciousness in Section III.

Finally, consciousness should not be confused with self-consciousness. There are indeed certain types of animals, such as humans, that are capable of extremely complicated forms of self-referential consciousness which would normally be described as self-consciousness. For example, I think conscious feelings of shame require that the agent be conscious of himself or herself. But seeing an object or hearing a sound, for example, does not require self-consciousness. And it is not generally the case that all conscious states are also self-conscious.

II. What are the relations between consciousness and the brain?

This question is the famous `mind-body problem'. Though it has a long and sordid history in both philosophy and science, I think, in broad outline at least, it has a rather simple solution. Here it is: Conscious states are caused by lower level neurobiological processes in the brain and are themselves higher level features of the brain. The key notions here are those of *cause* and *feature*. As far as we know anything about how the world works, variable rates of neuron firings in different neuronal architectures cause all the enormous variety of our conscious life. All the stimuli we receive from the external world are converted by the nervous system into one medium, namely, variable rates of neuron firings at synapses. And equally remarkably,

these variable rates of neuron firings cause all of the colour and variety of our conscious life. The smell of the flower, the sound of the symphony, the thoughts of theorems in Euclidian geometry -- all are caused by lower level biological processes in the brain; and as far as we know, the crucial functional elements are neurons and synapses.

Of course, like any causal hypothesis this one is tentative. It might turn out that we have overestimated the importance of the neuron and the synapse. Perhaps the functional unit is a column or a whole array of neurons, but the crucial point I am trying to make now is that we are looking for causal relationships. The first step in the solution of the mind-body problem is: brain processes *cause* conscious processes.

This leaves us with the question, what is the ontology, what is the form of existence, of these conscious processes? More pointedly, does the claim that there is a causal relation between brain and consciousness commit us to a dualism of `physical' things and `mental' things? The answer is a definite no. Brain processes cause consciousness but the consciousness they cause is not some extra substance or entity. It is just a higher level feature of the whole system. The two crucial relationships between consciousness and the brain, then, can be summarized as follows: lower level neuronal processes in the brain cause consciousness and consciousness is simply a higher level feature of the system that is made up of the lower level neuronal elements.

There are many examples in nature where a higher level feature of a system is caused by lower level elements of that system, even though the feature is a feature of the system made up of those elements. Think of the liquidity of water or the transparency of glass or the solidity of a table, for example. Of course, like all analogies these analogies are imperfect and inadequate in various ways. But the important thing that I am trying to get across is this: there is no metaphysical obstacle, no logical obstacle, to claiming that the relationship between brain and consciousness is one of causation and at the same time claiming that consciousness is just a feature of the brain. Lower level elements of a system can cause higher level features of that system, even though those features are features of a system made up of the lower level elements. Notice, for example, that just as one cannot reach into a glass of water and pick out a molecule and say `This one is wet', so, one cannot point to a single synapse or neuron in the brain and say `This one is thinking about my grandmother'. As far as we know anything about it, thoughts about grandmothers occur at a much higher level than that of the single neuron or synapse, just as liquidity occurs at a much higher level than that of single molecules.

Of all the theses that I am advancing in this article, this one arouses the most opposition. I am puzzled as to why there should be so much opposition, so I want to clarify a bit further what the issues are: First, I want to argue that we simply know as a matter of fact that brain processes cause conscious states. We don't know the details about how it works and it may well be a long time before we understand the details involved. Furthermore, it seems to me an

understanding of how exactly brain processes cause conscious states may require a revolution in neurobiology. Given our present explanatory apparatus, it is not at all obvious how, within that apparatus, we can account for the causal character of the relation between neuron firings and conscious states. But, at present, from the fact that we do not know *how* it occurs, it does not follow that we do not know *that* it occurs. Many people who object to my solution (or dissolution) of the mind-body problem, object on the grounds that we have no idea how neurobiological processes could cause conscious phenomena. But that does not seem to me a conceptual or logical problem. That is an empirical/theoretical issue for the biological sciences. The problem is to figure out exactly how the system works to produce consciousness, and since we know that in fact it does produce consciousness, we have good reason to suppose that are specific neurobiological mechanisms by way of which it works.

There are certain philosophical moods we sometimes get into when it seems absolutely astounding that consciousness could be produced by electro-biochemical processes, and it seems almost impossible that we would ever be able to explain it in neurobiological terms. Whenever we get in such moods, however, it is important to remind ourselves that similar mysteries have occurred before in science. A century ago it seemed extremely mysterious, puzzling, and to some people metaphysically impossible that life should be accounted for in terms of mechanical, biological, chemical processes. But now we know that we can give such an account, and the problem of how life arises from biochemistry has been solved to the point that we find it difficult to recover, difficult to understand why it seemed such an impossibility at one time. Earlier still, electromagnetism seemed mysterious. On a Newtonian conception of the universe there seemed to be no place for the phenomenon of electromagnetism. But with the development of the theory of electromagnetism, the metaphysical worry dissolved. I believe that we are having a similar problem about consciousness now. But once we recognize the fact that conscious states are caused by neurobiological processes, we automatically convert the issue into one for theoretical scientific investigation. We have removed it from the realm of philosophical or metaphysical impossibility.

III. Some Features of Consciousness

The next step in our discussion is to list some (not all) of the essential features of consciousness which an empirical theory of the brain should be able to explain.

Subjectivity.

As I mentioned earlier, this is the most important feature. A theory of consciousness needs to explain how a set of neurobiological processes can cause a system to be in a subjective state of sentience or awareness. This phenomenon is unlike anything else in biology, and in a sense it is one of the most amazing features of nature. We resist accepting subjectivity as a ground floor, irreducible phenomenon of nature because, since the seventeenth century, we have come to believe that science must be objective. But this involves a pun on the notion of

objectivity. We are confusing the *epistemic* objectivity of scientific investigation with the *ontological* objectivity of the typical subject matter in science in disciplines such as physics and chemistry. Since science aims at objectivity in the epistemic sense that we seek truths that are not dependent on the particular point of view of this or that investigator, it has been tempting to conclude that the reality investigated by science must be objective in the sense of existing independently of the experiences in the human individual. But this last feature, ontological objectivity, is not an essential trait of science. If science is supposed to give an account of how the world works and if subjective states of consciousness are part of the world, then we should seek an (epistemically) objective account of an (ontologically) subjective reality, the reality of subjective states of consciousness. What I am arguing here is that we can have an epistemically objective science of a domain that is ontologically subjective.

Unity.

It is important to recognize that in non-pathological forms of consciousness we never just have, for example, a pain in the elbow, a feeling of warmth, or an experience of seeing something red, but we have them all occurring simultaneously as part of one unified conscious experience. Kant called this feature `the transcendental unity of apperception'. Recently, in neurobiology it has been called `the binding problem'. There are at least two aspects to this unity that require special mention. First, at any given instant all of our experiences are unified into a single conscious field. Second, the organization of our consciousness extends over more than simple instants. So, for example, if I begin speaking a sentence, I have to maintain in some sense at least an iconic memory of the beginning of the sentence so that I know what I am saying by the time I get to the end of the sentence.

Intentionality

`Intentionality' is the name that philosophers and psychologists give to that feature of many of our mental states by which they are directed at, or about states of affairs in the world. If I have a belief or a desire or a fear, there must always be some content to my belief, desire or fear. It must be about something even if the something it is about does not exist or is a hallucination. Even in cases when I am radically mistaken, there must be some mental content which purports to make reference to the world. Not all conscious states have intentionality in this sense. For example, there are states of anxiety or depression where one is not anxious or depressed about anything in particular but just is in a bad mood. That is not an intentional state. But if one is depressed about a forthcoming event, that is an intentional state because it is directed at something beyond itself.

There is a conceptual connection between consciousness and intentionality in the following respect. Though many, indeed most, of our intentional states at any given point are unconscious, nonetheless, in order for an unconscious intentional state to be genuinely an intentional state it must be accessible in principle to consciousness. It must be the sort of

thing that could be conscious even if it, in fact, is blocked by repression, brain lesion, or sheer forgetfulness.

The distinction between the center and the periphery of consciousness

At any given moment of non-pathological consciousness I have what might be called a field of consciousness. Within that field I normally pay attention to some things and not to others. So, for example, right now I am paying attention to the problem of describing consciousness but very little attention to the feeling of the shirt on my back or the tightness of my shoes. It is sometimes said that I am unconscious of these. But that is a mistake. The proof that they are a part of my conscious field is that I can at any moment shift my attention to them. But in order for me to shift my attention to them, there must be something there which I was previously not paying attention to which I am now paying attention to.

The gestalt structure of conscious experience.

Within the field of consciousness our experiences are characteristically structured in a way that goes beyond the structure of the actual stimulus. This was one of the most profound discoveries of the Gestalt psychologists. It is most obvious in the case of vision, but the phenomenon is quite general and extends beyond vision. For example, the sketchy lines drawn in Fig. 1 do not physically resemble a human face. If we actually saw someone on the street that looked like that, we would be inclined to call an ambulance. The disposition of the brain to structure degenerate stimuli into certain structured forms is so powerful that we will naturally tend to see this as a human face. Furthermore, not only do we have our conscious experiences in certain structures, but we tend also to have them as figures against backgrounds. Again, this is most obvious in the case of vision. Thus, when I look at the figure I see it against the background of the page. I see the page against the background of the table. I see the table against the background of the floor, and I see the floor against the background of the room, until we eventually reach the horizon of my visual consciousness.

The aspect of familiarity

It is a characteristic feature of non-pathological states of consciousness that they come to us with what I will call the `aspect of familiarity'. In order for me to see the objects in front of me as, for example, houses, chairs, people, tables, I have to have a prior possession of the categories of houses, chairs, people, tables. But that means that I will assimilate my experiences into a set of categories which are more or less familiar to me. When I am in an extremely strange environment, in a jungle village, for example, and the houses, people and foliage look very exotic to me, I still perceive that as a house, that as a person, that as clothing, that as a tree or a bush. The aspect of familiarity is thus a scalar phenomenon. There can be greater or lesser degrees of familiarity. But it is important to see that non-pathological forms of consciousness come to us under the aspect of familiarity. Again, one way to consider

this is to look at the pathological cases. In Capgras's syndrome, the patients are unable to acknowledge familiar people in their environment as the people they actually are. They think the spouse is not really their spouse but is an imposter, etc. This is a case of a breakdown in one aspect of familiarity. In non-pathological cases it is extremely difficult to break with the aspect of familiarity. Surrealist painters try to do it. But even in the surrealist painting, the three-headed woman is still a woman, and the drooping watch is still a watch.

Mood

Part of every normal conscious experience is the mood that pervades the experience. It need not be a mood that has a particular name to it, like depression or elation; but there is always what one might call a flavour or tone to any normal set of conscious states. So, for example, at present I am not especially depressed and I am not especially ecstatic, nor indeed, am I what one would call simply `blah'. Nonetheless, there is a certain mood to my present experiences. Mood is probably more easily explainable in biochemical terms than several of the features I have mentioned. We may be able to control, for example, pathological forms of depression by mood-altering drugs.

Boundary conditions

All of my non-pathological states of consciousness come to me with a certain sense of what one might call their `situatedness'. Though I am not thinking about it, and though it is not part of the field of my consciousness, I nonetheless know what year it is, what place I am in, what time of day it is, the season of the year it is, and usually even what month it is. All of these are the boundary conditions or the situatedness of nonpathological conscious states. Again, one can become aware of the pervasiveness of this phenomenon when it is absent. So, for example, as one gets older there is a certain feeling of vertigo that comes over one when one loses a sense of what time of year it is or what month it is. The point I am making now is that conscious states are situated and they are experienced as situated even though the details of the situation need not be part of the content of the conscious states.

IV. Some Common Mistakes about Consciousness

I would like to think that everything I have said so far is just a form of common sense. However, I have to report, from the battlefronts as it were, that the approach I am advocating to the study of consciousness is by no means universally accepted in cognitive science nor even neurobiology. Indeed, until quite recently many workers in cognitive science and neurobiology regarded the study of consciousness as somehow out of bounds for their disciplines. They thought that it was beyond the reach of science to explain why warm things feel warm to us or why red things look red to us. I think, on the contrary, that it is precisely the task of neurobiology to explain these and other questions about consciousness. Why would anyone think otherwise? Well, there are complex historical reasons, going back at least

to the seventeenth century, why people thought that consciousness was not part of the material world. A kind of residual dualism prevented people from treating consciousness as a biological phenomenon like any other. However, I am not now going to attempt to trace this history. Instead I am going to point out some common mistakes that occur when people refuse to address consciousness on its own terms.

The characteristic mistake in the study of consciousness is to ignore its essential subjectivity and to try to treat it as if it were an objective third person phenomenon. Instead of recognizing that consciousness is essentially a subjective, qualitative phenomenon, many people mistakenly suppose that its essence is that of a control mechanism or a certain kind of set of dispositions to behavior or a computer program. The two most common mistakes about consciousness are to suppose that it can be analysed behavioristically or computationally. The Turing test disposes us to make precisely these two mistakes, the mistake of behaviorism and the mistake of computationalism. It leads us to suppose that for a system to be conscious, it is both necessary and sufficient that it has the right computer program or set of programs with the right inputs and outputs. I think you have only to state this position clearly to enable you to see that it must be mistaken. A traditional objection to behaviorism was that behaviorism could not be right because a system could behave as if it were conscious without actually being conscious. There is no logical connection, no necessary connection between inner, subjective, qualitative mental states and external, publicly observable behavior. Of course, in actual fact, conscious states characteristically cause behavior. But the behavior that they cause has to be distinguished from the states themselves. The same mistake is repeated by computational accounts of consciousness. Just as behavior by itself is not sufficient for consciousness, so computational models of consciousness are not sufficient by themselves for consciousness. The computational model of consciousness stands to consciousness in the same way the computational model of anything stands to the domain being modelled. Nobody supposes that the computational model of rainstorms in London will leave us all wet. But they make the mistake of supposing that the computational model of consciousness is somehow conscious. It is the same mistake in both cases.

There is a simple demonstration that the computational model of consciousness is not sufficient for consciousness. I have given it many times before so I will not dwell on it here. Its point is simply this: *Computation is defined syntactically.* It is defined in terms of the manipulation of symbols. But the syntax by itself can never be sufficient for the sort of contents that characteristically go with conscious thoughts. Just having zeros and ones by themselves is insufficient to guarantee mental content, conscious or unconscious. This argument is sometimes called `the Chinese room argument' because I originally illustrated the point with the example of the person who goes through the computational steps for answering questions in Chinese but does not thereby acquire any understanding of Chinese.[1] The point of the parable is clear but it is usually neglected. *Syntax by itself is not sufficient for semantic content.* In all of the attacks on the Chinese room argument, I have never seen anyone come out baldly and say they think that syntax is sufficient for semantic content.

However, I now have to say that I was conceding too much in my earlier statements of this argument. I was conceding that the computational theory of the mind was at least false. But it now seems to me that it does not reach the level of falsity because it does not have a clear sense. Here is why.

The natural sciences describe features of reality that are intrinsic to the world as it exists independently of any observers. Thus, gravitational attraction, photosynthesis, and electromagnetism are all subjects of the natural sciences because they describe intrinsic features of reality. But such features such as being a bathtub, being a nice day for a picnic, being a five dollar bill or being a chair, are not subjects of the natural sciences because they are not intrinsic features of reality. All the phenomena I named -- bathtubs, etc. -- are physical objects and as physical objects have features that are intrinsic to reality. But the feature of being a bathtub or a five dollar bill exists only relative to observers and users.

Absolutely essential, then, to understanding the nature of the natural sciences is the distinction between those features of reality that are intrinsic and those that are observer-relative. Gravitational attraction is intrinsic. Being a five dollar bill is observer-relative. Now, the really deep objection to computational theories of the mind can be stated quite clearly. Computation does not name an intrinsic feature of reality but is observer-relative and this is because computation is defined in terms of symbol manipulation, but the notion of a `symbol' is not a notion of physics or chemistry. Something is a symbol only if it is used, treated or regarded as a symbol. The Chinese room argument showed that semantics is not intrinsic to syntax. But what this argument shows is that syntax is not intrinsic to physics. There are no purely physical properties that zeros and ones or symbols in general have that determine that they are symbols. Something is a symbol only relative to some observer, user or agent who assigns a symbolic interpretation to it. So the question, `Is consciousness a computer program?', lacks a clear sense. If it asks, `Can you assign a computational interpretation to those brain processes which are characteristic of consciousness?' the answer is: you can assign a computational interpretation to anything. But if the question asks, `Is consciousness intrinsically computational?' the answer is: nothing is intrinsically computational. Computation exists only relative to some agent or observer who imposes a computational interpretation on some phenomenon. This is an obvious point. I should have seen it ten years ago but I did not.

Footnotes

***** An earlier version of this article has appeared in the publications of the CIBA Foundation. The theses advanced in this paper are presented in more detail and with more supporting argument in Searle, J.R. *The Rediscovery of the Mind*, MIT Press, 1992.

1. Searle, J.R., 'Minds, Brains, and Programs,' *Behavioral and Brain Sciences,* (1980) 3, 417-457.