JULIAN JAYNES'S SOFTWARE ARCHEOLOGY

DANIEL DENNETT

Daniel Dennett received his B.A. from Harvard and his D.Phil. from Oxford. Following academic positions at the University of California at Irvine, at Harvard, and at Oxford, he joined the Department of Philosophy at Tufts University, where he is currently Professor of Philosophy. Dr. Dennett has also held the position of Fellow at the Center for Advanced Study in the Behavioral Sciences at Stanford University. He lectures widely in North America, Great Britain and Europe. His publications include books such as Brainstorms, The Mind's I *(with Douglas Hofstadter), and most recently* Elbow Room, *and journal articles on the issues of mind, consciousness, self, hallucinations, and even grey matter. He has also written numerous reviews of theories of prominent philosophers, linguists, and psychologists, such as* Passing the Buck to Biology *on Noam Chomsky and* Why do we think what we do about why we think what we do? *on Nelson Goodman.*

What a philosopher would usually do on an occasion like this is to begin to launch into a series of devastating arguments, criticisms, and counter-examples, and I am not going to do that today, because in this instance I don't think it would be very constructive. I think first it is very important to understand Julian Jaynes's project, to see a little bit more about what the whole shape of it is, and delay the barrage of nitpicking objections and criticisms until we have seen what the edifice as a whole is. After all, on the face of it, it is preposterous, and I have found that in talking with other philosophers my main task is to convince them to take it seriously when they are very reluctant to do this. I take it very seriously, so I am going to use my time to try to describe what I take the project to be. Perhaps Julian will disavow the version of Julian Jaynes I am going to present, but at least the version I am going to present is one that I take very seriously.

Now, another thing that philosophers usually do on these occasions is demand definitions of consciousness, of mind, and of all the other terms. I am not going to do that either, because I don't think that would be constructive at this point. If I thought I could bring his entire project crashing down with one deft demand for an impossible definition, I would do it, but I don't think so.

Perhaps this is an autobiographical confession: I am rather fond of his way of using these terms; I rather like his way of carving up consciousness. It is in fact very similar to the way that I independently decided to carve up consciousness some years ago.

So what then is the project? The project is, in one sense, very simple and very familiar. It is bridging what he calls the "awesome chasm" between mere inert matter and the inwardness, as he puts it, of a conscious being. Consider the awesome chasm between a brick and a bricklayer. There isn't, in Thomas Nagel's (1974) famous phrase, anything that it is like to be a brick. But there is something that it is like to be a bricklayer, and we want to know what the conditions were under which there happened to come to be entities that it was like something to be in this rather special sense. That is the story, the developmental, evolutionary, historical story that Jaynes sets out to tell.

Now, if we are going to tell this story at all, obviously we are going to have to stretch our imaginations some, because if we think about it in our habitual ways, without trying to stretch our imaginations, we just end up with a blank; it is just incomprehensible. Sherry Turkle (1984), in her new book about computers, *The Second Self*, talks about the reactions small children have to computer toys when they open them up and look inside. What they see is just an absurd little chip and a battery and that's all. They are baffled at how *that* could possibly do what they have just seen the toy do. Interestingly, she says they look at the situation, scratch their heads for a while, and then they typically say very knowingly: "It's the battery!" (A grown-up version of the same fallacy is committed by the philosopher John Searle, 1980, when he, arriving at a similar predicament, says: "It's the mysterious causal powers of the brain that explain consciousness.") Suddenly facing the absurdly large gap between what we know from the inside about consciousness and what we see if we take off the top of somebody's skull and look in can provoke

such desperate reactions. When we look at a human brain and try to think of it as the seat of all that mental activity, we see something that is just as incomprehensible as the microchip is to the child when she considers it to be the seat of all the fascinating activity that she knows so well as the behaviour of the simple toy.

Now, if we are going to do this work at all, if we are going to try to fill this gap, we are going to have to talk about consciousness, because that is what is at one edge of this large terrain. It is fascinating to me to see how reluctant, how uncomfortable, most scientifically minded people are in talking about consciousness. They realize that some day they will have to talk about consciousness — unless they are going to do what the behaviourists tried so unconvincingly to do, just dismiss the problem as not really there. If you can't quite face "feigning anaesthesia" for the rest of your days, you are going to have to admit that consciousness is a phenomenon that needs explaining. We are going to have to talk about the ephemeral, swift, curious, metaphorical features of consciousness.

Many people say: "Some day, but not yet. This enterprise is all just premature." And others say: "Leave it to the philosophers (and look what a mess they make of it)." I want to suggest that it is not premature, that in fact there is no alternative but to start looking as hard as we can at consciousness first. If we don't look at consciousness and get clear about what the destination is, and instead try to work our way up by just thinking about how the brain is put together, we won't know where we are trying to get to from where we are and we will be hopelessly lost. This is commonly referred to as the defence of the top-down strategy, and in looking at Jaynes's book again this morning I find that in his introduction he has one of the clearest and most perspicuous defences of the top-down approach that I have ever come across:

> We can only know in the nervous system what we have known in behavior first. Even if we had a complete wiring diagram of the nervous system, we still would not be able to answer our basic question. Though we knew the connections of every tickling thread of every single axon and dendrite in every species that ever existed, together with all its neurotransmitters and how they varied in its billions of synapses of every brain that ever existed, we could still never — *not ever* — from a knowledge of the brain alone know if that brain contained a consciousness like our own. We first have to start from the top, from some conception of what consciousness is, from what our own introspection is. (Jaynes, 1976b, p. 18)

When I try to make this idea clear to other people I sometimes use a somewhat threadbare analogy with computers. If you want to know what a chess-playing computer does, forget about trying to understand it from the bottom up. If you don't understand the conceptual domain in which the topic is chess — moves and strategy — you'll never make sense of what happens by building up from an analysis of the registers and logico-arithmetic operations in the central processing unit. (I am going to put this comparison with computers to a number of other uses in talking about Jaynes's work.)

If we are going to use this top-down approach, we are going to have to be bold. We are going to have to be speculative, but there is good and bad speculation, and this is not an unparalleled activity in science. An area of great fascination in science is the speculative hypothesis-spinning about the very origins of life, of the first self-replicating creatures, in the primordial soup. That sort of research has to be speculative; there simply are no data imaginable anywhere in the world today, nor are there experiments that could tease out with any certainty how that process began, but if you let your imaginations wander and start speculating about how it might have begun, you can put together some pretty interesting stories. Soon you can even begin to test some of them. Although I know that this enterprise is looked askance upon by many hard-headed scientists, it is certainly one that I would defend as a valuable part of our scientific enterprise more largely seen.

The dangers of this top-down approach of course are many. Speculation is guided largely by plausibility, and plausibility is a function of our knowledge, but also of our bad habits, misconceptions, and bits of ignorance, so when you make a mistake it tends to be huge and embarrassing. That's the price you pay in playing this game. Some people have no taste for this, but we really can't do without it. Those scientists who have no taste for this sort of speculative exercise will just have to stay in the trenches and do without it, while the rest of us risk embarrassing mistakes and have a lot of fun.

Consider the current controversy in biology between the adaptationists and their critics, Stephen J. Gould and Richard Lewontin (1979)

at Harvard. Gould and Lewontin shake an angry and puritanical finger at the adaptationists for their "just-so stories," their "panglossian" assumptions, their willingness to engage in speculation where the only real test seems to be how imaginative you can be and how plausible a story you can tell. But see Dennett (1983). One of the responses that one can make is that there is no alternative; the fossil record is simply not going to provide enough information to provide a rigorous, bottom-up, scientific account of all the important moments in the evolution of species. The provable history has to be embellished and extrapolated with a good deal of adaptationists' thinking. What we need is a *just-so story.* This term comes, of course, from Kipling and is used as a term of abuse by Gould and Lewontin, but I am happy to say that some of those who are unblushing adaptationists have simply adopted it. They say, in effect: "That's right; what we are doing is telling just-so stories, and just-so stories have a real role to play in science."

Now, when Julian Jaynes tells his story he doesn't say that it is a just-so story. He doesn't say it is just a sort of guess at how it might have been. He claims to be telling the historical truth as best as he can figure it out.

But he is also clever enough to realize that if he doesn't have the details just right, then some other story which is in very important respects rather like it must be true. So the whole account is put together by a rather interesting amalgam of aprioristic thinking about *how it had to be,* historical sleuthing, and inspired guesswork. He uses aprioristic reasoning to establish that we have to have travelled from point A to point B by *some* route — where and when the twists come is an interesting empirical question. He flavours and deepens and constrains his aprioristic thinking about how it had to be with whatever he can find about how it was and he lets his imagination go as fertile as he can and clutches at whatever straws he can find in the "fossil record."

Now, it is good to be "modular" when you do this sort of theorizing, and Jaynes's account is remarkably modular. On page 221 of his book he presents a summary of seven factors which go into his account: (1) the weakening of the auditory by the advent of writing; (2) the inherent fragility of hallucinatory control; (3) the unworkableness of gods in the chaos of historical upheaval; (4) the positing of internal cause in the observation of difference in others; (5) the acquisition of narratization from epics; (6) the survival value of deceit; and (7) a modicum of natural selection.

He has given us seven factors in this passage, and I think he would agree that you could throw out any one of them and replace it with something that simply did the same work but was historically very different; his theory would survive otherwise quite intact. Moreover, there are many details in his theory which are, as he has noted today, optional. There are ideas which are plausible, indications of the detailed way that his account *might* run, but if they turn out to be wrong they can be jettisoned with small damage to the fabric of the whole theory.

I will just mention a few that are favourites of mine. He claims, for example, that there is a little evidence to suggest that in the period when bicameral people and conscious people coexisted, there was a policy of killing obdurately bicameral children, which may in fact have hastened the demise of the bicameral type. This possible policy of "eugenics" may have given the *cultural* evolutionary process he is describing a biological (or genetic) boost, speeding up the process. You don't need it (to make the just-so story work), but there is a little evidence to suggest something like that might have happened.

Another of his optional modules is the idea that the reason that the Greeks placed the mind in the breast (in the "heart") is that when you have to make decisions you go into a stressful state which makes you breathe harder and may even make your pulse quicken. You notice the turmoil in your breast and this leads to a localization fallacy. Jaynes suggests that's why some of the Greek terms for mind have their roots as it were in the chest instead of in the head. That might be true. It might not, but it's another of the optional modules.

I don't know if Jaynes would agree with me about how much optionality there is in his system. The module I would be most interested in simply discarding is the one about hallucinations. Now you might think that if you throw out his account of hallucinations you haven't got much left of Jaynes's theory, but in fact I think you would, although he would probably resist throwing that part away.

This, then, is why I think it is a mistake, as I said at the outset, simply to bash away at the weakest links in sight, because the weakest links are almost certainly not correct — but also not critical to the enterprise as a whole.

I want to turn now to an explanation of what I find most remarkable about Julian Jaynes's just-so story, by comparing it with another (and this was in fact touched upon by one of the earlier questions). In the 17th century Thomas Hobbes (1651) asked himself where morality came from. If you look at what he called the state of nature — if you look at animals in the wild, at lions hunting and killing wildebeests, for instance — there is no morality or immorality to be seen there at all. There is no good or evil. There is killing, there is pain, but there is no "ought," there is no right, there is no wrong, there is no sin. But then look at us; here we see the institution of morality permeating the entire fabric of our lives. How did we get from there to here? Hobbes had the same sort of problem as the problem that Jaynes is facing, and his answer of course was a famous just-so story. Back in the old days, he says, man lived in the state of nature and his life was "solitary, poor, nasty, brutish and short." Then there was a sort of crisis and people got together in a group and they formed a covenant or compact, and out of this morality was born. Right and wrong came into existence as a consequence of that social contract.

Now, as history, it is absurd. Hobbes didn't think otherwise. His just-so story was quite obviously a thought experiment, a rational reconstruction or idealization. The last thing Hobbes would have done would have been to look through cuneiform records to see exactly when this particular momentous occasion happened.

But now consider the following objection to Hobbes's account, but first applied to Julian Jaynes's work. In a review of Jaynes's book some years ago, Ned Block (1981) said the whole book made one great crashing mistake, what we sometimes call a "use mention" error: confusing a phenomenon with either the name of the phenomenon or the concept of the phenomenon. Block claimed that even if everything that Jaynes said about historical events were correct, all he would have shown was not that *consciousness* arrived in 1400 B.C., but that *the concept* of consciousness arrived in 1400 B.C. People were conscious long before they had the concept of consciousness, Block declared, in the same way that there was gravity long before Newton ever hit upon the concept of gravity. The whole book in Block's view was simply a great mistake. Concept does not equal phenomenon. You can't ride the concept of the horse!

Well, now, has Jaynes made that mistake? Let's ask if Hobbes made the same mistake. Hobbes says that morality came into existence out of the social contract. Now one might say, "What a stupid mistake Hobbes has made! Maybe the *concepts* of right and wrong didn't exist before the contract of which he speaks, but certainly right and wrong themselves did. That is, people did things that were nasty and evil before they had the concepts of it."

Right and wrong, however, are parts of morality, a peculiar phenomenon that *can't* predate a certain set of concepts, including the concepts of right and wrong. The phenomenon is *created* in part by the arrival on the scene of a certain set of concepts. It is not that animals just haven't noticed that they are doing things that are evil and good. Lacking the concept, they are not doing anything right or wrong; there isn't any evil or good in their world. It's only once you get in a certain conceptual environment that the phenomenon of right and wrong, the phenomenon of morality, exists at all.

Now, I take Jaynes to be making a similarly exciting and striking move with regard to consciousness. To put it really somewhat paradoxically, you can't have consciousness until you have the concept of conciousness. In fact he has a more subtle theory than that, but that's the basic shape of the move.

These aren't the only two phenomena, morality and consciousness, that work this way. Another one that Jaynes mentions is history, and at first one thinks, "Here's another use-mention error!" At one point in the book Jaynes suggests that history was invented or discovered just a few years before Herodotus, and one starts to object that of course there was history long before there were historians, but then one realizes that in a sense Jaynes is right. Is there a *history* of lions and antelopes? Just as many years have passed for them as for us, and things have happened to them, but it is very different. Their passage of time has not been conditioned by their recognition of the transition, it has not been conditioned and tuned and modulated by any reflective consideration of that very process. So history itself, our *having* histories, is in part a function of our recognizing that very fact. Other phenomena in this category are obvious: you can't have baseball before you have the concept of baseball, you can't have money before you have the concept of money.

I have used up as much time as I should use,

but I am going to say a few more words. If you want to pursue the interesting idea that consciousness postdates the arrival of a certain set of concepts, then of course you have to have in your conceptual armamentarium the idea that concepts themselves can be preconscious, that concepts do not require consciousness. Many have held that there is no such thing as the unconscious wielding of concepts, but Jaynes's account of the origins of consciousness depends on the claim that an elaboration of a conceptual scheme under certain social and environmental pressures was the *precondition* for the emergence of consciousness as we know it. This is, to my mind, the most important claim that Jaynes makes in his book. As he puts it, "The bee has a concept of the flower," but not a conscious concept. We have a very salient theoretical role for something which we might as well call concepts, but if you don't like it we can call them schmoncepts, concept-like things that you don't have to be conscious to have.

For instance, computers have them. They are not conscious — yet — but they have lots of concepts, and in fact one way of viewing artificial intelligence is as the attempt to design conceptual systems for those computers to use. In fact this is the way people in artificial intelligence talk all the time. For instance, they may note that they have to give a robot *some concept* of an obstacle so that it can recognize this and that as an obstacle in its environment. Having figured out what concepts to give the robot or the computer, you do some fancy software design, and then you say: Here's how we have realized the concept of *causation*, or *obstacle*, or *the passage of time*, or *other sources of information* or whatever. The idea of unconscious concepts is, as a computer scientist would say, a "winning" idea, and if it is hard for you to get used to it, then at least my recommendation (along with Jaynes) would be: try harder because it is a very useful idea.

After all, one way of casting this whole question (the way that I usually think about it) is not "How do we get from the bricks, amoebas, and then apes to us?" but "How in the world could you ever make a conscious automaton, how could you make a conscious robot?" The answer, I think, is not to be found in hypotheses about hardware particularly, but in software. What you want to do is design the software in such a way that the system has a certain set of concepts. If you manage to endow the system with the right sort of concepts, you create one of those *logical spaces* that Jaynes talks about.

This in fact is a ubiquitous way of talking in the field of artificial intelligence. Consider for instance the idea of LISP. LISP is a programming language. Once you have LISP, your whole vision of how a computer is put together, and what you can do with it, changes dramatically. All sorts of things become possible that weren't possible before. Logical spaces are created that didn't exist before and you could never find them in the hardware. Such a logical space is not in the hardware, it is not in the "organs"; it is purely at the software level. Now Jaynes, in his largest and most dispensable optional module, ties his entire theory to the structure of the brain and I am fascinated to know whether there is anything in that. But I am quite content to jettison the whole business, because what I think he is really talking about is a software characterization of the mind, at the level, as a computer scientist would say, of a *virtual machine*.

The underlying hardware of the brain is just the same now as it was thousands of years ago (or it may be just the same), but what had to happen was that the environment had to be such as to encourage the development, the emergence, of certain concepts, certain software, which then set in motion some sort of chain reaction. Jaynes is saying that when the right concepts settled into place in the preconscious "minds" of our ancestors, there was a sort of explosion, like the explosion in computer science that happens when you invent something like LISP. Suddenly you discover a new logical space, where you get the sorts of different behaviours, the sorts of new powers, the sorts of new problems that we recognize as having the flavour of human consciousness.

Of course, if that is what Jaynes's theory really is, it is no wonder he has to be bold in his interpretation of the tangible evidence, because this isn't just archeology he is doing: this is *software archeology,* and software doesn't leave much of a fossil record. Software, after all, is just concepts. It is abstract and yet, of course, once it is embodied it has very real effects. So if you want to find a record of major "software" changes in archeological history, what are you going to have to look at? You are going to have to look at the "printouts," but they are very indirect. You are going to have to look at texts, and you are going to have to look at the pottery shards and figurines as Jaynes does, because that is the only place you are going to find any trace. Now,

of course, maybe the traces are just gone, maybe the "fossil record" is simply not good enough.

Jaynes's idea is that for us to be the way we are now, there has to have been a revolution — almost certainly not an *organic* revolution, but a *software* revolution — in the organization of our information processing system, and that has to have come *after* language. That, I think, is an absolutely wonderful idea, and if Jaynes is completely wrong in the details, that is a darn shame, but something like what he proposes has to be right; and we can start looking around for better modules to put in the place of the modules that he has already given us.